Estimation non-paramétrique du quantile conditionnel et apprentissage semi-paramétrique : Applications en assurance et actuariat

Muhammad Anas Knefati

École doctorale SIIM nº 521 : Laboratoire de Mathématiques et

Applications (UMR 7348)

Doctorat Université de Poitiers

PROJET DE THÈSE

pour obtenir le grade de docteur délivré par

l'École Doctorale Sciences et Ingénierie pour l'Information

Muhammad Anas KNEFATI

le 19 novembre 2015

Estimation non-paramétrique du quantile conditionnel et apprentissage semi-paramétrique : Applications en assurance et actuariat

Directeur de thèse : **Farid BENINEL**

Jury

M. Michel DELECROIX,	Professeur	Examinateur
M. Valentin PATILEA,	Professeur	Rapporteur
M. Marian HRISTACHE,	Maître de conférence	Examinateur
M. Ali GANNOUN,	Professeur	Rapporteur
M. Christophe BIERNACKI,	Professeur	Examinateur
Mme. Anne BERTRAND,	Professeur	Examinateur
M. Pierre CHAUVET,	Professeur	Examinateur

A ma mère Asous et mon père Mahmoud,

A mon épouse Amaselle,

A mes frères et ma sœur,

A mes enfants Mahmoud et Layane,

A mes amis...

Remerciements

Je tiens à remercier Monsieur Farid BENINEL, mon directeur de recherches, pour m'avoir fait confiance malgré mes difficultés en français, pour m'avoir guidé, encouragé, conseillé et pour m'avoir délégué plusieurs responsabilités.

Je remercie également Monsieur le professeur Pierre CHAUVET, pour l'intérêt manifesté pour mes travaux et à mes interrogations, ainsi que pour les conseils et les corrections.

Je remercie, aussi, l'ensemble de l'équipe de l'IMA, à l'université catholique de l'ouest, pour m'avoir accueilli, pour la durée de ma thèse afin d'y réaliser mes recherches, et pour les conseils stimulants que j'y ai reçus.

Je remercie tous ceux sans qui cette thèse ne serait pas ce qu'elle est, aussi bien par les discussions que j'ai eu la chance d'avoir avec eux, que par leurs suggestions ou contributions. Je pense ici, en particulier, à Monsieur Marian HRISTACHE, maître de conférences à l'ENSAI, qui de plus me fait l'honneur de prendre part au Jury.

Messieurs les professeurs Valentin PATILEA et Ali GANNOUN qui ont accepté d'être rapporteurs sur cette thèse ; je les en remercie, de même que pour leur participation au Jury. Leurs nombreuses remarques et suggestions, m'ont permis d'améliorer la qualité de ce rapport, et je leur en suis très reconnaissant.

Monsieur le professeur Michel DELECROIX et Madame le professeur Anne BERTRAND qui me font l'honneur de participer au Jury de soutenance ; je les en remercie profondément.

Je tiens aussi à mentionner le plaisir que j'ai eu à travailler au sein de l'équipe de statistique de l'université Rennes 2, en tant qu'ATER. J'en remercie, ici, tous les membres, et en particulier, monsieur le Professeur Jacques BENASSENI.

Je remercie également, Monsieur Salim BOUZEBDA, maître de conférences à l'UTC, pour m'avoir invité à Compiègne et conseillé, dans la phase finale de ma thèse.

Mes remerciements à l'équipe de l'Ecole Doctorale "Sciences et Ingénierie pour l'Information(S2Mi)", pour m'avoir accepté comme doctorant et fourni toute l'aide nécessaire à l'aboutissement de cette thèse.

Mes remerciements à la faculté d'Alep en Syrie, qui a financé une grande partie de mes études.

Table des matières

Liste des figures

Liste des tableaux

Introduction

L'une des préoccupation en assurance-finance est l'identification et la quantification des risques. On s'intéresse notamment, à la mise en évidence des facteurs de risque, à la mesure de la dépendance des mesures de risques utilisées (VaR par exemple).

La Value-at-Risk (*VaR*) est devenue depuis les années 90, une mesure de risque commune aux organismes financiers car son utilisation est recommandée par le Comité de Bale(2004).

La *VaR* représente la perte maximale que peut subir un gestionnaire de portefeuille durant un certain horizon, avec un seuil de confiance donné.

Les organes financiers sont donc désormais contraints à calculer périodiquement (sur un horizon de 1 jour, 10 jours ou un an selon l'activité, avec un seuil de confiance de l'ordre de $\tau = 95\%$, $\tau = 99\%$ ou $\tau = 99.5\%$) leur *VaR* et à provisionner une réserve suffisante pour essuyer les pertes éventuelles mesurées par la *VaR*.

La *VaR* n'est autre que le quantile d'ordre τ de la distribution perte ou rendement.

Les méthodes classiques d'estimation de la *VaR* s'avèrent rapidement inefficaces surtout lorsque les distributions des actifs risqués constituant le portefeuille s'approchent d'une distribution asymétrique et à queue lourde.

La majorité des estimateurs utilisés pour estimer la *VaR* sont basés sur l'inversion numérique d'estimateurs à noyau de la distribution marginale non conditionnelle (puisque la *VaR* est un quantile de la distribution étudiée) ou sur les L-statistiques utilisant une combinaison linéaire pondérée de statistiques d'ordre basées sur des observations iid.

Dans cette thèse, on s'intéresse à l'estimation de la *VaR* conditionnelle. Etant donné un ensemble de variables exogènes décrivant l'état actuel et passé des marchés financiers ainsi que l'historique de la variable d'intérêt (variable endogène, qui peut être le rendement, le coût de sinistre ou autre variable...), on se propose d'estimer la *VaR* conditionnelle (quantile conditionnel) par des estimateurs non-paramétriques utilisant des noyaux asymétriques.

Nous appliquerons les résultats obtenus sur des données réelles que nous comparons avec les estimateurs usuels existants.

Cette thèse se compose de deux parties. La première partie est consacrée à l'estimation non-paramétrique du quantile conditionnel qui revêt une importance de plus en plus grandes dans plusieurs domaines comme la finance, l'assurance, l'économie,...

Cette partie est organisée en trois chapitres.

Le chapitre 1 est consacré à une intoduction sur la régression linéaire locale, présentant les méthodes les plus utilisée dans la littérature, pour estimer le paramètre de lissage. Ce chapitre sera la base pour les deux chapitres suivants.

Le chapitre 2 traite des méthodes existantes d'estimation non-paramétriques du quantile conditionnel. Ces méthodes non-paramétriques sont les plus utilisées dans la littérature.

Nous remarquons qu'elles sont basées sur une forme unifiée. Nous abordons, aussi, les méthodes existantes d'estimation des paramètres de lissage des ces estimateurs.

Nous proposons, ensuite, d'utiliser les réseaux de neurones à fonction radiale de base pour améliorer la qualité de l'estimateur du quantile conditionnel.
Les méthodes d'estimation abordées dans ce chapitre sont comparées au moyen d'expérience numérique sur des données simulés et des données réelles.
Le chapitre 3 est consacré à un nouvel estimateur du quantile conditionnel que nous proposons, cet estimateur repose sur l'utilisation d'un noyau asymétrique en x.
Nous montrons, sous certaines hypothèses que notre estimateur s'avère plus performant que les estimateurs usuels. Pour cela, nous faisons des comparaisons théoriques et d'autres numériques basé sur des données simulée et données réelles.
La deuxième partie de la thèse est consacrée à l'apprentissage supervisé, avec notamment l'objectif de traiter le risque de défaut, en remboursement. Cette thématique représente un vaste champ de recherche, intéressant différents domaines d'application, comme l'écologie, la médecine, la biologie, la banque, l'actuariat et assurance, l'informatique,...
Cette partie est composée de trois chapitres
Le chapitre 4 est une introduction à l'apprentissage statistique et les notions de base utilisées, dans cette partie.
Le chapitre 5 est une revue des méthodes conventionnelles de classification supervisée.
Le chapitre 6 est consacré au transfert d'un d'apprentissage statistique. Nous proposons une méthode du transfert d'un modèle d'apprentissage semi-paramétrique.
La performance de cette méthode est montrée par des expériences numériques sur des données morpho-métriques et des données de credit-scoring.

Première partie

Régression quantile conditionnel

Chapitre 1

Régression non-paramétrique

Sommaire

1.1 Introduction

La régression est l'un des outils, les plus utilisés en statistique. Elle est très pratique lorsqu'on s'intéresse à la relation entre une variable réponse Y et une covariable X. La regression peut aussi être utilisée pour prédire la valeur de la variable réponse, à partir de valeurs connues d'une ou plusieurs covariables (ou variables explicatives). Les applications de la régression, couvrent la plupart des domaines.

Soit (X_i, Y_i), $i = 1, ..., n$, des observations bivariées où les X_i représentent les observations de X et les Y_i représentent celles de Y. Dans le cas général, le modèle de régression prend la forme

$$Y = m(X) + \epsilon, \tag{1.1}$$

où $m(X) = \mathbb{E}(Y|X)$; $m(X)$ est inconnu et il faut l'estimer à partir des observations (X_i, Y_i). Les termes d'erreur ϵ sont aléatoires ; Ils indiquent qu'il n'existe pas de relation exacte entre la variable réponse Y et la variable explicative X. On suppose, aussi, que $\mathbb{E}(\epsilon|X = x) = 0$ et $\mathbb{V}ar(\epsilon|X = x) = \sigma^2(x)$.

Dans la littérature, il y a deux grandes approches pour estimer la moyenne conditionnelle $m(x)$: La régression paramétrique et la régression non-paramétrique. Dans le chapitre à suivre, on présente les deux approches, en insistant sur l'approche non-paramétrique.

1.2 Régression paramétrique

La fonction de régression m, prend une forme déterminée ne dépendant que du vecteur de paramètres $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$ *i.e.,* une fois que le vecteur $\boldsymbol{\beta}$ est déterminé, la fonction m est déterminée. Alors, dans la régression paramétrique, le modèle prend la forme

$$Y_i = m(X_i, \boldsymbol{\beta}) + \epsilon_i, \quad i = 1, ..., n \text{ et } \boldsymbol{\beta} \in \mathbb{R}^p.$$

Le modèle de régression paramétrique, le plus utilisé, est le modèle linéaire simple, donné par

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, ..., n, \tag{1.2}$$

où les termes d'erreur sont supposés non corrélés, de moyenne nulle et de variance σ^2.

Dans le cas général où l'on a plusieurs variables explicatives $X^1, ..., X^p$, on introduit le modèle linéaire

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.3}$$

où $\mathbf{Y} = (Y_1, ..., Y_n)^t$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^t$ et $\mathbf{X} = (X_i^j)_{1 \le i \le n,\ 1 \le j \le p}$ est une matrice de dimension $n \times p$, pour laquelle la ligne i correspond au $i^{\text{ème}}$ individu et la colonne j correspond à la $j^{\text{ème}}$ variable explicative ; $\boldsymbol{\beta}$ est le vecteur des coefficients de régression, inconnu. Lorsque le rang de $\mathbf{X}$ est égal à p, l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y},$$

avec $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ et $\mathbb{V}ar(\hat{\boldsymbol{\beta}}) = \boldsymbol{\sigma}^2\mathbf{X}^t\mathbf{X}$.
Si les erreurs ϵ_i sont gaussiennes, le modèle (1.3) est un modèle de régression linéaire gaussienne ; La variable réponse est, dans ce cas, gaussienne ; Plus précisément $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\sigma}^2\mathbf{X}^t\mathbf{X})$.

La régression linéaire possède l'avantage d'être facile à interpréter, elle permet des tests statistiques sur les paramètres. Cependant, il arrive que le modèle linéaire ne soit pas approprié. Dans ce cas, on peut proposer un modèle non linéaire, à mêmes paramètres. Mais cette alternative peut, elle aussi, s'avérer inappropriée.

Il est, alors, préférable d'introduire un modèle, beaucoup plus flexible : la régression non-paramétrique. Pour cette approche, le modèle (1.1) est pris en compte, sans aucune hypothèse sur la forme de la fonction de régression m.

1.3 Régression non-paramétrique

Cette approche ne postule pas d'hypothèse quant à la forme de la fonction de régression ; De plus, elle s'utilise en présence de données de types divers. Dans la littérature, on rencontre plusieurs estimateurs non-paramétriques, pour estimer la moyenne conditionnelle, parmi lesquels, les estimateurs à noyaux sont les plus utilisés.

1.3.1 Estimateur de Nadaraya-Watson

L'estimateur de Nadarya-Watson (NW) est proposé par NADARAYA [1964] et par WATSON [1964]. Afin de présenter l'idée de cet estimateur, rappelons que

$$\begin{aligned} m(x) &= \mathbb{E}(\mathrm{Y}|\mathrm{X}=x), \\ &= \int y f(y|x)\,dy, \\ &= \int y \frac{f(x,y)}{f_{\mathrm{X}}(x)}\,dy, \end{aligned}$$

avec $f_{\mathrm{X}}(x)$, $f(x,y)$ et $f(y|x)$ respectivement la densité marginale de X, la densité jointe de (X, Y) et la densité conditionnelle de Y sachant X.

Afin d'estimer $m(x)$, on remplace $f(x,y)$ par son estimateur non-paramétrique, à noyaux

$$\hat{f}(x,y) = \frac{1}{nhh'} \sum_{i=1}^{n} \mathrm{K}\left(\frac{x-\mathrm{X}_i}{h}\right)\mathrm{K}_1\left(\frac{y-\mathrm{Y}_i}{h'}\right),$$

où K et K_1 sont des noyaux symétriques, h et h' sont les paramètres de lissage, respectivement en x et y.

On remplace la densité $f_{\mathrm{X}}(x)$ par

$$\hat{f}_{\mathrm{X}}(\mathbf{x}) = \frac{1}{nh} \sum_{i=1}^{n} \mathrm{K}\left(\frac{\mathbf{x}-\mathrm{X}_i}{h}\right).$$

On obtient, après simplification, la formule de l'estimateur de Nadaraya-Watson

$$\hat{m}_{\mathtt{NW}}(\mathbf{x}) = \sum_{i=1}^{n} \omega_i \mathrm{Y}i, \tag{1.4}$$

avec

$$\omega_i = \frac{\mathrm{K}\left(\frac{x-\mathrm{X}_i}{h}\right)}{\sum_{j=1}^{n} \mathrm{K}\left(\frac{x-\mathrm{X}_j}{h}\right)}.$$

De la formule (1.4), il est clair qu'il est fonction linéaire des réalisations de Y.

1.3.2 Estimateur de Gasser-Müller

Cet estimateur, proposé par GASSER et MÜLLER [1979], est donné par la formule

$$\hat{m}_{\mathtt{GM}}(x) = \sum_{i=1}^{n} \omega_i \mathrm{Y}_i \tag{1.5}$$

avec $\omega_i = \frac{1}{n}\int_{s_{i-1}}^{s_i} \mathrm{K}(\frac{x-u}{h})du$, $s_i = \frac{x_i+x_{i+1}}{2}$, pour $i = 1, ..., n-1$; s_0 et s_n sont respectivement les limites inférieure et supérieure de l'intervalle, contenant x.
On montre aisément que les estimateurs `NW` et `GM` sont solutions du problème de minimisation, selon le critère des moindres carrés pondérés

$$\arg\min_{c\in\mathbb{R}} \sum_{i=1}^{n} \omega_i (\mathrm{Y}_i - c)^2,$$

avec $\omega_i = \mathrm{K}(\frac{\mathrm{X}_i - x}{h})$ pour l'estimateur de `NW` et $\omega_i = \int_{s_{i-1}}^{s_i} \mathrm{K}(\frac{x-u}{h})du$, pour l'estimateur `GM`. Ainsi, les estimateurs `NW` et `GM` approchent, localement, l'espérance conditionnelle par une constante.

1.3.3 Régression polynomiale locale

Du cas simple où les estimateurs `NW` et `GM` approchent localement, par une constante, on suggère d'approcher localement la fonction m par un polynôme de degré p, au voisinage de x *i.e.*,

$$m(\mathrm{X}_i) = \sum_{j=0}^{p} \beta_j (\mathrm{X}_i - x)^j, \text{ pour } \mathrm{X}_i \in (x-h, x+h),$$

.
Ainsi, l'on a à minimiser, selon $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^t$, le critère des moindres carrés pondérés

$$\sum_{i=1}^{n} \left(\mathrm{Y}_i - \sum_{j=0}^{p} \beta_j (\mathrm{X}_i - x)^j \right)^2 \mathrm{K}(\frac{\mathrm{X}_i - x}{h}), \tag{1.6}$$

où K est un noyau symétrique.

L'écriture matricielle de la solution de ce problème est donnée par

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^t \mathbf{W}_x \mathbf{Z})^{-1} \mathbf{Z}^t \mathbf{W}_x \mathbf{Y}, \text{ où} \tag{1.7}$$

- $\mathbf{W}_x = \mathrm{diag}(\mathrm{K}(\frac{\mathrm{X}_1 - x}{h}), ..., \mathrm{K}(\frac{\mathrm{X}_n - x}{h}))$,
- $\mathbf{Z} = ((\mathrm{X}_i - x)^j)_{1\le i\le n,\ 0\le j\le p}$,
- $\mathrm{Y}^t = (\mathrm{Y}_1, \ldots, \mathrm{Y}_n)$.

Sous l'hypothèse de différentiabilité de la fonction m, en x, l'estimateur de $m^{(j)}(x)$, dérivée d'ordre j de la moyenne conditionnelle , est donné par identification de deux polynômes de degré p, respectivement le polynôme obtenu par développement limité et le polynôme estimation des moindres carrés pondérés *i.e.*,

$$\sum_{j=0}^{p} \frac{m_j(x)}{j!} (\mathrm{X}_i - x)^j = \sum_{j=0}^{p} \hat{\beta}_j (\mathrm{X}_i - x)^j.$$

Par, suite

$$\hat{m}^{(j)}(x) = j!\hat{\beta}_j = j!\mathbf{e}_j^t (\mathbf{Z}^t \mathbf{W}_x \mathbf{Z})^{-1} \mathbf{Z}^t \mathbf{W}_x \mathbf{Y}, \quad j = 0, ..., p, \tag{1.8}$$

où $\mathbf{e}_j \in \mathbb{R}^{p+1}$, indicateur de la $j^{\text{ème}}$ composante.

Propriétés asymptotiques

Les propriétés de l'estimateur $\hat{m}^{(j)}(x)$ défini en 1.8, sont obtenues de RUPPERT et WAND [1994] et FAN et GIJBELS [1996].
Posons

$$\mu_j = \int u^j \mathrm{K}(u)du \quad , \quad \nu_j = \int u^j \mathrm{K}^2(u)du,$$

Soient S, $\overline{\mathrm{S}}$ et S^* les matrices carrées, d'ordre $(p+1)$ et de terme général, respectivement

$$\mathrm{S}_{ij} = \mu_{j+l}, \quad \bar{\mathrm{S}}_{ij} = \mu_{j+l+1}, \quad \mathrm{S}^*_{ij} = \nu_{j+l}.$$

Soit $\mathbf{c}_p = (\mu_{p+1}, ..., \mu_{2p+1})^t$, $\bar{\mathbf{c}}_p = (\mu_{p+2}, ..., \mu_{2p+2})^t$.

En posant $\mathbf{X}^{(n)} = (\mathrm{X}_1, ..., \mathrm{X}_n)$ et en supposant

- $f(x) \neq 0$ et $f(.)$, $m^{(p+1)}(.)$ et $\sigma^2(.)$ continues au voisinage de x,
- $h \to 0$ et $nh \to \infty$,

la variance conditionnelle asymptotique

$$\mathrm{Var}\left(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)}\right) = \mathbf{e}_{j+1}^t \mathrm{S}^{-1}\mathrm{S}^*\mathrm{S}^{-1}\mathbf{e}_{j+1}\frac{j!\sigma^2(x)}{nh^{1+2j}f(x)}. \tag{1.9}$$

Le biais conditionnel asymptotique, pour $(p-j)$ impair,

$$\mathrm{Biais}\left(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)}\right) = \mathbf{e}_{j+1}^t \mathrm{S}^{-1} c_p \frac{j!}{(p+1)!} m^{(p+1)}(x)h^{p-j+1} + o_p(h^{p-j+1}) \ ; \tag{1.10}$$

Et pour $(p-j)$ pair,

$$\mathrm{Biais}\left(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)}\right) = \mathbf{e}_{j+1}^t \mathrm{S}^{-1} \bar{c}_p \frac{j!}{(p+2)!}\left\{m^{(p+1)}(x) + (p+2)m^{(p+1)}(x)\frac{f'(x)}{f(x)}\right\}h^{p-j+2} + o_p(h^{p-j+2}). \tag{1.11}$$

Noyaux équivalents

En posant

$$\mathbf{R} = \mathbf{Z}^t\mathbf{W}_x\mathbf{Z}$$

On peut réécrire l'estimateur $\hat{\beta}_j$:

$$\begin{aligned}\hat{\beta}_j = \mathbf{e}_{j+1}^t\hat{\boldsymbol{\beta}} &= \mathbf{e}_{j+1}^t\mathbf{R}^{-1}\mathbf{Z}^t\mathbf{W}_x\mathbf{Y}, \\ &= \sum_{i=1}^n \mathrm{W}_j^n\left(\frac{\mathrm{X}_i - x}{h}\right)\mathrm{Y}_i, \end{aligned} \tag{1.12}$$

où $\mathrm{W}_j^n(u) = \mathbf{e}_{j+1}^t\mathbf{R}^{-1}\mathbf{H}_p(uh)\mathrm{K}(u)/h$ et $\mathbf{H}_p(z) = (1, z, ..., z^p)^t$.

Cette écriture de l'estimateur $\hat{\boldsymbol{\beta}}$ illustre le fait que la régression polynômiale locale est très similaire aux estimateurs à noyaux classiques. De plus, la fonction de poids W_j^n dépend des valeurs x, pour lesquelles, on calcule l'estimateur, ainsi que de leur position (à l'intérieur ou au bord du support).

C'est la raison pour laquelle, la régression polynômiale corrige, automatiquement, les problèmes de bordures (FAN et GIJBELS [1996]). Ces avantages n'existent pas, dans le cas

des estimateurs classiques comme `NW` et `GM`.

Aussi,les fonctions de poids W_j^n vérifient les conditions des moments discrets :

$$\sum_{i=1}^{n}(X_i - x)^l W_j^n(\frac{X_i - x}{h}) = \delta_{j,l}, \quad 0 \le j, l \le p. \tag{1.13}$$

Pour aller plus loin sur l'importance de la régression polynomiale locale, on définit le noyau équivalent(FAN et GIJBELS [1996])

$$K_{j,p}^*(u) = \mathbf{e}_{j+1}^t S^{-1}\mathbf{H}_p(u)K(u) = (\sum_{l=0}^{p}\eta_{jl}u^l)K(u) \tag{1.14}$$

où $S^{-1} = (\eta_{jl})_{0\le j,l\le p}$. Ce noyau vérifie

$$\int u^l K_{j,p}^*(u)du = \delta_{j,l} \quad 0 \le j, l \le p \tag{1.15}$$

FAN et GIJBELS [1996] montrent que

$$W_j^n(u) = \frac{1}{nh^{j+1}f(x)}K_{j,p}^*(u)\{1 + o_p(1)\},$$

et donc, en utilisant les noyaux équivalents,

$$\begin{aligned}\hat{\beta}_j &= \sum_{i=1}^{n} W_j^n(\frac{X_i - x}{h})Y_i, \\ &= \frac{1}{nh^{j+1}f(x)}\sum_{i=1}^{n} K_{j,p}^*(\frac{X_i - x}{h})Y_i\{1 + o_p(1)\}.\end{aligned} \tag{1.16}$$

j	p	$K_{j,p}^*(u)$
0	1	$K(u)$
0	3	$\frac{\mu_4 - \mu_2 u^2}{\mu_4 - \mu_2^2}K(u)$
1	2	$\frac{1}{\mu_2}K(u)$
2	3	$\frac{u^2 - \mu_2}{\mu_4 - \mu_2^2}K(u)$

TABLEAU 1.1 – Quelques formes de noyaux équivalents $K_{j,p}^*(u)$

Aussi, il est possible de reformuler le biais conditionnel asymptotique et la variance conditionnelle asymptotique, de la régression polynomiale locale, en utilisant les noyaux équivalents.

En utilisant les formules (1.15) et (1.16),

$$\text{Biais}\left(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)}\right) = \frac{j!}{(p+1)!}\mu_{p+1}(K_{j,p}^*)m^{(p+1)}(x)h^{p-j+1} + o_p(h^{p-j+1}), \tag{1.17}$$

$$\text{Var}\left(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)}\right) = j!^2\frac{\nu_0(K_{j,p}^*)\sigma^2(x)}{nh^{1+2j}f(x)} + o_p(\frac{1}{nh^{1+2j}}), \tag{1.18}$$

où $\mu_{p+1}(K_{j,p}^*) = \int u^{p+1}K_{j,p}^*(u)du$, $\nu_0(K_{j,p}^*) = \int {K_{j,p}^*}^2(u)du$.

La fenêtre optimale

Le paramètre le plus important, déterminant la qualité de l'estimateur $\hat{m}^{(j)}(x)$, est le paramètre de lissage h, appelé aussi fenêtre. La taille des données utilisées, pour construire l'estimateur, en dépend.

Dans la littérature, on rencontre deux types de fenêtre :

(i) Les fenêtres constantes, fixées indépendamment des données (X_i, Y_i) et des valeurs x du support de f (la densité marginale de X) pour lesquelles on calcule l'estimateur $\hat{m}^{(j)}(x)$;

(ii) Les fenêtres variables. On distingue les fenêtres variables globales, ne dépendant que des observations et les fenêtres variables localement $h(x)$, dépendant des valeurs x, pour lequelles on calcule l'estimateur.

Une petite fenêtre h réduit le biais de l'estimateur, mais augmente sa variance ; Tandis qu'une grande fenêtre, réduit sa variance mais augmente son biais. Or, le problème est de choisir h minimisant, simultanément, biais et variance.

Dans ce but, on utilise comme critère à minimiser, l'erreur quadratique moyenne `MSE` (dans le cas, de fenêtres variables localement) ou l'erreur quadratique moyenne intégrée `MISE` (dans le cas, de fenêtres variables globales).

Précisémment, la fenêtre optimale locale est obtenue par minimisation de l'erreur quadratique moyenne

$$\texttt{MSE}(h) = \left(\text{Biais}(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)})\right)^2 + \text{Var}(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)})$$

La fenêtre optimale locale est approchée par minimisation de l'erreur quadratique moyenne asymptotique. En utilisant (1.17) et (1.18), puis en effectuant la minimisation on obtient

$$h_{\text{opt}}(x) = \text{C}_{j,p}(\text{K}) \left(\frac{\sigma^2(x)}{\left\{m^{(p+1)}(x)\right\}^2 f(x)} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}, \tag{1.19}$$

où

$$\text{C}_{j,p}(\text{K}) = \left(\frac{(p+1)!^2(2j+1)\nu_0(\text{K}_j^*)}{2(p-j+1)\mu_{p+1}^2(\text{K}_j^*)} \right)^{\frac{1}{2p+3}},$$

$$\mu_{p+1}(\text{K}_j^*) = \int u^{p+1}\text{K}_j^*(u)\,du, \qquad \nu_0(\text{K}_j^*) = \int \text{K}_j^{*2}(u)\,du.$$

La fenêtre globale peut être obtenue par minimisation de l'erreur quadratique moyenne intégrée `MISE`

$$\texttt{MISE}(h) = \int \left\{ \left(\text{Biais}(\hat{m}^{(j)}(x)|\mathbf{X}^{(n)})\right)^2 + \text{Var}(\hat{m}^{(j)}(x)|\text{X}^{(n)}) \right\} \omega(x)\,dx$$

où ω est une fonction de poids positive. En utilisant (1.17) et (1.18) et après minimisation, on obtient la fenêtre globale optimale, asymptotique

$$h_{\text{opt}} = \text{C}_{j,p}(\text{K}) \left(\frac{\int \sigma^2(x)\omega(x)/f(x)\,dx}{\int \left\{m^{(p+1)}(x)\right\}^2 \omega(x)\,dx} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}. \tag{1.20}$$

Pour une fonction poids $\omega(x)$ égale la densité $f(x)$, on obtient

$$h_{\text{opt}} = \mathrm{C}_{j,p}(\mathrm{K}) \left(\frac{\int \sigma^2(x)dx}{\int \left\{m^{(p+1)}(x)\right\}^2 f(x)dx} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}. \tag{1.21}$$

Effets de bord

Pour la plupart des estimateurs qui dépendent d'un paramètre de lissage, la vitesse de convergence sur les points aux bords est plus faible que pour ceux à l'intérieur. Dans la littérature, ce problème est désigné par *effets de bord*. Il est donc utile de faire quelques modifications pour traiter ce problème d'effets de bord. La régression polynomiale locale est encore attractive dans ce cas là. TIBSHIRANI et HASTIE [1987] ont remarqué empiriquement que la régression locale linéaire corrige automatiquement les effets de bord. Ce fait a été montré théoriquement par FAN et GIJBELS [1992]. RUPPERT et WAND [1994] ont étendu ce résultat au cas des estimateurs polynomiaux locaux.

Noyau optimal

Les mesures `MSE` et `MISE` pour lesquelles on a calculé les formules des fenêtres (1.19) et (1.20) dépendent de la quantité

$$\mathrm{T}_{j,p}(\mathrm{K}) = \left|\mu_{p+1}(\mathrm{K}_j^*)\right|^{2j+1} \nu_0^{p-j+1}(\mathrm{K}_j^*) \tag{1.22}$$

FAN et *al.* [1997] ont montré le théorème suivant.

Théorème 1.3.1 *Le noyau d'Epanechnikov,* $\mathrm{K}(u) = \frac{3}{4}(1-u^2)_+$*, est le meilleur choix parmi toutes les fonctions de poids positives, symétriques et lipschitziennes, au sens qu'il minimise la quantité* $\mathrm{T}_{j,p}(\mathrm{K})$.

Bien que le choix du noyau n'ait pas une grande influence sur la performance de l'estimateur $\hat{m}^{(j)}(x)$, FAN et GIJBELS [1996] recommandent d'utiliser le noyau d'Epanechnikov, fondés sur le théorème précédent ; mais aussi sur la rapidité de calcul de l'estimateur, avec ce noyau.

Choix de l'ordre du polynôme

Comme il est mentionné par FAN et GIJBELS [1996], si on augmente le degré du polynôme p, on réduit le biais de l'estimateur mais on augmente sa variabilité.
A même variance, entre deux polynômes de degré consécutifs, celui à degré impair constituera un meilleur estimateur, au sens du biais.

Cas particulier : Régression linéaire locale

On appelle régression linéaire locale, l'estimateur par polynôme local linéaire de la moyenne conditionnelle, c'est-à-dire l'estimateur $\hat{m}(x) = \hat{\beta}_0$ avec

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \left(\mathrm{Y}_i - \beta_0 - \beta_1(\mathrm{X}_i - x)\right)^2 \mathrm{K}\left(\frac{\mathrm{X}_i - x}{h}\right) \tag{1.23}$$

où K est un noyau et h est le paramètre de lissage. La solution de ce problème vient de (1.7) en mettant $j = 0$ et $p = 1$:

$$\hat{m}(x) = \sum_{i=1}^{n} \omega_i(x)\mathrm{Y}_i \tag{1.24}$$

où

$$\omega_i(x) = \frac{S_2(x) - S_1(x)(x - X_i)}{S_2(x)S_0(x) - S_1^2(x)} K(\frac{X_i - x}{h})$$

et $S_l = \frac{1}{n}\sum_{j=1}^{n}(x - X_j)^l K(\frac{X_j - x}{h})$, pour $l = 0, 1, 2$.
On peut obtenir les propriétés asymptotiques de cet estimateur en mettant encore $j = 0$ et $p = 1$ dans (1.17), (1.18), (1.19) et (1.21) :

$$\text{Biais}\left(\hat{m}^{(j)}(x)|X_1, ..., X_n\right) = \mu_2(K) m''(x) h^2 + o_p(h^2) \tag{1.25}$$

$$\text{Var}\left(\hat{m}^{(j)}(x)|X_1, ..., X_n\right) = \frac{\mu_0(K)\mu_0(K^2)\sigma^2(x)}{nhf(x)} + o(\frac{1}{nh}) \tag{1.26}$$

La fenêtre optimale locale :

$$h_{opt}(x) = \left(\frac{\mu_0(K^2)\sigma^2(x)}{\{\mu_2(K)m''(x)\}^2 f(x)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{1.27}$$

où $\mu_2(K) = \int u^2 K(u)du$ et $\mu_0(K^2) = \int K^2(u)du$.

La fenêtre globale :

$$h_{\text{opt}} = \frac{\mu_0(K^2)}{\mu_2(K)^2}\left(\frac{\int \sigma^2(x)dx}{\int \{m''(x)\}^2 f(x)dx}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{1.28}$$

Le tableau (1.2) résume les propriétés asymptotiques des estimateurs : Nadaraya-Watson (`NW`), Gasser-Müller (`GM`) et Régression Linéaire Locale (`RLL`)

Méthode	Biais	Variance
`NW`	$(m''(x) + 2\frac{m'(x)f'(x)}{f(x)})b_n$	V_n
`GM`	$m''(x)b_n$	$1.5V_n$
`RLL`	$m''(x)b_n$	V_n

TABLEAU 1.2 – Propriétés asymptotiques des estimateurs `NW`, `GM` et `RLL`

où $b_n = \frac{1}{2}\mu_2(K)$ et $V_n = \frac{\mu_0(K^2)\sigma^2(x)}{nhf(x)}$.
Le tableau 1.2 illustre la supériorité de la régression linéaire locale, comparativement aux estimateurs `NW` et `GM`.

1.3.4 Choix de la fenêtre

Dans la littérature, on rencontre deux types de méthodes pour estimer la fenêtre optimale : les méthodes de type *plug in* et les méthodes de type validation croisée.

Méthodes de type *plug-in*

Ces méthodes consistent en l'estimation des quantités inconnues intervenant dans l'expression de la fenêtre globale (Voir formule 1.28) et à les y injecter.
On mentionne deux méthodes de type *plug-in* : La règle *rule of thumb* proposée par FAN et GIJBELS [1996] et la méthode *Plug-in* directe proposée par RUPPERT et *al.* [1995].

L'algorithme *rule of thumb* de Fan et Gijbels

Cette heuristique a été proposée par FAN et GIJBELS [1996], dans le contexte de la régression non-paramétrique ; Elle permet d'estimer, de façon paramétrique, les quantités inconnues dans l'expression de la fenêtre optimale. Si son implémentation est simple et rapide, elle peut fréquemment , néanmoins, donner un mauvais choix de fenêtre.
Toutefois, elle peut être utilisée, quand le choix de la fenêtre n'a pas un rôle très important ou comme choix initial.

L'algorithme :

— En utilisant un polynome d'ordre $p \geq 2$, pour estimer la fonction moyenne conditionnelle, on obtient :
$$\hat{m}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x + \ldots + \hat{\alpha}_p x^p.$$
où le vecteur $(\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_p)^t$ est la solution du problème d'optimisation suivant :
$$\min_{\alpha \in \mathbb{R}^5} \sum_{i=1}^{n} (Y_i - \sum_{k=0}^{p} \alpha_k x^k)^2.$$

— On estime σ^2 par $\hat{\sigma}^2$ qui est la somme des carrés des résidus standardisés obtenue de cet estimateur paramétrique.

— Un estimateur de $m''(x)$ est donc obtenu par la dérivée seconde de $\hat{m}(x)$:
$$\hat{m}''(x) = \sum_{k=2}^{p} k(k-1)\hat{\alpha}_k x^{k-2}$$
.

— L'expression $\int \sigma^2(x)dx$ intervenant dans la fenêtre optimale est estimé par $\hat{\sigma}^2(b-a)$ où $[a,b]$ est le support de la densité de X ;

— Le dénominateur $\int \{m''(x)\}^2 f(x)dx$ est donc estimé par $\frac{1}{n}\sum_{i=1}^{n} \left(\hat{m}''(X_i)\right)^2$.

— L'estimation *rule of thumb* de h_{opt} est donc donnée par
$$\hat{h}_{rot} = \frac{\mu_0(K^2)}{\mu_2(K)^2} \left(\frac{\hat{\sigma}^2(b-a)}{\sum_{i=1}^{n} (\hat{m}''(X_i))^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

Méthode directe : L'algorithme de Ruppert, Seather et Wand

RUPPERT et *al.* [1995] estiment $\int \{m''(x)\}^2 f(x)dx$ par $\frac{1}{n}\sum_{i=1}^{n} \hat{m}''(X_i)^2$, et en utilisant 1.8 (pour le cas $j = 2$),
$$\hat{m}''(X_i) = 2\mathbf{e}_3^t(\mathbf{Z}^t\mathbf{W}_x\mathbf{Z})^{-1}\mathbf{X}WY$$
où $\widehat{m}(x) = \widehat{m}(x;g)$ est un estimateur polynomial cubique, dépendant de la fenêtre g.

Pour estimer σ^2, on peut prendre
$$\hat{\sigma}_p^2 = \frac{1}{n-r}\sum_{i=1}^{n} [Y_i - \hat{m}(X_i;\lambda)]^2,$$
où r est tel que $(n-r)\sigma^2 = \mathbb{E}(\sum_{i=1}^{n} [Y_i - \hat{m}(X_i;\lambda)]^2 | \mathbf{X}^{(n)})$.

RUPPERT et *al.* [1995] utilisent une heuristique *rule of thumb* pour estimer les deux fenêtres g et λ, utilisées dans cette méthode.
Sous le logiciel `R`, on peut mettre en oeuvre la méthode de RUPPERT et *al.* [1995] , en utilisant la fonction `dpill` du package `KernSmooth` qui choisit la fenêtre pour la régression linéaire locale ($p = 1$) et pour des noyaux gaussiens.

Validation croisée

La validation croisée est une technique populaire pour estimer la fenêtre optimale, dans le contexte de la régression à noyau et l'estimation de la densité (voir par exemple HÄRDLE [1990]). L'idée de cette méthode consiste à mettre de côté l'une des observations pour valider le modèle et à utiliser le reste pour construire le modèle.

Dans le contexte de la régression à noyau ($m(x) = \mathbb{E}(Y|X = x)$), le critère de la validation croisée est défini par

$$CV(h) = \sum_{i=1}^{n} (Y_i - \hat{m}_{-i}(X_i))^2 \omega(X_i)$$

où $\hat{m}_{-i}$ est l'estimateur de m sans utiliser la $i^{\text{ème}}$ observation ($i = 1, ..., n$), et ω est une fonction de poids permettant d'éviter les difficultés liées à la division par 0 ou à une faible vitesse de convergence, due aux effets de bords. La fenêtre optimale est obtenue par minimisation de $CV(h)$, selon h. Cette méthode *leave-one-out*. Elle a le défaut d'être très consommatrice de temps de calcul.
WAHBA [1977] et WAHBA et CRAVEN [1978] proposent une version améliorée, pour optimiser le temps de calcul : La fenêtre optimale h est obtenue par minimisation selon h, du critère

$$GCV(h) = \frac{MASE(h)}{(1 - trH/n)^2}$$

où $MASE(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2$ et H est une matrice carrée de taille n qui vérifie $\hat{Y} = HY$ pour $\hat{Y} = (\hat{m}(X_1), ..., \hat{m}(X_n))^t$ et $Y = (Y_1, ..., Y_n)^t$.

Il est, par ailleurs, possible d'accélérer le temps de calcul, en considérant la version "leave (l+1) out" de la validation croisée. Dans ce cas, on met de côté l observations, pour valider le modèle et on utilise le reste pour construire le modèle. Le critère de cette version est

$$CV_{l+1}(h) = \sum_{i=1}^{n} (Y_i - \hat{m}_i(X_i))^2 \omega(X_i)$$

où $\hat{m}_i$ est l'estimateur de m sans utiliser les données $(X_i, Y_i), ..., (X_{i+l}, Y_{i+l})$. La fenêtre h est donc choisie telle qu'elle minimise $CV_{l+1}(h)$.

YAO et TONG [1998] proposent une version rapide de la validation croisée. Sommairement, il suffit de partager les observations en deux parties *i.e.*, $\{(X_i, Y_i), 1 \leq i \leq k\}$ et $\{(X_i, Y_i), k+1 \leq i \leq n\}$; On utilise la première partie pour construire l'estimateur $\hat{m}_{k,h}$, et la seconde partie pour construire le critère de la validation croisée

$$CV_{YT}(h) = \frac{1}{n-k} \sum_{t=k+1}^{n} \{Y_i - \hat{m}_{k,h}(X_i)\}^2 \omega(X_i) \tag{1.29}$$

Soit h_k le paramètre qui minimise le critère CV_{YT} sur l'intervalle ouvert $]ak^{-\frac{1}{5}-\epsilon_0}, bk^{-\frac{1}{5}+\epsilon_0}[$, où $0 < a < b < \infty$ et $\epsilon_0 \in (0, \frac{1}{150})$. Puisque la fenêtre optimale h de l'estimateur $\hat{m}(.)$ est

$$h = h_k(\frac{k}{n})^{\frac{1}{5}}$$

Il est clair que cette version de la validation croisée est plus rapide que la version classique. Cette dernière nécessite $n(n-1)$ évaluations du noyau comparativement à $k(n-k)$, pour la version de Yao et Tong.

Mais si la courbe inconnue a une structure compliquée, ces méthodes échouent à estimer la fenêtre optimale de l'estimateur $\hat{m}$. Pour cela, YAO et TONG [1998] ont proposé aussi de modifier le dernier critère de la validation croisé à fin d'obtenir une fenêtre variable qui est préférable dans ce cas. Ce critère est défini comme suit :

$$\mathrm{LCV}_{\mathrm{YT}}(h) = \frac{1}{n-k} \sum_{t=k+1}^{n} \{\mathrm{Y}_t - \hat{m}_{k,h}(\mathrm{X}_t)\}^2 \omega(\mathrm{X}_t - x) \tag{1.30}$$

la fenêtre optimale h de l'estimateur $\hat{m}(.)$ est donnée par

$$h = h_k(x)\left(\frac{k}{n}\right)^{\frac{1}{5}}$$

où $h_k(x) = \arg\min_{h \in \mathcal{H}_k} \mathrm{LCV}_{\mathrm{YT}}(h)$.

1.4 Exemple sur données réelles

On va utiliser la régression linéaire locale sur une série de mesures de l'accélération de la tête (head acceleration) lors d'un accident de moto simulé. Ces données sont récupérées à partir du package `MASS` du logiciel `R` et elles contiennent deux colonnes. La première, qu'on note X, représente le temps en millisecondes après l'impact, et la deuxième, notée Y, représente l'accélération mesurée en g. Les méthodes *plug-in* CV, CV_3, CV_5, GCV et $\mathrm{CV}_{\mathrm{YT}}$ sont utilisées pour choisir la fenêtre h. La figure (1.1) montre la courbe de la régression linéaire locale avec ces méthodes de choix de la fenêtre. Tandis que la figure (1.2) représente les courbes des différents critères de validations croisées utilisés.

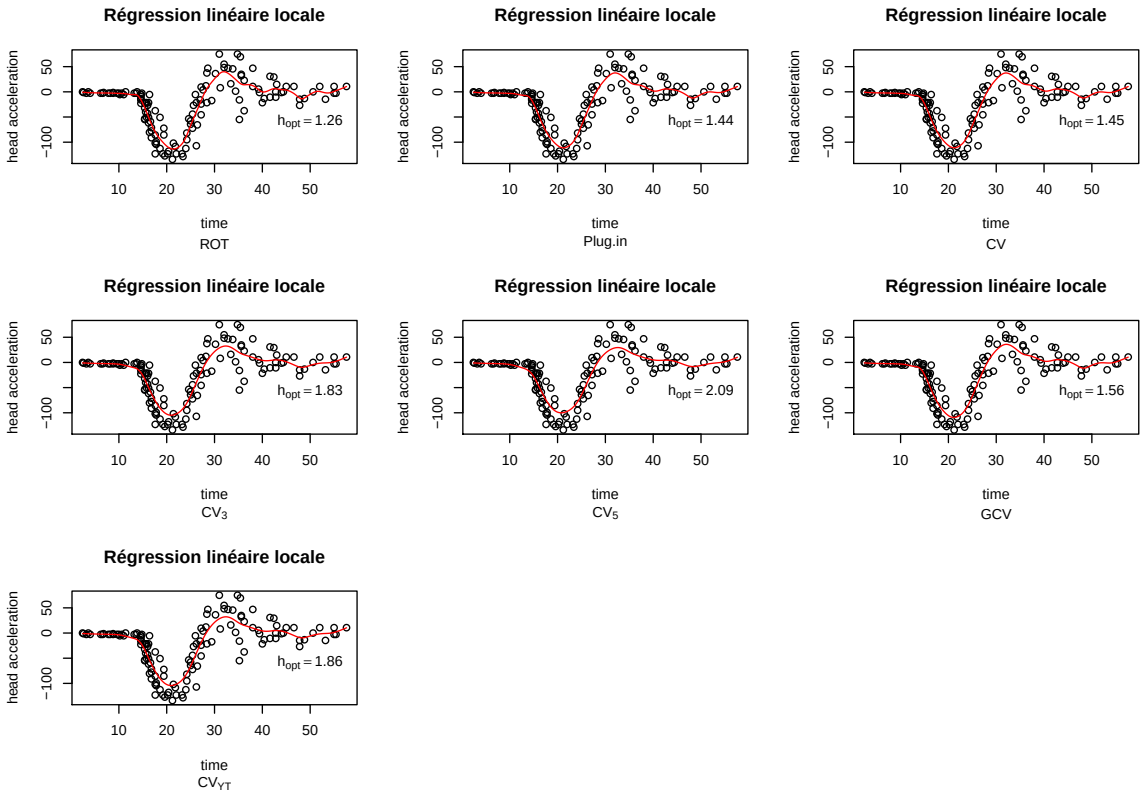


FIGURE 1.1 – Régression linéaire locale, selon plusieurs méthodes, pour le choix de fenêtre

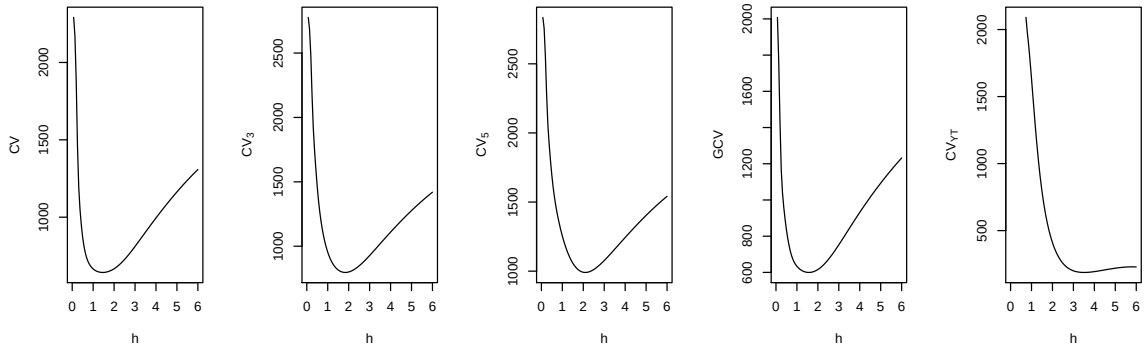


FIGURE 1.2 – Courbes de validation croisée

1.5 Références

FAN, J., T. GASSER, I. GIJBELS, M. BROCKMANN et J. ENGEL. 1997, «Local polynomial regression : optimal kernels and asymptotic minimax efficiency», *Annals of the Institute of Statistical Mathematics*, vol. 49, nº 1, p. 79–99. 12

FAN, J. et I. GIJBELS. 1992, «Variable bandwidth and local linear regression smoothers», *The Annals of Statistics*, p. 2008–2036. 12

FAN, J. et I. GIJBELS. 1996, «Local polynomial modelling and its applications», . 9, 10, 12, 13, 14

GASSER, T. et H.-G. MÜLLER. 1979, *Kernel estimation of regression functions*, Springer. 8

HÄRDLE, W. 1990, *Applied nonparametric regression*, vol. 27, Cambridge Univ Press. 15

NADARAYA, E. A. 1964, «On estimating regression», *Theory of Probability & Its Applications*, vol. 9, nº 1, p. 141–142. 7

RUPPERT, D., S. J. SHEATHER et M. P. WAND. 1995, «An effective bandwidth selector for local least squares regression», *Journal of the American Statistical Association*, vol. 90, nº 432, p. 1257–1270. 13, 14

RUPPERT, D. et M.-P. WAND. 1994, «Multivariate weighted least squares regression», *V Annals*. 9, 12

TIBSHIRANI, R. et T. HASTIE. 1987, «Local likelihood estimation», *Journal of the American Statistical Association*, vol. 82, nº 398, p. 559–567. 12

WAHBA, G. 1977, «Practical approximate solutions to linear operator equations when the data are noisy», *SIAM Journal on Numerical Analysis*, vol. 14, nº 4, p. 651–667. 15

WAHBA, G. et P. CRAVEN. 1978, «Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation.», *Numerische Mathematik*, vol. 31, p. 377–404. 15

WATSON, G. S. 1964, «Smooth regression analysis», *Sankhyā : The Indian Journal of Statistics, Series A*, p. 359–372. 7

YAO, Q. et H. TONG. 1998, «Cross-validatory bandwidth selections for regression estimation based on dependent data», *Journal of Statistical Planning and Inference*, vol. 68, nº 2, p. 387–415. 15, 16

Chapitre 2

Régression Quantile

Sommaire

2.1 Motivation

L'estimation du quantile et du quantile conditionnel revêt une importance, de plus en plus grande, dans les domaines de la finance, de l'assurance, de l'économie, de la biologie et de la médecine. A titre d'exemples, la crise immobilière apparue à la fin des années 2000 a provoqué la chute de toutes les places boursières de la planète ; L'allongement de la durée de vie et ses conséquences sur la faillite des systèmes de retraite.

Ces évènements atypiques mettent en avant le problème de la mauvaise identification et quantification des risques ainsi que le problème de la dépendance des mesures de risque utilisées (comme VaR, par exemple) et concernant l'évolution des marchés financiers en fonction du temps et en fonction du contexte économique et politique.

Ces risques financiers et économiques ont accéléré la recherche sur les mesures de solvabilité et de gestion des risque (Bale II,Solvency II). La *Value at Risk* ou VaR, consistant en le quantile de la perte maximale que peut s'autoriser un gestionnaire de portefeuille, sur un certain horizon, est devenue depuis les années 90, une mesure de risque commune aux organismes financiers suite à la recommandation de son utilisation par le Comité de Bale dans le rapport DI BASILEA PER LA VIGILANZA BANCARIA [2004].

En général, la prévision du risque dépend de la distribution de la perte conditionnelle, de l'information actuelle sur l'environnement de l'investissement (financier, politique) ; Plus précisémment, il s'agit de prédire les valeurs futures du marché, les volatilités et les corrélations. La VaR dépend des évènements passés et cela montre l'importance du quantile conditionnel (voir CAI et WANG [2008]).

Une autre application du quantile conditionnel, consiste en la construction des intervalles de prédiction, d'une une valeur prochaine, connaissant une partie des valeurs passées récentes comme série temporelle stationnaire (voir GRANGER et *al.* [1989]).

2.2 Introduction

Dans ce travail, on s'intéresse à l'étude des méthodes d'estimation non-paramétriques du quantile conditionnel. Plusieurs méthodes, pour estimer le quantile conditionnel, ont été proposées ; On les regroupe en deux grandes catégories :

— Les méthodes directes : Elles sont basées sur l'utilisation de la fonction de perte $\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\})$, pout $\tau \in (0,1)$), introduite par Koenker et Bassett(1978).
KOENKER et BASSETT JR [1978] introduisent ces estimateurs, en utilisant un modèle paramétrique pour estimer le quantile conditionnel.
KOENKER et BASSETT JR [1978] et FAN et *al.* [1994] proposent un estimateur non-paramétrique linéaire local, pour les observations *iid*.
Yu et Jones(1997,1998) approfondissent l'étude de cet estimateur non-paramétrique et proposent une méthode pour déterminer la fenêtre à utiliser. Ils comparent la méthode de régression constante locale et la méthode linéaire.
CAI et XU [2008] étendent la méthode de YU et JONES [1998], conçue pour des observations *iid*, aux données temporelles.

— Les méthodes indirectes : Elles consistent à déterminer par des méthodes numériques, un "estimateur inverse" de la fonction de répartition $F(y|x)$. Plus précisement, on détermine un estimateur de $F(y|x)$ et ensuite à résoudre en y l'équation $\widehat{F}(y|x) = \alpha$.

Parmi les estimateurs de la fonction de répartition, il y a celui de Nadaraya-Watson (1964) ; mais celui-ci souffre d'un grand biais et d'effets de bord, non négligeables.
FAN et *al.* [1996] introduisent un estimateur à double noyau, connu sous le nom d'estimateur de Yu et Jones, bien détaillé dans l'article de YU et JONES [1998], qui montrent l'absence d'effets de bord .
Cet estimateur présente des défauts, non compatibles avec les propriétés d'une fonction de répartition (non monotones et valeurs, en dehors de $[0,1]$).

Pour surmonter ces difficultés, HALL et *al.* [1999] proposent `WNM`, étendant l'estimateur (`NW`), en multipliant le noyau de `NW` par d'autre poids, choisis pour réduire le biais et effacer les effets de bord de NW. Cet estimateur est basé sur le principe de la vraisemblance empirique. CAI [2002] en a étudié les propriétés asymptotique, pour des observations d'une série temporelle. Malgré ces améliorations Cet estimateur WNW a le défaut de ne pas être continu en y, et donc de ne pas être differentiable en y.

CAI et WANG [2008] proposent un autre estimateur, plus efficace que l'estimateur `WNW`, ayant les propriétés de l'estimateur de `WNW` et de plus, continu en y. Cet estimateur est un estimateur à double noyau, comme l'estimateur de Yu et Jones ; Il est construit en utilisant les poids de `WNW`.

La fenêtre (ou les fenêtres) utilisée, dans les estimateurs non-paramétriques du quantile conditionnel, joue un rôle important dans le contrôle du lissage. Néanmoins, ce problème reste encore peu étudié. ABBERGER [1998] traite de ce problème et utilise la méthode de la validation croisée pour les données *iid*, dans le cas où l'estimateur, de la fonction de répartition, est l'estimateur `NW`.
YU et JONES [1998] proposent une méthode *plug-in* pour déterminer la fenêtre optimale de l'estimateur linéaire local du quantile et aussi, pour déterminer les deux fenêtres utilisées dans l'estimateur à double noyau de Yu et Jones.
CAI et WANG [2008] utilisent le critère, proposé dans CAI et TIWARI [2000] pour estimer la fenêtre optimale de l'estimateur de `WNW` et de l'estimateur à double noyau, qu'ils introduisent.
ATTAR [2008] utilise une méthode *plug-in* itérative, pour déterminer la fenêtre optimale de l'estimateur linéaire local et celle de l'estimateur à double noyau de Yu et Jones.

2.3 Quantile et Quantile conditionnel : Définition

2.3.1 Quantile

Soit Y une variable aléatoire réelle, F sa fonction de répartition et $\tau \in]0,1[$. Le quantile d'ordre τ de la variable Y, noté $q_Y(\tau)$, est la plus petite valeur y vérifiant $F(y) \geq \tau$ *i.e.,*

$$q_Y(\tau) := F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}. \tag{2.1}$$

Aussi, cette définition peut être présentée sous forme d'un optimum :

$$q_Y(\tau) = \arg\min_{a\in\mathbb{R}} \mathbb{E}\big[\rho_\tau(Y-a)\big], \tag{2.2}$$

où ρ_τ est la fonction de perte de Koenker et Basset(1978), donnée par

$$\rho_\tau(u) = u(\tau - \mathbb{1}\{u<0\}). \tag{2.3}$$

En effet,

$$\mathbb{E}\big[\rho_\tau(Y-a)\big] = (\tau-1)\int_{-\infty}^{a}(y-a)dF(y) + \tau\int_{a}^{\infty}(y-a)dF(y).$$

Par différentiation, par rapport à a, on obtient

$$0 = (1-\tau)\int_{-\infty}^{a} dF(y) - \tau\int_{a}^{\infty} dF(y);$$

Et par suite, $F(a) = \tau$.
Ainsi, tous les éléments de l'ensemble $\{a : F(a) = \tau\}$ minimisent l'espérance de la perte. Si F est monotone, la solution est unique $\hat{a} = F^{-1}(\tau)$; Sinon, on a une infinité de quantiles d'ordre τ (constituant un intervalle) ; On choisit le plus petit, car la fonction quantile est continue à gauche.

2.3.2 Quantile conditionnel

Soit X, Y deux variables aléatoires réelles, $F(y|x)$ la fonction de répartition conditionnelle de Y sachant $X = x$, et $\tau \in]0,1[$.
D'une manière similaire à la définition du quantile, le quantile conditionnel $q_Y(\tau|x)$ (noté simplement $q_\tau(x)$) est défini par

$$q_\tau(x) = \inf\{y \in \mathbb{R} : F(y|x) \geq \tau\}; \tag{2.4}$$

Ou de manière équivalente,

$$q_\tau(x) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\{\rho_\tau(Y-\theta)|X=x\}, \tag{2.5}$$

avec ρ_τ est la fonction perte définie par l'équation (2.3).

Pour revenir à la fonction ρ_τ, celle-ci peut se présenter sous forme de plusieurs écritures équivalentes :

$$\begin{aligned}
\rho_\tau &= u(\tau - \mathbb{1}(u<0)),\\
&= \tau u\mathbb{1}(u>0) + (\tau-1)u\mathbb{1}(u<0),\\
&= \frac{1}{2}|u| + (\tau - \frac{1}{2})u.
\end{aligned}$$

On vérifie aisément que $\rho_\tau(u) \to \infty$ quand $|u| \to \infty$ et que $\rho_\tau(0) = 0$.
Pour $\tau = 0.5$, le quantile conditionnel est la médiane conditionnelle.
La figure (2.1) présente le graphe de la fonction ρ_τ, pour différents niveaux de τ.

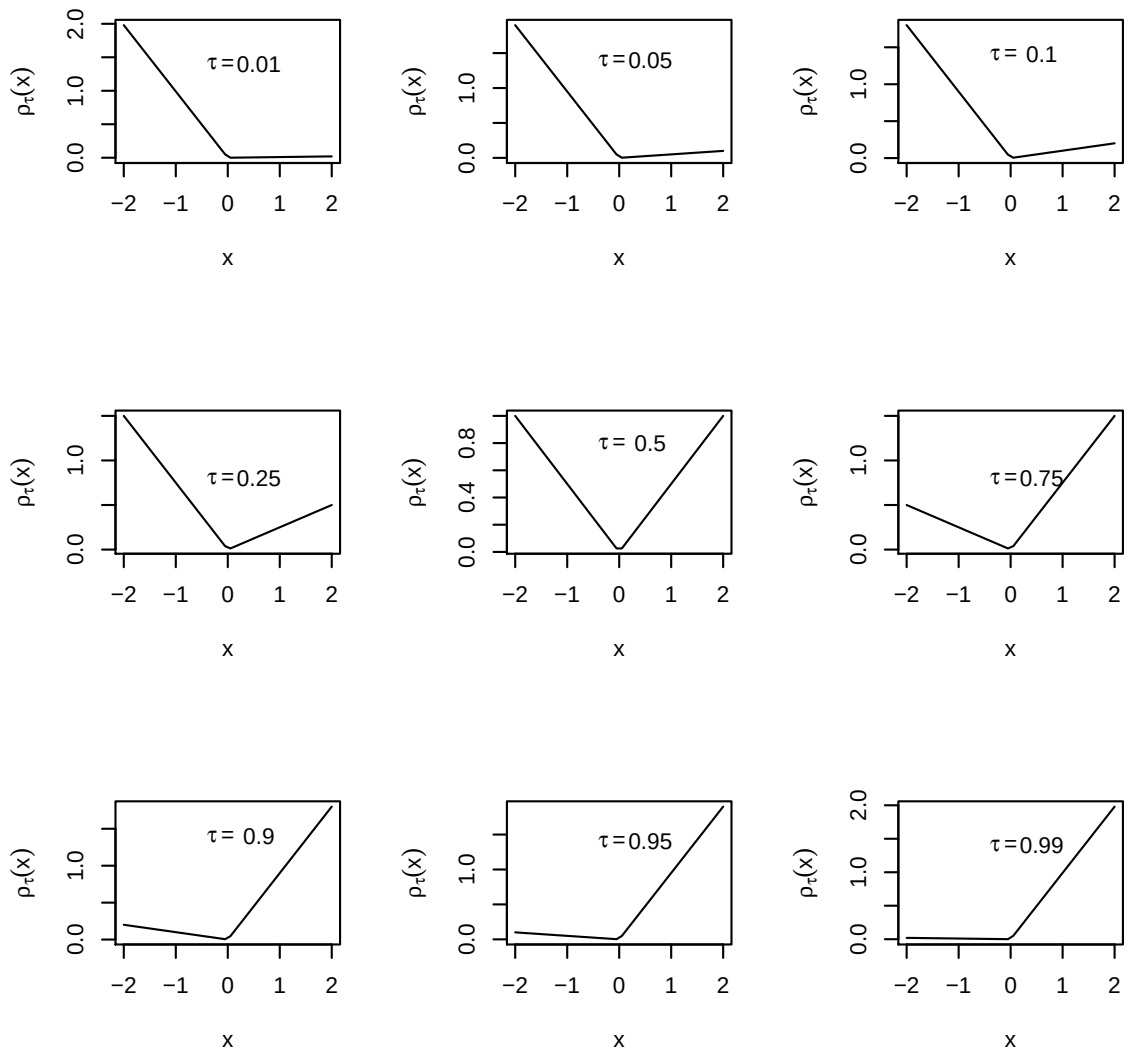


FIGURE 2.1 – Graphe de la fonction ρ_τ, pour différants niveaux τ

.

2.4 Méthodes d'estimation du quantile conditionnel

Soit $\{(X_i, Y_i), 1 \leq i \leq n\}$, des observations *iid*, avec $X_i \in \mathbb{R}$, Y_i est la variable réponse. On peut classer les méthodes d'estimation du quantile conditionnel, selon deux approches :

— L'approche directe ;

— L'approches indirecte.

2.5 Approches directes

Ces approches sont étroitement liées à la définition (2.5). Elle ne nécessitent pas l'estimation de la fonction de répartition conditionnelle. Ces approches peuvent être classées aussi en méthodes paramétriques et méthodes non-paramétriques.

2.5.1 Méthodes paramétriques

Régression quantile paramétrique

La régression quantile paramétrique a été introduite la première fois par KOENKER et BASSETT JR [1978]. En substituant dans l'équation 2.2, l'espérance par la moyenne empirique correspondante *i.e.*, $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - a)$, le problème de recherche du quantile revient en le problème d'optimisation

$$\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - a).$$

Et puisque la moyenne de l'échantillon peut être considérée comme la solution du problème d'optimisation

$$\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2,$$

en exprimant la moyenne conditionnelle de Y sachant $X = x$ comme $\mu(x) = x\beta$, alors β peut être estimé par la solution du problème

$$\min_{\beta\in\mathbb{R}} \sum_{i=1}^{n} (y_i - x_i\beta)^2.$$

De manière analogue, puisque le quantile d'ordre τ est la solution de

$$\min_{\alpha\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha),$$

On peut spécifier le quantile conditionnel d'ordre τ comme $q_y(\tau|x) = x\beta(\tau)$ avec $\beta(\tau)$ solution du problème

$$\min_{\beta\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta).$$

Cette méthode est ce qu'on appelle la régression quantile paramétrique.

Régression quantile non-paramétrique

Dans plusieurs situations, la régression quantile paramétrique échoue ; Dans ce cas, on utilise la régression non-paramétrique. KOENKER et BASSETT JR [1978] propose la régression quantile non-paramétrique, en utilisant une version non pondérée. Le quantile conditionnel $q_\tau(x)$ est une fonction en x. Si on suppose que $q_\tau(x)$ est une fonction dérivable d'ordre $p+1$; Au voisinage de x_0, on peut l'approcher, selon Taylor, par

$$q_\tau(x) = \sum_{j=1}^{p} \frac{q_Y^{(j)}(\tau|x_0)}{j!}(x-x_0)^j.$$

Cette approximation permet d'utiliser une régression locale pondérée (*i.e.,* à noyaux), pour estimer le quantile conditionnel. Cela consiste en le problème d'optimisation :

$$\begin{aligned} &\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\rho_\tau\left(Y_i-\sum_{j=0}^{p}(X_i-x_0)^j\beta_j(x_0)\right)K(\frac{X_i-x_0}{h})\right\} \\ = &\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\left(|Y_i-\sum_{j=0}^{p}(X_i-x_0)^j\beta_j(x_0)|+(2\tau-1)(Y_i-\sum_{j=0}^{p}(X_i-x_0)^j\beta_j(x_0))\right)K(\frac{X_i-x_0}{h})\right\} \end{aligned}$$

La solution $\widehat{\boldsymbol{\beta}}$ de ce problème permet l'estimation de $q_\tau(x_0)$ et de ses dérivées :

$$\hat{q}_\tau^{(j)}(x_0) = j!\hat{\beta}_j(x_0), \quad j=0,\ldots,p.$$

Ce problème d'optimisation ne peut être résolu analytiquement ; Il doit l'être de façon numérique.
Sous le logiciel R, par exemple, la fonction `rq` du package `quantreg` calcule la solution de ce problème.
L'estimateur $\hat{q}_Y(\tau|x)$ est l'estimation par régression quantile polynomiale locale. On présente, dans ce qui suit, deux cas particuliers.

1) **Régression constante locale :**
Si $p=0$, on obtient la méthode dite de la constante locale :

$$\widehat{q}_\tau(x) = \arg\min_a \sum_{i=1}^{n}\rho_\tau(Y_i-a)K(\frac{x-X}{h})$$

l'estimateur du quantile conditionnel est donné par $\widehat{q}_\tau(x)=\hat{a}$. On s'intéresse, dans ce qui suit, aux propriétés de cet estimateur *i.e.,* la variance, l'erreur quadratique moyenne MSE (Mean Square Error) et la fenêtre de lissage optimale. Ces résultats sont tirés de JONES et HALL [1990].
Les propriétés des estimateurs pour des valeurs x se trouvant à l'intérieur des bornes, diffèrent des estimateurs en des valeurs x se situant à l'extérieur de ces bornes. On distinguera, les deux cas de figure.

Cas 1 : x est un point intérieur
Soit

- g la densité de X ;
- $\mu_2(K)=\int u^2K(u)du$, $\mu_0(K^2)=\int K^2(u)du$;

- $\alpha(\mathrm{K}) = \int \mathrm{G}(u)(1-\mathrm{G}(u))du$;
- $\mathrm{F}^{ab}(q_\tau(x))) = \dfrac{\partial^{ab}}{\partial^a z \partial^b y}\mathrm{F}(y,z)|_{x,q_\tau(x)}$.

Considérons les conditions ci-après :

(a1) Pour x et y fixés, les dérivées partielles de $\mathrm{F}(x,y)$, $f(x,y)$ et $g(x)$ existent, sont bornées et continues à l'intérieur et au bord des données ;

(a2) $g(x) > 0$ et la densité conditionnelle $f(y|x) > 0$ est bornée ;

(a3) Le noyau K(.) est symétrique, borné et à support compact ;

(a4) $h \to 0$ et $nh \to \infty$, quand $n \to \infty$.

A l'intérieur des bornes et sous les conditions (a1), (a2), (a3) et (a4) ci-dessus, le biais asymptotique est

$$\mathrm{B}iais(\widehat{q}_\tau(x)) = \frac{1}{2}\mu_2(\mathrm{K})\left\{-\frac{\mathrm{F}^{20}(q_\tau(x)}{f(q_\tau(x)|x)} + 2\frac{g'(x)q'_\tau(x)}{g(x)}\right\}.$$

La variance asymptotique est

$$\mathrm{V}ar(\hat{q}_\tau(x)) = \frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{nhg(x)f^2(q_\tau(x)|x)}$$

Par addition du carré du biais asymptotique et de la variance asymptotique, on obtient l'erreur quadratique moyenne asymptotique (ou AMSE) :

$$\mathtt{AMSE}(\hat{q}_\mathrm{Y}(\tau;x)) = \frac{1}{4}\mu_2^2(\mathrm{K})\left\{-\frac{\mathrm{F}^{20}(q_\tau(x)|x}{f(q_\tau(x)|x)} + 2\frac{g'(x)q'_\tau(x)}{g(x)}\right\}^2 + \frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{nhg(x)f^2(q_\tau(x)|x)}$$

La fenêtre optimale est obtenue par minimisation de AMSE(h), selon h

$$h_{opt}(x) = \frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)f(q_\tau(x)|x)^2\mathrm{B}_\tau(x)},$$

avec

$$\mathrm{B}_\tau(x) = -\frac{\mathrm{F}^{20}(q_\tau(x)|x)}{f(q_\tau(x)|x)} + 2\frac{g'(x)q'_\tau(x)}{g(x)} \tag{2.6}$$

Cas 2 : x est un point frontière

Maintenant, on suppose que le support de $g(x)$ est l'intervalle $[0,1]$ et que l'on veuille calculer l'estimateur aux points problématiques situés à gauche ou à droite *i.e.*, $x = hc$, $0 < c < 1$ ou $x = 1 - hc$. Si le support du noyau K est $[0,1]$, alors, le biais asymptotique

$$\mathrm{B}iais(\hat{q}_\mathrm{Y}(\tau;x)) = h\frac{a_1(c;\mathrm{K})}{a_2(c;\mathrm{K})}\left\{-\frac{\mathrm{F}^{10}(q\tau(0^+)|0^+)}{f(q_\tau(0^+)|0^+)}\right\}$$

La variance asymptotique est donnée par

$$\mathrm{V}ar(\hat{q}_\mathrm{Y}(\tau;x)) = \frac{\tau(1-\tau)\mathrm{V}(c;\mathrm{K})}{nhg(0^+)f^2(q_\tau(0^+)|0^+)},$$

avec $a_l(c;\mathrm{K}) = \int_{-1}^{c} u^l \mathrm{K}(u)du$ et pour simplifier $\psi(0^+) = \lim_{x\to 0}\psi(x)dx$ pour toute fonction ψ.

L'AMSE peut être facilement trouvée à partir du biais et de la variance.

2) **Régression linéaire locale** :
Ici, on approche localement, Y par un pôlynome de degré $p = 1$; Il s'agit donc, de la régression linéaire locale *i.e.,* du problème d'optimisation

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_{i=1}^{n} \rho_\tau(Y_i - a - b(X_i - x))K(\frac{x - X}{h})$$

où l'estimateur du quantile conditionnel $\hat{q}_Y(\tau; x) = \hat{a}$.

Comme pour le cas de la régression par constante locale, on s'intéresse aux propriétés de cet estimateur, données dans FAN et *al.* [1994].

Sous les conditions (a), à l'intérieur des bornes, le biais asymptotique est

$$Biais(\hat{q}_\tau(x)) = \frac{1}{2} h^2 \mu_2(K) q''_\tau(x).$$

La variance asymptotique est

$$Var(\hat{q}_\tau(x)) = \frac{\tau(1-\tau)\mu_0(K^2)}{nhg(x)f^2(q_\tau(x)|x)}$$

L'AMSE est donc donnée par

$$AMSE(\hat{q}_\tau(x)) = \frac{1}{4} h^4 \mu_2^2(K) q''_\tau(x)^2 + \frac{\tau(1-\tau)\mu_0(K^2)}{nhg(x)f^2(q_\tau(x)|x)} \tag{2.7}$$

Par minimisation de AMSE(h), selon h, on obtient la fenêtre optimale

$$h_{opt}(x) = \frac{\tau(1-\tau)\mu_0(K^2)}{\mu_2^2(K) g(x) q"_\tau(x)^2 f(q_\tau(x)|x)^2}$$

Sans perte, si $x = hc, \quad 0 < c < 1$ est un point au bord gauche, et g et K de support, respectivement $[0,1]$ et $[-1,1]$, on obtient comme biais

$$Biais(\hat{q}_\tau(x)) = \frac{1}{2} h^2 \alpha_c(K) q''_\tau(0^+);$$

Et comme variance asymptotique

$$Var(\hat{q}_\tau(x)) = \frac{\tau(1-\tau)\beta_c(K)}{nhg(0^+)f^2(q_\tau(0^+)|0^+)},$$

où

- $\alpha_c(K) = \frac{a_2^2(c;K) - a_1(c;K)a_3(c;K)}{a_0(c;K)a_2(c;K) - a_1^2(c;K)}$,
- $\beta_c(K) = \frac{\int_{-1}^{c} \{a_2(c;K) - a_1(c;K)u\}^2 K(u)du}{\{a_0(c;K)a_2(c;K) - a_1(c;K)\}^2}$,
- $a_l(c;K) = \int_{-1}^{c} u^l K(u)du.$

YU et JONES [1997] comparent la régression constante locale et la régression linéaire locale : Les deux méthodes donnent des estimations voisines, pour x à l'intérieur. Mais au bord, la régression linéaire locale s'avère meilleure que la régression constant (biais respectivement d'ordre $O(h^2)$ et $o(h)$.

2.6 Approches implicites

Les méthodes étudiées, dans cette partie, visent à l'estimation du quantile tel que défini en (2.4). Pour ces méthodes, l'estimation de la fonction de répartition conditionnelle et son inversion se font par des méthodes numériques.
Ces approches peuvent, elles aussi, être classées en méthodes paramétriques et non paramétriques.

2.6.1 Méthodes paramétriques

Pour ces méthodes, on connaît la forme de la fonction de répartition conditionnelle F comme fonction dépendant du paramètre θ. Un estimateur de F est alors obtenu en remplaçant, dans la formule de F, θ par son estimateur $\hat{\theta}$.
Comme $q_\tau(x) = F_\theta^{-1}(\tau|x)$, alors on estime le quantile conditionnel $q_\tau(x)$, par

$$\hat{q}_\tau(x) = F_{\hat{\theta}}^{-1}(\tau|x).$$

On peut mentionner, comme exemple d'utilisation de cette approche, les modèles `GARCH` et `RiskMetrics`.

2.6.2 Méthodes non-paramétriques

Quand on ne connaît pas la forme de la fonction de répartition, on utilise les méthodes d'estimation non-paramétriques. L'estimation du quantile conditionnel, se fait à l'aide d'estimateurs à noyaux.
On présente, dans ce qui suit, les trois principaux estimateurs à noyaux, rencontrés dans la littérature.

Estimateur à double noyaux proposé par Yu et Jones

Fan et *al.* [1996] introduisent un estimateur à double noyau, pour estimer la fonction de répartition conditionnelle ; Yu et Jones [1998] étudient, avec plus de détails cet estimateur.
Il repose sur le lien entre la fonction de la densité conditionnelle et le problème de régression non paramétrique. Plus précisément, pour h_2 au voisinage de 0,

$$\mathbb{E}\left(L_{h_2}(Y_i - y)\middle| X_i = x\right) \approx f(y|x), \quad i = 1, ..., n,$$

avec $f(y|x)$ la densité conditionnelle de Y sachant $X = x$, h_2 une fenêtre dans la direction y, $L_{h_2}(u) = L(u/h_2)/h_2$ et L est un noyau symétrique.

En posant $Y_i^*(y) = L_{h_2}(y - Y_i)$, la densité conditionnelle $f(y|x)$ peut être considérée comme une régression non-paramétrique de la variable observée $Y_i^*(y)$ (en X_i).
On peut, donc, considérer le problème de minimisation, selon a et b,

$$\sum_{i=1}^{n} \{Y_i^*(y) - a - b(X_i - x)\}^2 K\left(\frac{x - X_i}{h_1}\right), \tag{2.8}$$

avec

- K(.) une fonction noyau,

- $h_1 = h_1(n) > 0$ est une fenêtre en direction x, vérifiant $h_1 \to 0$ et $nh_1 \to \infty$, quand $n \to \infty$.

Comme vu, en section (1.3.3), la solution de ce problème est donnée par

$$\hat{f}(y|x) = \sum_{i=1}^{n} \omega_i(x, h_1) Y_i^*(y), \tag{2.9}$$

avec

- $\omega_i(x, h_1) = \frac{[S_2(x) - (x - X_i)S_1(x)]}{S_2(x) - S_1(x)^2} K(\frac{x - X_i}{h_1}), \quad i = 1, \ldots, n$
- $S_j(x) = \sum_{i=1}^{n} K(\frac{x - X_i}{h_1})(x - X_i)^j, \quad j = 0, 1, 2.$

Dans le cas particulier de l'equation (1.13), ces poids vérifient les conditions des moments discrets :

$$\sum_{i=1}^{n} \omega_i(x, h)(X_i - x)^j = \delta_{0j} = \begin{cases} 1 & \text{si } j = 0\,; \\ 0 & \text{sinon.} \end{cases} \tag{2.10}$$

Un estimateur $F(y|x)$ est donc

$$\hat{F}(y|x) = \int_{-\infty}^{y} \hat{f}(u|x)\,du = \sum_{i=1}^{n} \omega_i(x, h_1) G(\frac{y - Y_i}{h_2})$$

où G(.) est la fonction de répartition de L(.).

Cet estimateur est linéaire local et puisqu'il contient deux noyaux, on l'appelle estimateur à double noyau linéaire local. Il a le défaut de prendre, parfois, des valeurs à l'extérieur de l'intervalle $[0, 1]$, mais il n'est pas difficile de trouver un algorithme pour calculer son inverse(Yu et Jones [1998]).
Il est clair aussi que $\hat{F}(y|x)$ est continue et differentiable par rapport à y avec $\hat{F}(-\infty|x) = 0$ et $\hat{F}(\infty|x) = 1$. L'estimateur du quantile conditionnel est donc donné par

$$\hat{q}_Y(\tau|x) = \hat{F}^{-1}(\tau|x). \tag{2.11}$$

L'estimateur $\hat{F}(y|x)$ peut ne pas être monotone en y, et afin de traiter ce problème pendant l'implementation, Yu et Jones [1998] ont proposé, dans ce cas, l'algorithme suivant. On prend pour la médiane conditionnelle($\tau = \frac{1}{2}$) n'importe quelle valeur qui vérifie l'équation (2.11). Pour $\tau > \frac{1}{2}$, on prend $\hat{q}_\tau(x)$ qui est la valeur la plus élevée telle qu'elle vérifie (2.11).
On passe maintenant aux propriétés asymptotiques de cet estimateur qui sont récupérées de Yu et Jones [1998]. Considérons les conditions

b1) Le noyau L(.) est symétrique, borné et il est à support compact.

b2) $h_1 \to 0$, $nh_1 \to \infty$, et $h_2 \to 0$, $nh_2 \to \infty$, quand $n \to \infty$

b3) $h_2 = o(h_1)$

Sous les conditions a1),a2),a3),a4), b1),b2), b3) et à l'intérieur des bornes,
le biais asymptotique est

$$Biais(\hat{q}_\tau(x)|x)) = \frac{1}{2}\{\mu_2(K)\frac{F^{20}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_1^2 + \mu_2(W)\frac{F^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_2^2\};$$

Et la variance asymptotique est

$$Var(\widehat{q}_\tau(x)|x)) = \frac{\mu_0(K^2)\mu_0(K^2)}{nh_1 g(x) f^2(q_\tau(x)|x)}(\tau(1-\tau) - h_2 f(q_\tau(x)|x)\alpha(W)).$$

L'AMSE est donc donnée par

$$\begin{aligned} \text{AMSE}(\widehat{q}_\tau(x)|x)) &= \frac{1}{4}\{\mu_2(K)\frac{F^{20}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_1^2 + \mu_2(W)\frac{F^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_2^2\}^2 \\ &+ \frac{\mu_0(K^2)}{nh_1 g(x) f^2(q_\tau(x)|x)}(\tau(1-\tau) - h_2 f(q_\tau(x)|x)\alpha(W)). \end{aligned}$$

Pour $h_1 >> h_2$, on obtient la fenêtre optimale de h_1, en minimisant $\text{AMSE}(h_1)$ selon fonction de h_1 :

$$h_{1,opt}(x) = \left(\frac{\tau(1-\tau)n\mu_0(K^2)}{\mu_2^2(K)g(x)F^{02}(q_\tau(x)|x)^2}\right)^{\frac{1}{5}}.$$

Maintenant, si $x = hc$ $(0 < c < 1)$ un point au bord gauche, g et K de support, respectivement $[0,1]$ et $[-1,1]$, le biais asymptotique est

$$Biais(\hat{q}_\tau(x)) = \frac{1}{2}\{\alpha_c(K)\frac{F^{20}(q_\tau(0^+)|0^+)}{f(q_\tau(0^+)|0^+)}h_1^2 + \mu_2(W)\frac{F^{02}(q_\tau(0^+)|0^+)}{f(q_\tau(0^+)|0^+)}h_2^2\};$$

Et la variance asymptotique est

$$Var(\hat{q}_\tau(x)) = \frac{\beta_c(K)}{nhg(0^+)f^2(q_\tau(0^+)|0^+)}(\tau(1-\tau) - h_2 f(q_\tau(0^+)|0^+)\alpha(W)).$$

Cet estimateur se caractérise par l'absence d'effet de bord et une bonne efficacité (voir YU et JONES [1998]) ; Cependant, il a le défaut de dépendre de $\widehat{F}(y|x)$ pouvant prendre des valeurs, en dehors de $[0,1]$ et ne pas être monotone croissant. Ces défauts posent un problème pour estimer le quantile conditionnel, par la méthode d'inversion.
Afin de surmonter ces difficultés, HALL et *al.* [1999] et CAI [2002] proposent l'estimateur dit de Nadaraya-Watson pondéré(WNW).

Estimateur de Nadaraya-Watson pondéré (WNW)

HALL et *al.* [1999], en vue d'affaiblir le biais, proposent l'estimateur Nadaraya-Watson pondéré de $F(y|x)$ en remplaçant les poids classiques $\omega_i(x,h) = \frac{K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}$, par des poids vérifiant les conditions des moments discrets (2.10).
Ce changement de pondération permet d'obtenir un estimateur $\hat{F}(y|x)$ équivalent à l'estimateur de Yu et Jones, monotone en y et à valeurs dans [0 1].

$$\hat{F}(y|x) = \sum_{i=1}^n \omega_i(x,h)\mathbb{1}(Y_i \leq y), \tag{2.12}$$

où les poids $\{\omega_i(x,h)\}$ sont donnés par

$$\omega_i(x,h) = \frac{p_i(x)K_h(x - X_i)}{\sum_{i=1}^n p_i(x)K_h(x - X_i)}, \tag{2.13}$$

avec $K_h(.) = K(./h)/h$ pour un noyau symétrique K, sous les contraintes

c1) $\sum_{i=1}^n p_i(x) = 1$,

c2) $\sum_{i=1}^{n} p_i(x)(\mathrm{X}_i - x)\mathrm{K}_h(\mathrm{X}_i - x) = 0$.

Les fonctions $p_i(x)$ qui vérifient ces conditions ne sont pas uniques, et une façon pour les préciser est d'utiliser le principe de la vraisemblance empirique : la somme logarithmique $\sum_{j=1}^{n} \log\{p_j(x)\}$ est maximisée sous les contraintes c1) et c2).
Ainsi, ce problème d'optimisation sous les contraintes c1) et c2) revient à maximiser

$$\sum_{j=1}^{n} \log\{p_j\} + \lambda_1 \sum_{j=1}^{n} p_j + \lambda_2 \sum_{j=1}^{n} p_j(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)$$

où λ_1 et λ_2 sont des multiplicateurs de Lagrange et $p_j = p_j(x)$. En dérivant par rapport à p_j , on obtient,

$$\frac{1}{p_j} + \lambda_1 + \lambda_2(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x) = 0\ ;$$

Ce qui donne

$$p_j = -[\lambda_1(1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x))]^{-1},$$

avec $\lambda = \frac{\lambda_2}{\lambda_1}$.

La contrainte unité sur le total des poids conduit à

$$p_j = \frac{\left[1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)\right]^{-1}}{\sum_{j=1}^{n}\left[1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)\right]^{-1}}.$$

En substituant dans l'équation contrainte c2), on obtient l'équation

$$\sum_{j=1}^{n} \frac{(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)}{1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)} = 0,$$

équivalente à

$$n - \sum_{j=1}^{n} \frac{1}{1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)} = 0\ ;$$

Par suite,

$$p_j = \frac{1}{n}\left(\frac{1}{1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)}\right). \tag{2.14}$$

Le paramètre auxiliaire λ dépend à la fois des observations et de x ; Il maximise la fonction

$$\mathrm{L}(\lambda) = -\sum_{j=1}^{n} \log\{1 + \lambda(\mathrm{X}_j - x)\mathrm{K}_h(\mathrm{X}_j - x)\}.$$

$\mathrm{L}(\lambda)$ est non linéaire en λ, on utilise un algorithme de programmation non linéaire, comme celui de Newton-Raphson, pour la détermination du maximum.

L'estimateur `WNW` a de meilleures propriétés que la méthode linéaire locale *i.e.,* un biais plus faible et absence d'effet de bords. Ainsi, l'estimateur `WNW` préserve les qualités de l'estimateur `NW` et l'améliore.

L'estimateur du quantile conditionnel est, donc donné par

$$\hat{q}_{\mathrm{Y}}(\tau|x) = \hat{\mathrm{F}}^{-1}(\tau|x).$$

Les propriétés de cet estimateur sont étudiées dans CAI [2002]. Nous en reprenons, les résultats essentiels. On se place, sous les hypothèses (a1), (a2),(a3) et (a4).

A l'intérieur des bornes, le biais asymptotique est

$$Biais(\hat{q}_\tau(x)) = -\frac{1}{2}h^2\mu_2(K)\frac{F^{20}(\tau;x)|x)}{f(q_\tau(x)|x)},?$$

alors que la variance asymptotique est

$$Var(\hat{q}_\tau(x)) = \frac{\tau(1-\tau)\mu_0(K^2)}{nhg(x)f^2(q_\tau(x)|x)}.$$

L'AMSE est donc donnée par

$$AMSE(\hat{q}_\tau(x)) = \frac{1}{4}h^4\left(\frac{\mu_2(K)F^{20}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right) + \frac{\tau(1-\tau)\mu_0(K^2)}{nhg(x)f^2(q_\tau(x)|x)}.$$

Par minimisation de AMSE(h), par rapport à h, on obtient la fenêtre optimale

$$h_{opt}(x) = \frac{\tau(1-\tau)\nu}{\mu_2^2(K)g(x)F^{02}(q_\tau(x)|x)^2}.$$

Soit $x = hc$ $(0 < c < 1)$ un point au bord gauche, g et K de support, respectivement $[0,1]$ et $[-1,1]$ respectivement, on pose

$$L_c(\lambda) = \int_{-1}^{c}\frac{uK(u)}{1-\lambda uK(u)}du.$$

On note λ_c la solution de $L_c(\lambda) = 0$.

Le biais asymptotique, aux bords,est

$$Biais(\hat{q}_\tau(x)) = -\frac{1}{2}h^2\frac{\beta_2(c)F^{02}(q_\tau(0^+)|0^+)}{2\beta_1(c)f(q_\tau(0^+)|0^+)};$$

La variance asymptotique, au bord,est

$$Var(\hat{q}_\tau(x)) = \frac{\tau(1-\tau)\beta_0(0)}{\beta_1^2(c)nhg(0^+)f^2(q_\tau(0^+)|0^+)},$$

avec

- $\beta_0(c) = \int_{-1}^{c}\frac{u^2K(u)}{1-\lambda_c uK(u)}du,$
- $\beta_j(c) = \int_{-1}^{c}\frac{K^j(u)}{(1-\lambda_c uK(u))^j}du, \quad j = 1,2.$

Cet estimateur n'est pas continu en y et donc non differentiable en y, c'est pour pallier à ce défaut que CAI et WANG [2008] proposent un estimateur à double noyau possèdant les qualités des deux estimateurs précédents (L'estimateur de Yu et Jones, à double noyau et l'estimateur WNW).

Estimateur à double noyaux de Cai et Wang

Afin de rassembler les bonnes propriétés (monotonie , continuité , différentiabilité , appartenance à $[0,1]$) et pas d'effets de bord), CAI et WANG [2008] proposent un estimateur de la fonction de densité conditionnelle $f(y|x)$, et un estimateur de la fonction de répartition conditionnelle $\mathrm{F}(y|x)$:

$$\hat{f}(y|x) = \sum_{j=1}^{n} \omega_j(x,h_1)\mathrm{Y}_j^*(y),$$

où $\omega_j(x,h_1)$ est donné par 2.13 , et $\mathrm{Y}_j^*(y) = \mathrm{W}_{h_2}(y-\mathrm{Y}_j)$.

$$\hat{\mathrm{F}}(y|x) = \int_{-\infty}^{y} \hat{f}_c(y|x)dy = \sum_{j=1}^{n} \omega_j(x,h_1)\mathrm{G}_{h_2}(y-\mathrm{Y}_j),$$

où G est la fonction de répartition d'un autre noyau L.
Cet estimateur est linéaire local, et puisqu'il contient deux noyaux, on l'appelle estimateur à double noyau linéaire local. On note que, si $p_j(x)$ dans (2.13) est constant pour tout j ou si $\lambda = 0$, alors $\hat{f}(y|x)$ est l'estimateur de type `NW` classique utilisé par SCAILLET [2005]. Il est clair que $\hat{f}(y|x)$ est une fonction de densité et alors $\hat{\mathrm{F}}(y|x)$ est une fonction de répartition continue et différentiable en y. L'estimateur du quantile conditionnel est, donc donné par $\hat{q}_\mathrm{Y}(\tau|x) = \hat{\mathrm{F}}^{-1}(\tau|x)$.

On s'intéresse maintenant aux propriétés de cet estimateur, données par CAI et WANG [2008]. Sous les conditionns a1),a2),a3),a4),b1),b2),b3), à l'intérieur des bornes, le biais asymptotique est

$$Biais(\hat{q}_\tau(x)) = \frac{1}{2}\{\mu_2(\mathrm{K})\frac{\mathrm{F}^{20}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_1^2 - \mu_2(\mathrm{W})\frac{\mathrm{F}^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_2^2\};$$

La variance asymptotique est

$$Var(\hat{q}_\tau(x)) = \frac{\mu_0(\mathrm{K}^2)}{nh_1 g(x) f^2(q_\tau(x)|x)}(\tau(1-\tau) + h_2 f(q_\tau(x)|x)\alpha(\mathrm{W})).$$

L'`AMSE` est, donc donnée par

$$\begin{aligned}\mathrm{AMSE}(\hat{q}_\tau(x)) &= \frac{1}{4}\{\mu_2(\mathrm{K})\frac{\mathrm{F}^{20}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_1^2 - \mu_2(\mathrm{W})\frac{\mathrm{F}^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}h_2^2\}^2 \\ &+ \frac{r(k)}{nh_1 g(x) f^2(q_\tau(x)|x)}(\tau(1-\tau) + h_2 f(q_\tau(x)|x)\alpha(\mathrm{W})).\end{aligned}$$

Pour $h_1 >> h_2$, on obtient la fenêtre optimale de h_1 en minimisant $\mathrm{AMSE}(h_1)$ selon h_1 :

$$h_{1,opt}(x) = \left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)\mathrm{F}^{02}(q_\tau(x)|x)^2}\right)^{\frac{1}{5}}$$

Maintenant si $x = hc$ $(0 < c < 1)$, un point au bord gauche, avec g et K ayant leur support dans $[0,1]$ et $[-1,1]$, le biais asymptotique est

$$Biais(\hat{q}_\tau(x)) = \frac{1}{2}\{\alpha_c(\mathrm{K})\frac{\mathrm{F}^{20}(q_\tau(0^+)|0^+)}{f(q_\tau(0^+)|0^+)}h_1^2 + \mu_2(\mathrm{W})\frac{\mathrm{F}^{02}(q_\tau(0^+)|0^+)}{f(q_\tau(0^+)|0^+)}h_2^2\};$$

La variance asymptotique est

$$Var(\hat{q}_\tau(x)) = \frac{\beta_c(\mathrm{K})}{nhg(0^+)f^2(q_\tau(0^+)|0^+)}(\tau(1-\tau) - h_2 f(q_\tau(0^+)|0^+)\alpha(\mathrm{W})).$$

Récapitulatif des propriétés

Dans les tableaux (2.1), (2.2), et (2.3), on résume les propriétés asymptotiques à l'intérieur des bornes pour les estimateurs précédents. Il est clair, au vu de ces tableaux, que les proporiétés des estimateurs `RLL`, `DYYJ`, `WNW`, `CW` sont équivalentes lorsque $h_2 << h_1$.

TABLEAU 2.1 – Le biais asymptotique des estimateurs

Estimateur	Biais
`RLC`	$\frac{1}{2}h^2\mu_2(\mathrm{K})\left\{-\frac{\mathrm{F}^{20}(q_\tau(x)\|x)}{f(q_\tau(x)\|x)}+2\frac{g'(x)q'_\tau(x)}{g(x)}\right\}$
`RLL`	$\frac{1}{2}h^2\mu_2(\mathrm{K})q"_\tau(x)$
`DNYJ`	$\frac{1}{2}\left\{\mu_2(\mathrm{K})\frac{\mathrm{F}^{20}(q_\tau(x)\|x)}{f(q_\tau(x)\|x)}h_1^2+\mu_2(\mathrm{W})\frac{\mathrm{F}^{02}(q_\tau(x)\|x)}{f(q_\tau(x)\|x)}h_2^2\right\}$
`WNW`	$-\frac{1}{2}h^2\mu_2(\mathrm{K})\frac{\mathrm{F}^{20}(q_\tau(x)\|x)}{f(q_\tau(x)\|x)}$
`CW`	$\frac{1}{2}\left\{\mu_2(\mathrm{K})\frac{\mathrm{F}^{20}(q_\tau(x)\|x)}{f(q_\tau(x)\|x)}h_1^2-\mu_2(\mathrm{W})\frac{\mathrm{F}^{02}(q_\tau(x)\|x)}{f(q_\tau(x)\|x)}h_2^2\right\}$

TABLEAU 2.2 – La variance asymptotique des estimateurs

Estimateur	Variance
`RLC`	$\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{nhg(x)f^2(q_\tau(x)\|x)}$
`RLL`	$\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{nhg(x)f^2(q_\tau(x)\|x)}$
`DNYJ`	$\frac{\mu_0(\mathrm{K}^2)}{nh_1g(x)f^2(q_\tau(x)\|x)}(\tau(1-\tau)-h_2f(q_\tau(x)\|x)\alpha(\mathrm{W}))$
`WNW`	$\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{nhg(x)f^2(q_\tau(x)\|x)}$
`CW`	$\frac{\mu_0(\mathrm{K}^2)}{nh_1g(x)f^2(q_\tau(x)\|x)}(\tau(1-\tau)+h_2f(q_\tau(x)\|x)\alpha(\mathrm{W}))$

TABLEAU 2.3 – La fenêtre optimale des estimateurs

Estimateur	Fenêtre optimale
`RLC`	$h_{opt}=\left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)f(q_\tau(x)\|x)^2\mathrm{B}_\tau(x)}\right)^{\frac{1}{5}}$
`RLL`	$h_{opt}=\left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)q''_\tau(x)^2f(q_\tau(x)\|x)^2}\right)^{\frac{1}{5}}$
`DNYJ`	$h_{1,opt}=\left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)\mathrm{F}^{02}(q_\tau(x)\|x)^2}\right)^{\frac{1}{5}}$
`WNW`	$h_{opt}=\left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)\mathrm{F}^{02}(q_\tau(x)\|x)^2}\right)^{\frac{1}{5}}$
`CW`	$h_{1,opt}=\left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})g(x)\mathrm{F}^{02}(q_\tau(x)\|x)^2}\right)^{\frac{1}{5}}$

2.7 Forme unifiée

Tous les estimateurs présentés, auparavant, dépendent de la solution

$$\arg\min_{\beta\in\mathbb{R}^p}\sum_{j=1}^{n}\rho(\Omega(y,Y_j)-q_j(\beta))\omega_j,$$

pour des choix particuliers de ρ, Ω, β, q_j et ω_j.

1) Si ρ est la fonction perte (ou coût) introduite par Koenker et Basset *i.e.*, $\rho(u)=\rho_\tau(u)=u(\tau-\mathbb{1}(u<0))$, $\Omega(y,Y_j)=Y_j$, $q_j(\beta)=\sum_{j=0}^{p}(X_j-x)^j\beta_j(x)$ et $\omega_j=K(\frac{X_j-x}{h})$ où K est un noyau, on a alors la régression quantile polynomiale.

2) Si ρ est la perte quadratique $\rho(z)=z^2$, $\Omega(y,Y_j)=G(\frac{y-Y_j}{h_2})$, $q_j(\beta)=\beta_0+\beta_1(X_j-x)$ et $\omega_j=K(\frac{X_j-x}{h_1})$ où K est un noyau et G la fonction de répartition d'un autre noyau, on a alors l'estimateur à double noyau de Yu et Jones.

3) Si ρ est la fonction perte quadratique, $\Omega(y,Y_j)=\mathbb{1}(Y_j\leq y)$, $q_j(\beta)=\beta_0$ et $\omega_j=p_j(x)K_h(X_j-x)$ où $p_j(x)$ est définie par (2.14), on a alors l'estimateur de WNW.

4) Si ρ est la perte quadratique, $\Omega(y,Y_j)=G(\frac{y-Y_j}{h_2})$, $q_j(\beta)=\beta_0$ et $\omega_j=p_j(x)K_h(X_j-x)$ avec $p_j(x)$ défini par (2.14), et G est la fonction de répartition d'un autre noyau, on a alors l'estimateur à double noyau de Cai et Wang.

2.8 Choix de la fenêtre

Dans cette section, nous présentons quatre méthodes de détermination de la fenêtre optimale : la Règle *rule of thumb*, une méthode *plug-in* itérative, la méthode de validation croisée et le critère Akaiké non-paramétrique.

2.8.1 Règle *rule of thumb*

YU et JONES [1998] proposent cette heuristique pour estimer les fenêtres, dans les directions x et y, de l'estimateur à double noyau et de la régression linéaire locale.

On choisit, comme fenêtre h (ou h_1) selon la direction x,

$$h=h_{\text{moy}}\left(\frac{\tau(1-\tau)}{\phi(\Phi^{-1}(\tau))^2}\right)^{\frac{1}{5}}$$

où h_{moy} est la fenêtre obtenue, dans la régression à noyau de Y sur X. Les fonctions ϕ et Φ sont respectivement la densité et la fonction de répartition de la loi normale standard.

Pour les estimateurs à double noyau, on choisit comme fenêtre h_2 selon la direction y,

$$h_2=\begin{cases} max\left(\frac{h_{1,\frac{1}{2}}^5}{h_1^3},\frac{h_1}{10}\right), & \text{si } h_{1,\frac{1}{2}}<1\,;\\ \frac{h_{1,\frac{1}{2}}^4}{h_1^3}, & \text{sinon.}\end{cases}$$

avec $h_{1,\frac{1}{2}}=\left(\frac{\pi}{2}\right)^{\frac{1}{5}}h_{\text{moy}}$.

2.8.2 Méthode *plug-in* itérative

ATTAR [2008] propose un algorithme itératif pour estimer la fenêtre optimale de la régression quantile linéaire locale et un autre algorithme itératif pour estimer la fenêtre optimale de l'estimateur à double noyau de Yu et Jones.
Ces deux algorithmes se basent sur une même demarche ; Le principe de cette démarche est de remplacer les quantités inconnues dans l'expression de la fenêtre globale par des estimateurs appropriés.
L'idée est similaire à celle proposée par GASSER et *al.* [1991], dans la contexte de la régression à noyau. La fenêtre optimale globale de la régression quantile linéaire locale minimise l'erreur quadratique moyenne intégrée asymptotique

$$\mathtt{AMISE}(\hat{q}_\tau(x)) = \int \mathrm{AMSE}(\hat{q}_\tau(x))\omega(x)dx$$

où ω est une fonction de poids positive. De (2.7), on a l'expression de $\mathrm{AMSE}(\hat{q}_\tau(x))$, on la remplace puis on dérive par rapport h pour obtenir la fenêtre optimale globale

$$h_{opt} = \left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})} \frac{\int \frac{\omega(x)}{g(x)f^2(q_\tau(x)|x)}dx}{\int q''_\tau(x)\omega(x)dx} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

On prenant $\omega(x) = g(x)$, on obtient

$$h_{opt} = \left(\frac{\tau(1-\tau)\mu_0(\mathrm{K}^2)}{\mu_2^2(\mathrm{K})} \frac{\int \frac{1}{f^2(q_\tau(x)|x)}dx}{\int q''_\tau(x)g(x)dx} \right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{2.15}$$

On estime les quantités inconnues $q_\tau(x)$, $q''_\tau(x)$ et $f(q_\tau(x)$ à l'aide d'une procédure itérative avec des estimateurs appropriés.
On estime $q_\tau(x)$ par la régression quantile linéaire locale *i.e.*, $\hat{q}_\tau(x) = \hat{\beta}_0$ où $\widehat{\beta}_0$ resulte du problème d'optimisation

$$\min_{\beta\in\mathbb{R}^2} \sum_{i=1}^{n} \rho_\tau(\mathrm{Y}_i - \beta_0 - \beta_1(\mathrm{X}_i - x))\mathrm{K}(\frac{\mathrm{X}_i - x}{h}). \tag{2.16}$$

Pour estimer $q''_\tau(x)$, on utilise la régression quantile cubique locale comme solution du problème

$$\min_{\beta\in\mathbb{R}^4} \sum_{i=1}^{n} \rho_\tau(\mathrm{Y}_i - \beta_0 - \beta_1(\mathrm{X}_i - x) - \beta_2(\mathrm{X}_i - x)^2 - \beta_3(\mathrm{X}_i - x)^3)\mathrm{K}_2(\frac{\mathrm{X}_i - x}{h}). \tag{2.17}$$

où K_2 est un noyau. On estime donc $q''_\tau(x)$ par $2\hat{\beta}_2$.

Pour estimer $f(y|x)$, on utilise l'estimateur de FAN et *al.* [1996] :

$$\hat{f}(y|x) = \frac{\sum \mathrm{K}_1(\frac{\mathrm{X}_i - x}{h'_1})\mathrm{W}_1(\frac{\mathrm{Y}_i - y}{h'_2})}{h'_2 \sum \mathrm{K}_1(\frac{\mathrm{X}_i - x}{h'_1})} \tag{2.18}$$

où K_1 et W_1 sont des noyaux. Ils proposent la règle *rule of thumb* pour choisir h_2 :

$$h_2 = c s_\mathrm{Y} n^{\frac{-1}{5}}$$

où $c = 1.06$ pour le noyau gaussien, $c = 2.34$ pour le noyau Epanechnikov, et s_Y est l'écart type de Y.

Les intégrales intervenant dans la formule de h_{opt} sont estimées par

$$\begin{aligned} \mathrm{I} &= \int \frac{1}{f(q_\tau(x)|x)^2 dx} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{\hat{f}(\hat{q}_\tau(x_i)|x_i)^2}, \\ \mathrm{J} &= \int q''(x)^2 dx = \frac{1}{m}\sum_{i=1}^{m} \hat{q}''(x_i)^2. \end{aligned}$$

Ainsi, l'algorithme *plug-in* itératif pour estimer h_{opt} est le suivant :

1) On choisit comme fenêtre initiale $h^{(0)} = \frac{1}{n}$;
2) Pour chaque point x_i (i=1,...,n),
   - On calcule l'estimation $\hat{q}_\tau(x_i) = \hat{\beta_0}(x_i)$ où $\hat{\beta}_0$ est la solution du problème (2.16) avec comme fenêtre $h^{(0)}$,
   - on calcule l'estimation $\hat{f}(\hat{q}_\tau(x_i)|x_i)$, en utilisant l'estimateur (2.18), au point $y = \hat{q}_\tau(x_i)$,
   - on calcule l'estimation $\hat{q}''_\tau(x_i) = 2\hat{\beta}_2(x_i)$ où $\hat{\beta}_2$ est la solution de (2.17), avec comme fenêtre $h^{(0)} * n^{\frac{1}{10}}$;
3) On calcule

$$\begin{aligned} \mathrm{I}^{(1)} &= \frac{1}{m}\sum_{i=1}^{m} \frac{1}{\hat{f}(\hat{q}_\tau(x_i)|x_i)^2}, \\ \mathrm{J}^{(1)} &= \frac{1}{m}\sum_{i=1}^{m} \hat{q}''(x_i)^2 = \frac{4}{m}\sum_{i=1}^{m} \hat{\beta}_2^2. \end{aligned}$$

4) On remplace $h^{(0)}$ par

$$h^{(1)} = \left(\frac{\mu_0(\mathrm{K}^2)\tau(1-\tau)\mathrm{I}^{(1)}}{\mu_2^2(\mathrm{K})\mathrm{J}^{(1)}} \right)^{\frac{1}{5}} n^{\frac{-1}{5}};$$

5) On répète les étapes 2), 3), 4) et 5) jusqu'à ce que la différence, entre deux fenêtres successives, soit négligeable.

2.8.3 Validation croisée

Validation croisée classique

Dans le contexte de la régression quantile, Koenker et al.(1992) proposent le modèle

$$\mathrm{Y}_i = q_\tau(\mathrm{X}_i) + \mathrm{U}_i, \text{ avec } \mathrm{P}(\mathrm{U}_i < 0) = \tau,$$

où $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$.

ABBERGER [1998] adapte la validation croisée *leave-one-out* à la régression quantile, pour des données *iid*, en utilisant le critère

$$\mathrm{CV}(h) = \sum_{j=1}^{n} \rho(\mathrm{Y}_j - \hat{q}_\tau^{(-j)}(\mathrm{X}_j))\omega(\mathrm{X}_j),$$

où $\hat{q}^{(-j)}(.)$ est l'estimateur du quantile conditionnel, basé sur toutes les observations sauf celle d'ordre j.

La fenêtre optimale est obtenue par minimisation de $\mathrm{CV}(h)$ selon h et $q_\tau(x)$ par $\hat{\mathrm{F}}^{(-1)}(\tau|x)$, avec

$$\hat{\mathrm{F}}(y|x) = \frac{\sum_{i=1}^{n} \mathbb{1}(\mathrm{Y}_i \leq y)\mathrm{K}(\frac{\mathrm{X}_i - x}{h})}{\sum_{i=1}^{n} \mathrm{K}(\frac{\mathrm{X}_i - x}{h})},$$

où K est un noyau ; On peut aussi utiliser cette méthode pour estimer la fenêtre optimale dans la direction x, pour les estimateurs précédents.

Pour les estimateurs à double noyau, la fenêtre h_2 (dans la direction y) n'a pas une grande influence sur le biais et la variance, selon Fan et *al.* [1996]. Bien que coûteuse en temps, nous retenons leur règle *i.e.,* la règle *rule of thumb* :

$$h_2 = \frac{\mu_0(\mathrm{W}^2)}{\mu_2(\mathrm{W})} \mathrm{S}_\mathrm{Y} n^{\frac{-1}{5}},$$

où S_Y est l'écart-type de Y.

Validation croisée généralisée

On peut utiliser le critère de la validation croisée généralisée proposée par Yao et Tong [1998] dans le contexte de la régression à noyau (déjà expliquée dans la section (1.3.4)), en l'adaptant à la régression quantile. Pour cela, on partage les données $\{(\mathrm{X}_j, \mathrm{Y}_j),\ 1 \leq j \leq n\}$ en deux sous-échantillons : $\{(\mathrm{X}_j, \mathrm{Y}_j),\ 1 \leq j \leq k\}$ et $\{(\mathrm{X}_j, \mathrm{Y}_j),\ k+1 \leq j \leq n\}$.

On utilise le premier sous-échantillon pour construire l'estimateur $\hat{q}_{k,h}$, et le second pour spécifier le critère de la validation croisée ; Par la suite, on cherche h minimisant ce critère *i.e.,*

$$h_k = \arg\min_{h \in \mathcal{H}_k} \frac{1}{n-k} \sum_{j=k+1}^{n} \left\{ \rho_\tau \left[\mathrm{Y}_j - \hat{q}_{k,h}(\mathrm{X}_j) \right] \omega(\mathrm{X}_j) \right\}. \tag{2.19}$$

Puisque la fenêtre optimale h de l'estimateur $\hat{q}_\tau(.)$ est d'ordre $n^{-\frac{1}{5}}$, on peut poser $h = h_k(\frac{k}{n})^{\frac{1}{5}}$.

2.8.4 Critère de CAI : Extension non-paramétrique du critère d'Akaike

Cai et Tiwari [2000] introduit un critère d'information, à l'instar du critère AIC d'Akaike ou du citère BIC, adapté au cas non pramétrique : h minimise

$$\mathtt{CAI}(h) = log\{\mathtt{MASE}\} + \psi(\mathrm{tr}(\mathrm{H}), n), \tag{2.20}$$

avec

- $\mathtt{MASE}(h) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathrm{Y}_i - q(\mathrm{Y}_i|\mathrm{X}_i))$,
- $\psi(x,n) = 2\left(\frac{x+1}{n-x-2}\right)$.

Dans ce qui suit, $\mathbf{H}$ est une matrice carrée d'ordre n telle que $\hat{\mathrm{Z}} = (\hat{\mathrm{F}}(y|\mathrm{X}_1), \ldots, \hat{\mathrm{F}}(y|\mathrm{X})) = \mathbf{H}(h)\mathrm{Z}$, avec $tr(\mathbf{H}) = \sum_{i=1}^{n} \frac{p_i(\mathrm{X}_i)\mathrm{W}(0)}{\sum_{j=1}^{n} p_i(\mathrm{X}_j)\mathrm{W}((\mathrm{X}_i - \mathrm{X}_j)/h)}$; Pour l'estimateur de WNW,

$$\mathrm{Z}^t = (\mathbb{I}(\mathrm{Y}_1 \leq y), \ldots, \mathbb{I}(\mathrm{Y}_n \leq y)).$$

Tandis que pour l'estimateur de Cai, où h_2 est défini comme précédemment (par la règle *rule of thumb*), on minimise le critère AIC(h) dans (2.20),

$$Z^t = (G(\frac{y - Y_1}{h_2}), \ldots, G(\frac{y - Y_n}{h_2})).$$

2.9 Amélioration de l'estimateur non-paramétrique du quantile conditionnel par Réseaux de neurones

Les estimateurs non-paramétriques de quantiles conditionnels ont le défaut d'une variance élevée, quand le paramètre de lissage est petit. C'est pourquoi nous proposons (KNEFATI et *al.*) une méthode consistant à utiliser un réseau de neurones à fonction radiale de base, pour améliorer la qualité de l'estimateur choisi. Rappelons, la définition du modèle réseaux de neurones à fonction radiale de base (RBF).
Definition : Une fonction radiale $\phi_{\mathbf{c}}$ ($\mathbb{R}^p \mapsto \mathbb{R}^+$) est une fonction dépendant d'un centre $\mathbf{c}$, d'une norme $\|.\|$ et d'une fonction de base ψ ($\mathbb{R}^+ \mapsto \mathbb{R}^+$) telle que $\phi_{\mathbf{c}}(\mathbf{x}) = \psi(\|\mathbf{x} - \mathbf{c}\|)$.
La norme est en général la norme euclidienne ; À titre d 'exemple de fonction ψ,

- la fonction gaussienne : $\psi(t) = e^{-(\epsilon t)^2}$,
- la fonction quadratique : $\psi(t) = \sqrt{1 + (\epsilon t)^2}$.

2.9.1 Le modèle du RBF

Soit $\mathbf{x}_i = (x_i^{(1)}, \ldots, x_i^{(p)})$, l'entrée du réseau, et $y_i = m(\mathbf{x}_i)$ la sortie de réseau, $1 \leq i \leq n$, m est une fonction réelle inconnue à approcher. L'idée des réseaux de neurones est de chercher un estimateur $\hat{m}$ de m tel que $\hat{m}(\mathbf{x}_i) = y_i$.

Un RBF s'écrit

$$\hat{m}(x) = \sum_{j=1}^{k} \omega_j \psi(\|x - c_j\|), \quad x \in \mathbb{R}^p \tag{2.21}$$

où

- k est le nombre de neurones ;
- $\mathbf{c}_j$ est le centre de la $j^{\text{ème}}$ neurone ;
- $(\omega_j)_{j=1,\ldots,k}$ sont les poids ;
- ψ est une fonction de base.

Ici, on s'intéresse au cas unidimensionnel ($p = 1$). Dans ce cas, l'estimation est

$$\hat{m}(x) = \sum_{j=1}^{k} \omega_j \psi(|x - c_j|), \quad x \in \mathbb{R}. \tag{2.22}$$

Cette estimation est entièrement déterminée, dés lors que sont calculés, les centres c_j ($1 \leq j \leq k$) et les poids ω_i ($1 \leq i \leq n$). Pour ajuster le modèle aux entrées $\mathbf{x}_i$, il faut que

$$y_i = \hat{m}(\mathbf{x}_i) = \sum_{j=1}^{k} \omega_j \psi(|x_i - c_j|), \quad i = 1, \ldots, n\,;$$

Cela revient à résoudre le système linéaire (en les ω_i) à n équations et k inconnues :

$$\mathbf{G}_c \mathbf{W} = \mathbf{Y}$$

où $\mathbf{G}_c$ est une matrice de dimension $n \times k$ d'éléments $\psi(|x_i - c_j|)$, $\mathbf{W}$ est le vecteur des poids ω_i de taille n, et $\mathbf{Y}$ est le vecteur des sorties y_i. Le calcul des centres et poids revient à résoudre le problème d'optimisation

$$\min_{\mathbf{W}\in\mathbb{R}^n, \mathbf{c}\in\mathbb{R}^k} \|\mathbf{Y} - \mathbf{G}_c \mathrm{W}\|^2. \tag{2.23}$$

2.9.2 Algorithme CCENTER : Détermination des centres et du nombre de neurones

Pour calculer les centres, nous avons implémenté sous le logiciel R, un algorithme comparable à la méthode robuste proposée dans le *package* de Matlab (*Neural Networks Toolbox*). L'algorithme que nous utilisons, pour estimer les centres et les poids, est le suivant :

0. **En entrée de l'algorithme** :
   Les observations $\{(x_i, y_i),\ \ i = 1, \ldots, n\}$,
   le seuil d'erreur $Seuil.Err$ à atteindre,
   le nombre maximal de neurones $K.Max$,
   $k = 0$.
1. Stocker $(y_1, ..., y_n)$ dans un autre vecteur $(y_1^*, ..., y_n^*)$.
2. Tant que l'erreur totale $Err = \frac{1}{n}\sum_{i=1}^n (\hat{m}(x_i) - y_i^*)^2$ est supérieure à $Seuil.Err$ et $k <$ $K.Max$,
   a. calculer la sortie du réseau (2.22) et l'erreur $r_i = |\hat{m}(x_i) - y_i^*|$, pour chaque entrée ;
   b. Trouver l'entrée d'erreur maximale *i.e.*, $\ell = \arg\max_{i\in\{1,...,n\}} r_i$;
   c. Ajouter une fonction radiale ayant pour centre $c_k = x_\ell$ et $k \leftarrow k + 1$;
   d. Recalculer le vecteur de poids (W) (cf. paragraphe 2.9.3) ;
   e. Recalculer $\hat{m}(x_i) = \sum_{j=1}^k \omega_j \psi(|x_i - c_j|)$;
   f. Mettre $y_i^* = 0$ et $\hat{m}(x_\ell) = 0$.
3. **En sortie** : Le nombre final de neurones k et les centres $c_j,\ \ j = 1, \ldots, k$.

2.9.3 Calcul des poids

Une fois les centres calculés, la matrice $\mathbf{G}_c = \mathbf{G}$ est complètement déterminée. Si la matrice carrée $\mathbf{G}$ est de rang k *i.e.*, $\mathbf{G}^t\mathbf{G}$ inversible, alors la solution du problème d'optimisation (2.23) nous donne le vecteur de poids $\mathbf{W} = (\mathbf{G}^t\mathbf{G})^{-1}\mathbf{G}^t\mathbf{Y}$.

Si la matrice $\mathbf{G}^t\mathbf{G}$ est irrégulière, on peut ajouter le terme de régularisation $\|\boldsymbol{\Gamma}\mathbf{W}\|$ au problème (2.23), où $\boldsymbol{\Gamma}$ est en général de la forme $\lambda \mathrm{I}_n$ avec I_n la matrice Identité de taille n et $\lambda \in \mathbb{R}$ à fixer.
On a alors, dans ce cas, le problème d'optimisation suivant :

$$\min_{\mathbf{W}\in\mathbb{R}^n, c\in\mathbb{R}^k} \left(\|\mathbf{Y} - \mathbf{G}\mathbf{W}\|^2 + \|\boldsymbol{\Gamma}\mathbf{W}\|^2\right),$$

où le vecteur de poids $\mathbf{W}$ est donné par $\mathbf{W} = (\mathbf{G}^t\mathbf{G} + \boldsymbol{\Gamma}^t\boldsymbol{\Gamma})^{-1}\mathbf{G}^t\mathbf{Y}$.

2.9.4 Algorithme CQRBF : Algorithme RBF d'estimation du quantile conditionnel

On présente, dans ce qui suit, notre algorithme RBF pour estimer le quantile conditionnel. On se place, dans le cadre du modèle $Y = m(x) + \epsilon$, où ϵ est le terme d'erreur de fonction de répartition F_ϵ connue.

0. **En entrée de l'algorithme** :
   Les observations $\{(X_i, Y_i),\ \ i = 1, \ldots, n\}$,
   τ : le niveau du quantile,
   x : une valeur du support de l'unique covariable X,

1. On calcule $\widehat{q}_\tau(x)$ par l'estimateur de YU et JONES [1998] et donc $\widehat{q}_\tau(x) = \hat{q}_\tau^{YJ}(x)$;
2. On calcule $\hat{m}^{YJ}(X_i) = \hat{q}_\tau^{YJ}(X_i) - F_\epsilon^{-1}(\tau),\ \ 1 \leq i \leq n$.
3. On utilise la méthode RBF (expliquée en section 2.9.1 avec en entrée $X_i,\ \ 1 \leq i \leq n$ et $y_i = \hat{m}^{YJ}(X_i)$. Ainsi, $\widehat{m}(x) = \hat{m}^{RBF}(x)$.
4. **En sortie** : L'estimation du quantile conditionnel *i.e.,*

$$\hat{q}_\tau(x) = \hat{m}^{RBF}(x) + F_\epsilon^{-1}(\tau). \tag{2.24}$$

2.10 Expériences numériques

2.10.1 Données simulées

Pour chaque modèle, on simule 100 échantillons de taille $n = 250$, $n = 500$ et $n = 1000$. Pour chaque échantillon, on calcule le quantile conditionnel, au niveau τ =0.01, 0.1, 0.5, 0.9, 0.99, au moyen des cinq estimateurs présentés : La régression constante locale (RQC), la régression linéaire locale (RQL), l'estimateur à double noyau de Yu et Jones (YJ), l'estimateur de Nadaraya-Watson (WNW) et l'estimateur à double noyau de Cai et Wang (CW).

La qualité de ces estimateurs est mesurée par la moyenne des valeurs absolues des erreurs *i.e.,*

$$\text{MADE} = \frac{1}{n^*} \sum_{i=1}^{n^*} |\hat{q}_\tau(x_i) - q_\tau(x_i)|.$$

où $x_i,\ \ i = 1, \ldots, n^*$ $(n^* = 40)$ sont des points sur l'intervalle de X. L'estimation des fenêtres est réalisée au moyen de la règle *rule of thumb* ; Les noyaux utilisés sont gaussiens.

Exemple 1 : Modèle ARCH
On considère, comme dans le travail de CAI et WANG [2008], le modèle ARCH

$$Y_i = 0.9 \sin(2.5 X_i) + \sigma(X_i)\epsilon_i,$$

avec $X_i = Y_{i-1}$ où $\sigma(x) = 0.8\sqrt{1.2 + x^2}$ et les erreurs $\epsilon_i,\ \ i = 1, \ldots, n$ *iid* de loi $\mathcal{N}(0,1)$. On estime le quantile conditionnel sur l'intervalle $[-2, 2]$.

Les figures 2.2, 2.3, 2.4, 2.5, et 2.6 restituent les résultats de la simulation, pour les cinq niveaux choisis, du quantile conditionnel.

les figures 2.7, 2.8 et 2.9 représentent les boxplots des erreurs MADE pour chacun des estimateurs.

Le tableau 2.4 donne la médiane de ces erreurs. Il est clair d'après ce tableau et les figures que pour τ loin de 0 et de 1, il n'y a pas de différence significative entre les cinq estimateurs.

Lorsque τ est au voisinage de 0 ou de 1 (Voir par exemple les cas $\tau = 0.01$ et $\tau = 0.99$), l'estimateur RQL est souvent, le plus efficace à l'intérieur du domaine de X, tandis que les estimateurs YJ, WNW et CW, le sont aux bords.

Exemple 2 : Modèle de Weibull
Le modèle est

$$Y_i = 2 + X_i + 2cos(X_i) + \epsilon_i,$$

avec X_i de Weibull, de paramètres $(\lambda, k) = (1, 1.5)$ et ϵ_i de loi exponentielle de moyenne 1. Le vrai quantile pour ce modèle est la fonction $q_\tau(x) = 2 + x + 2\cos(x) + \zeta^{-1}(\tau)$, avec $\zeta(.)$ la fonction de répartition de la loi exponentielle de moyenne 1. On estime le quantile conditionnel sur l'intervalle $[0,\ 2.5]$.

Des simulations analogues à celles réalisées, pour le modèle ARCH ont été réalisées. Les figures 2.10, 2.11, 2.12, 2.13,2.15, 2.16 , 2.17 et 2.14 et le tableau refmederrweib en restituent les résultats.
On aboutit aux mêmes conclusions que précédemment.

2.10.2 Données *maturation cérébrale*

Dans KNEFATI et *al.* [2014] nous nous intéressons aux facteurs pouvant influencer la maturation célébrale anormale chez les prématurés. Les données consistent en 416 enregistrements (Extrait de données PMSI, 2003 et 2004).
La variable explicative est l'age gestationnel (en jours) du nouveau né. La variable à expliquer est la durée totale en pourcentage de la longueur de l'intervalle IBI (intervalle temporel sur lequel on mesure la maturation célébrale chez le nouveau né), cette durée est mesurée en secondes.

L'objectif est de détecter les données anormales en traçant les deux courbes de référence $(x, \widehat{q}_\tau(x))$ et $(x, \widehat{q}_{1-\tau}(x))$. Les observations hors de la région comprise entre ces deux courbes, sont hors norme. Ici, nous utilisons les estimateurs YJ et QRBF de $q_\tau(x)$, avec $\tau = 0.1$ (cf. 2.18).

2.11 Conclusion

L'estimation du quantile conditionnel est basée, soit sur des méthodes indirectes reposant sur l'inversion numérique d'estimateurs de la fonction de répartition conditionnelle, soit sur des méthodes directes.

Parmi ces méthodes, la régression quantile linéaire locale est la plus simple à implémenter, et souvent la plus efficace, comme vu au travers des simulations ; Il arrive cependant,

que cet estimateur soit le moins efficace aux bords, quand τ est proche de 0 ou de 1.

Aussi, il y a plusieurs méthodes pour choisir les paramètres de lissage, utilisés pour ces estimateurs : la règle *rule of thumb*, la méthode *plug-in* itérative, la validation croisée (avec ses différents critères) et la méthode Akaiké non-paramétrique. Parmi ces méthodes, la règle de Yu et Jones est la plus utilisée, car simple à implémenter et fournissant des résultats satisfaisants.

Ainsi, pour estimer le qunatile conditionnel, nous recommandons d'utiliser la régression quantile linéaire locale, couplée avec la règle *rule of thumb*, pour choisir la fenêtre.

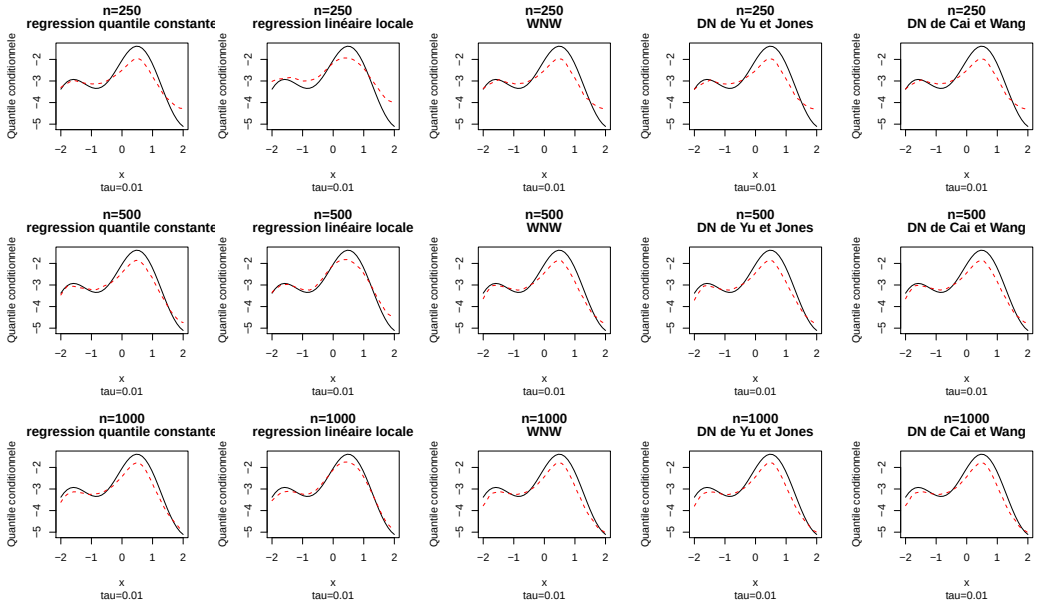


FIGURE 2.2 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.01$, pour les données simulées selon le modèle ARCH

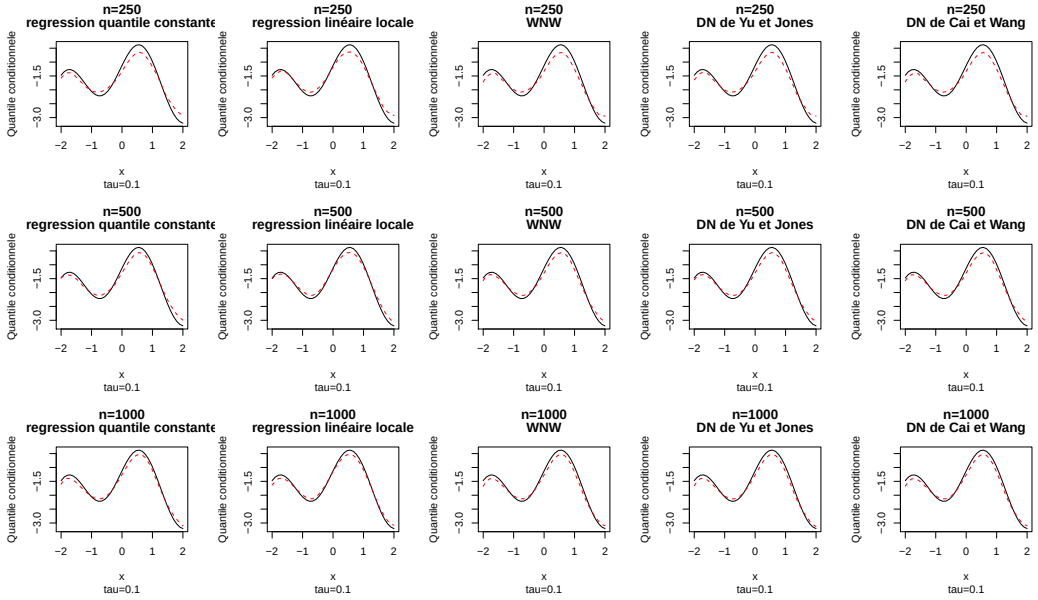


FIGURE 2.3 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.1$, pour les données simulées selon le modèle ARCH.

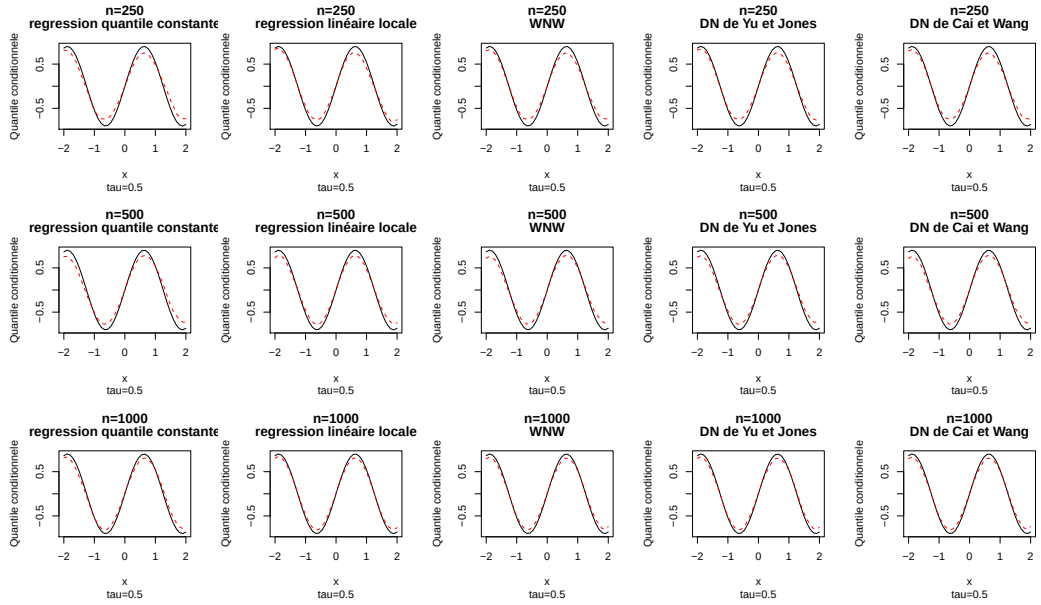


FIGURE 2.4 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.5$, pour les données simulées selon le modèle ARCH.

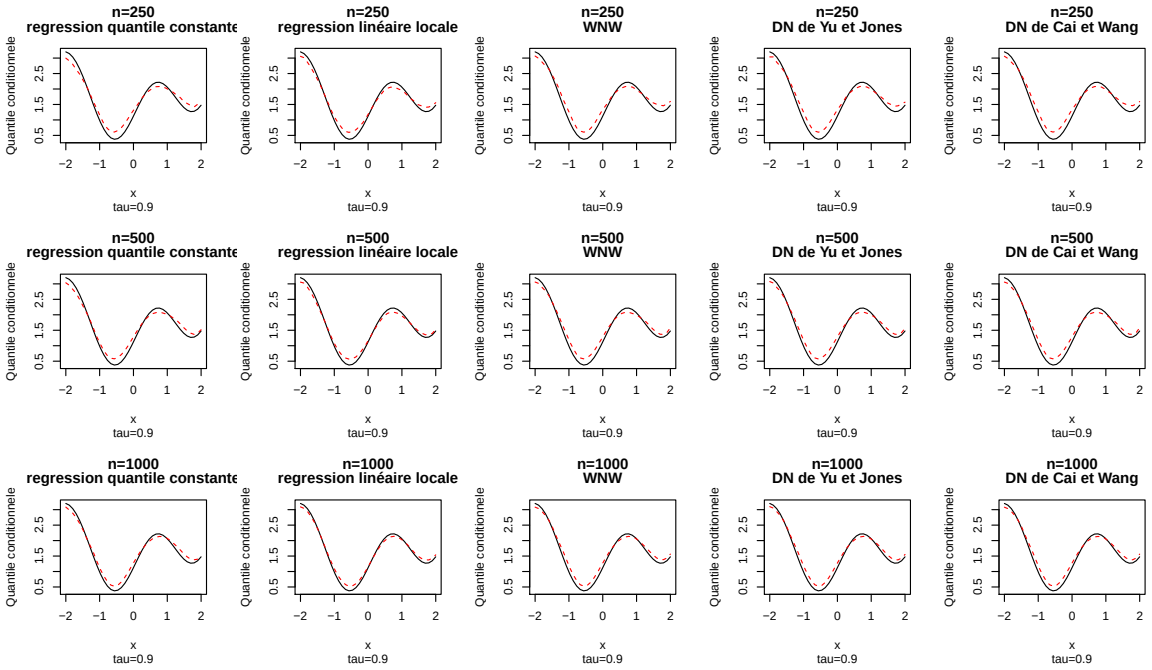


FIGURE 2.5 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.9$, pour les données simulées selon le modèle ARCH

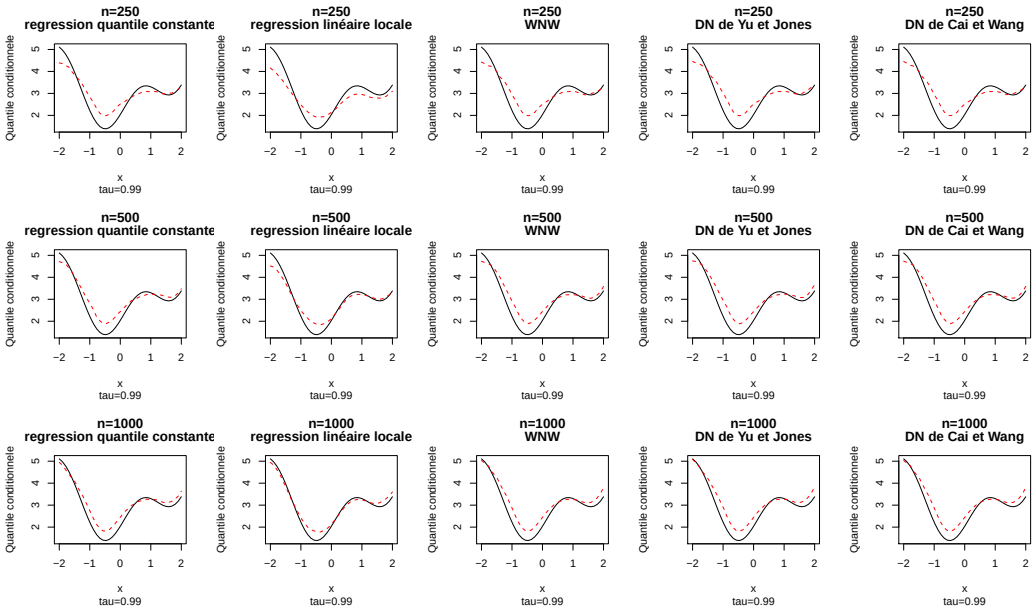


FIGURE 2.6 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.99$, pour les données simulées selon le modèle ARCH.

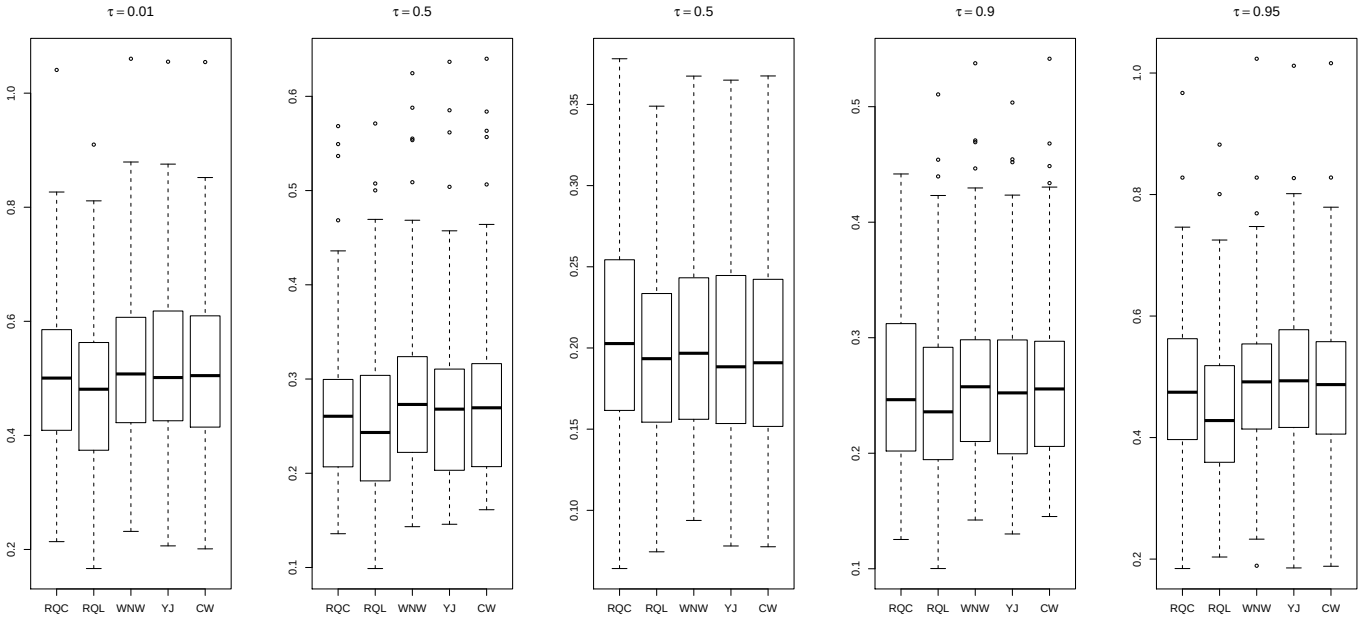


FIGURE 2.7 – Boxplots des erreurs MADE des cinq estimateurs, pour les données simulées selon le modèle ARCH, avec n=250.

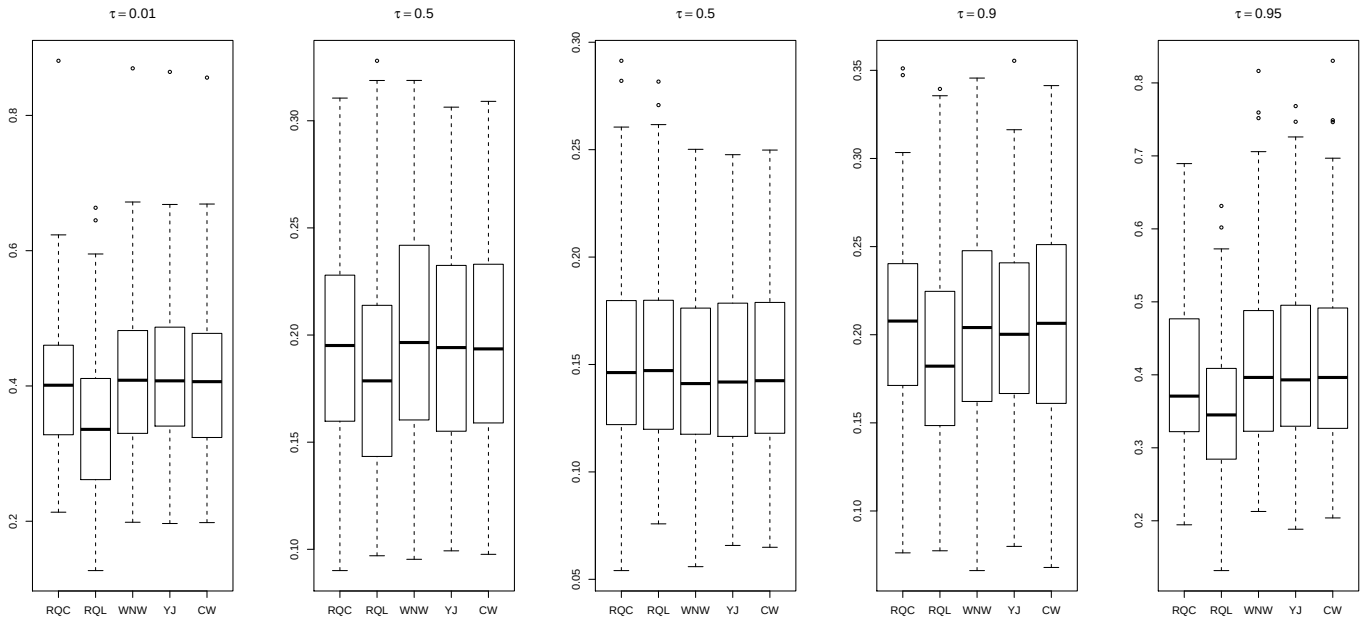


FIGURE 2.8 – Boxplots des erreurs MADE des cinq estimateurs, pour les données simulées selon le modèle ARCH, avec n=500.

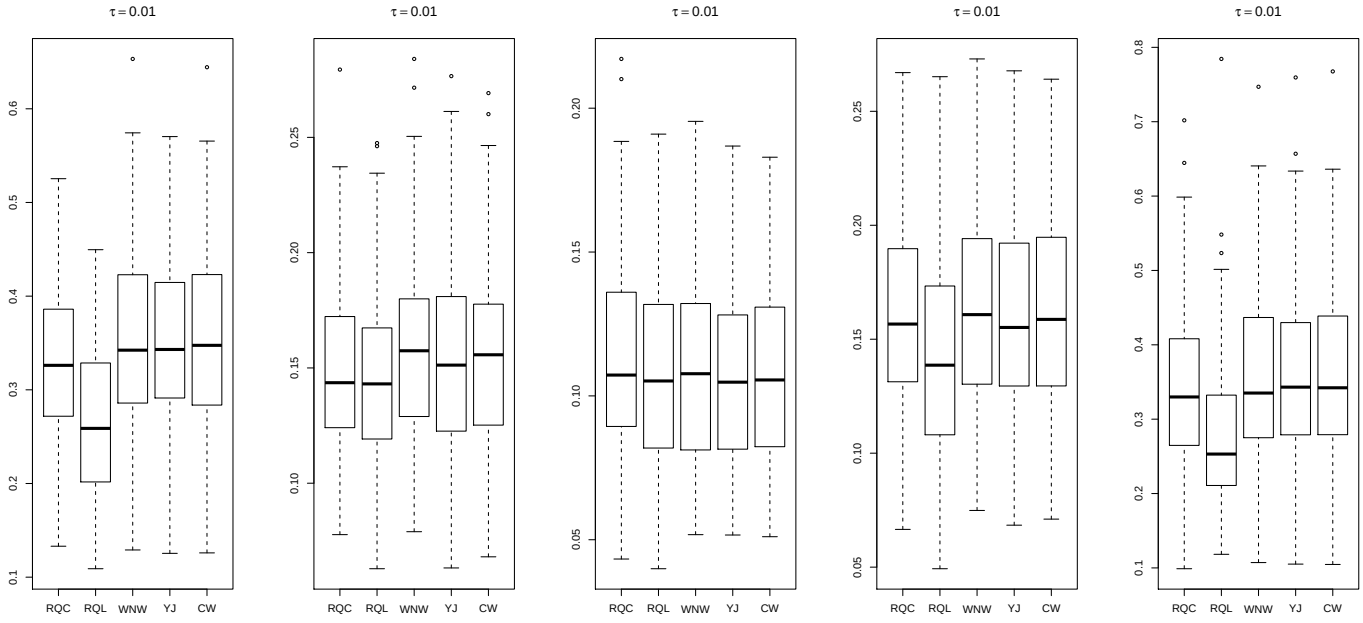


FIGURE 2.9 – Boxplots des erreurs MADE des cinq estimateurs, pour les données simulées selon le modèle ARCH, avec n=1000.

n	tau	RQC	RQL	WNW	YJ	CW
250	0.01	0.50	0.48	0.51	0.50	0.50
	0.10	0.26	0.24	0.27	0.27	0.27
	0.50	0.20	0.19	0.20	0.19	0.19
	0.90	0.25	0.24	0.26	0.25	0.26
	0.99	0.47	0.43	0.49	0.49	0.49
500	0.01	0.40	0.34	0.41	0.41	0.41
	0.10	0.20	0.18	0.20	0.19	0.19
	0.50	0.15	0.15	0.14	0.14	0.14
	0.90	0.21	0.18	0.20	0.20	0.21
	0.99	0.37	0.35	0.40	0.39	0.40
1000	0.01	0.33	0.26	0.34	0.34	0.35
	0.10	0.14	0.14	0.16	0.15	0.16
	0.50	0.11	0.11	0.11	0.10	0.11
	0.90	0.16	0.14	0.16	0.16	0.16
	0.99	0.33	0.25	0.34	0.34	0.34

TABLEAU 2.4 – Médiane des erreurs MADE, pour chaque estimateur, pour les données simulées selon le modèle ARCH.

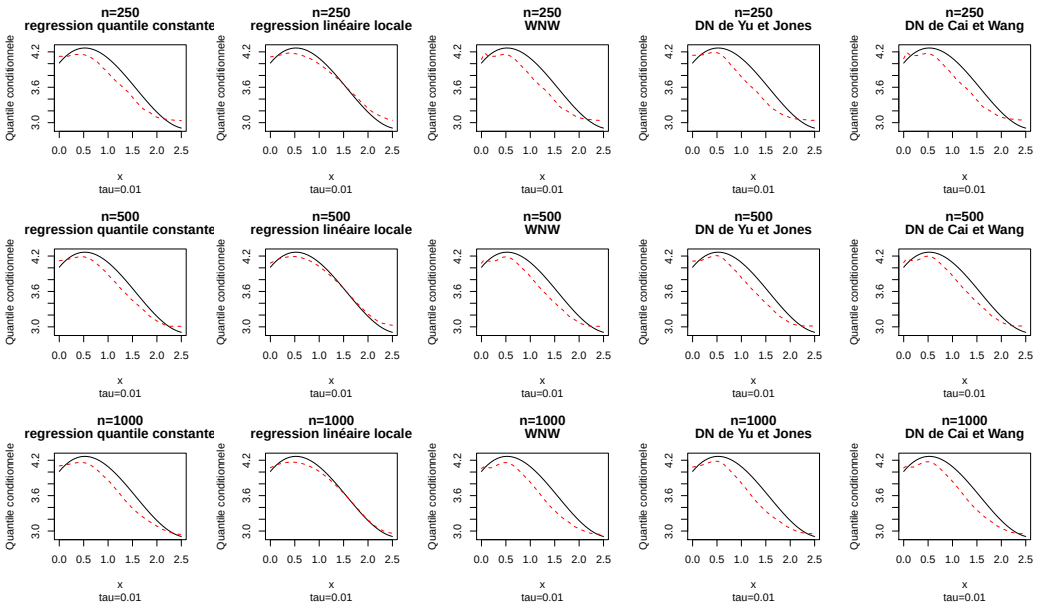


FIGURE 2.10 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.01$, pour les données simulées selon la loi de Weibull.

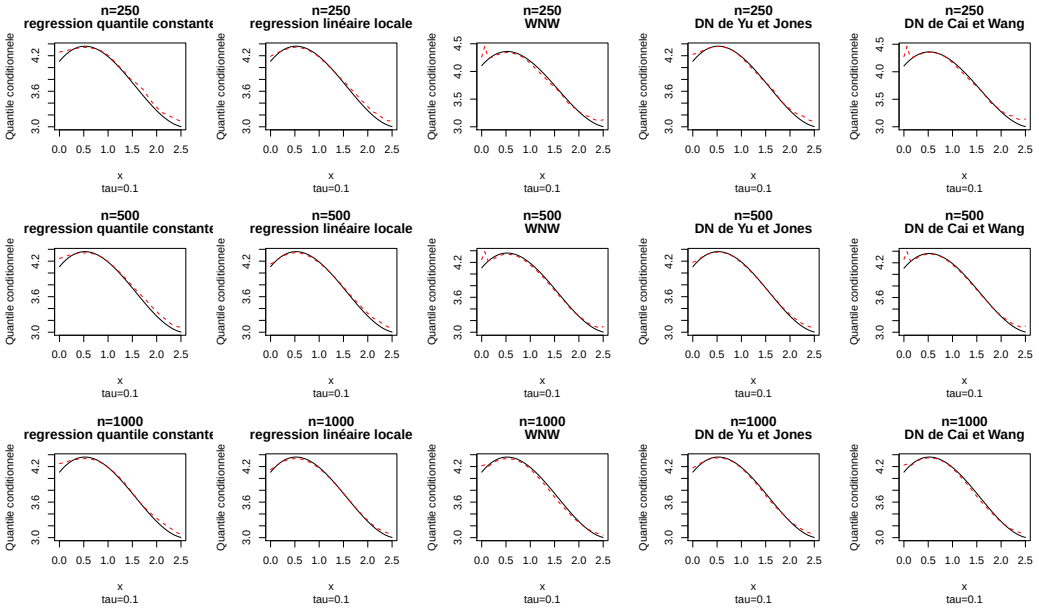


FIGURE 2.11 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.1$, pour les données simulées selon la loi de Weibull.

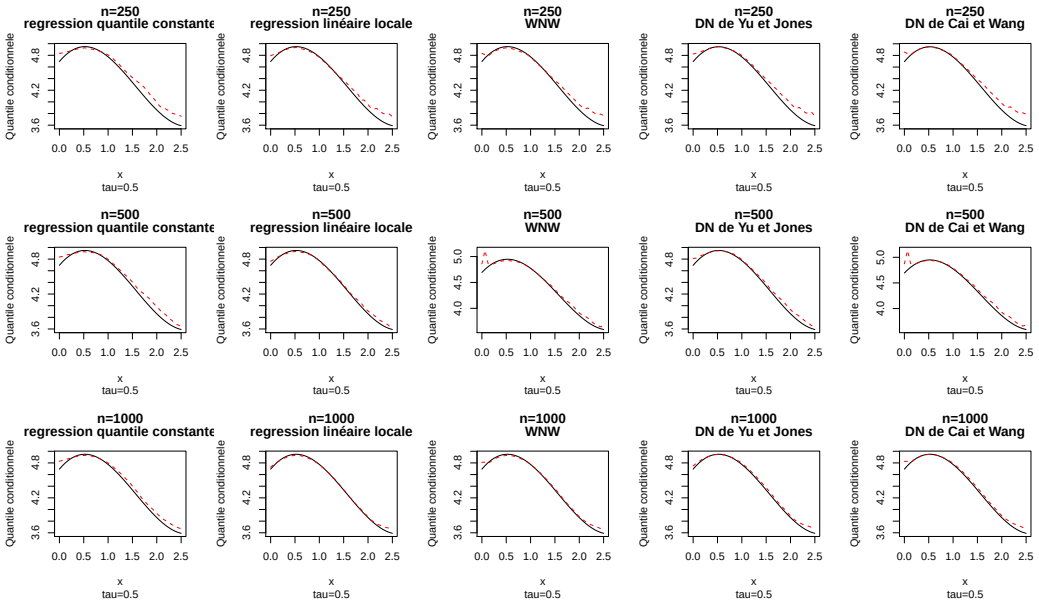


FIGURE 2.12 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.5$, pour les données simulées selon la loi de Weibull.

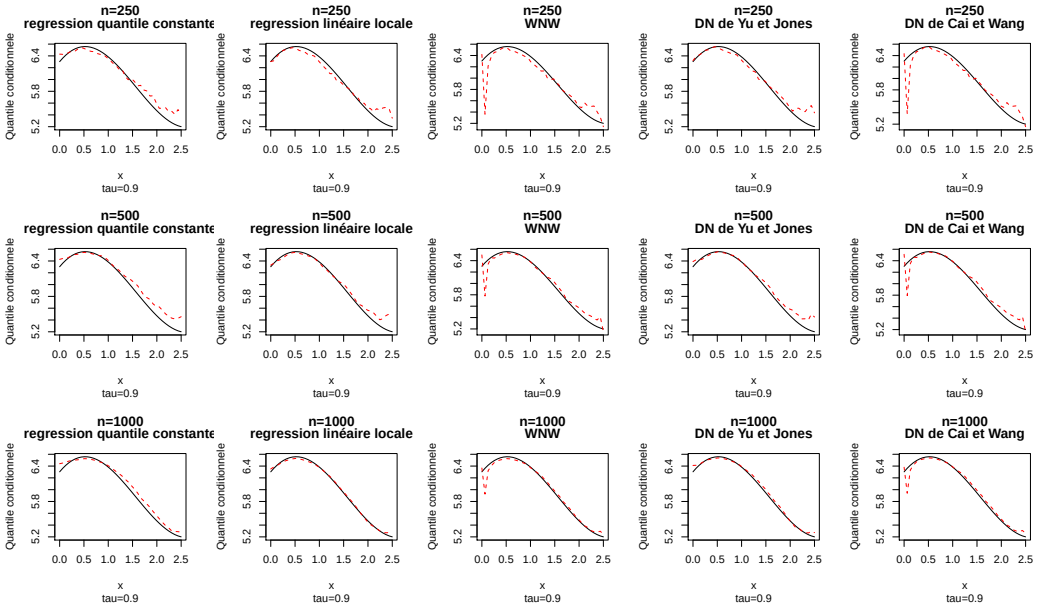


FIGURE 2.13 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.9$, pour les données simulées selon la loi de Weibull.

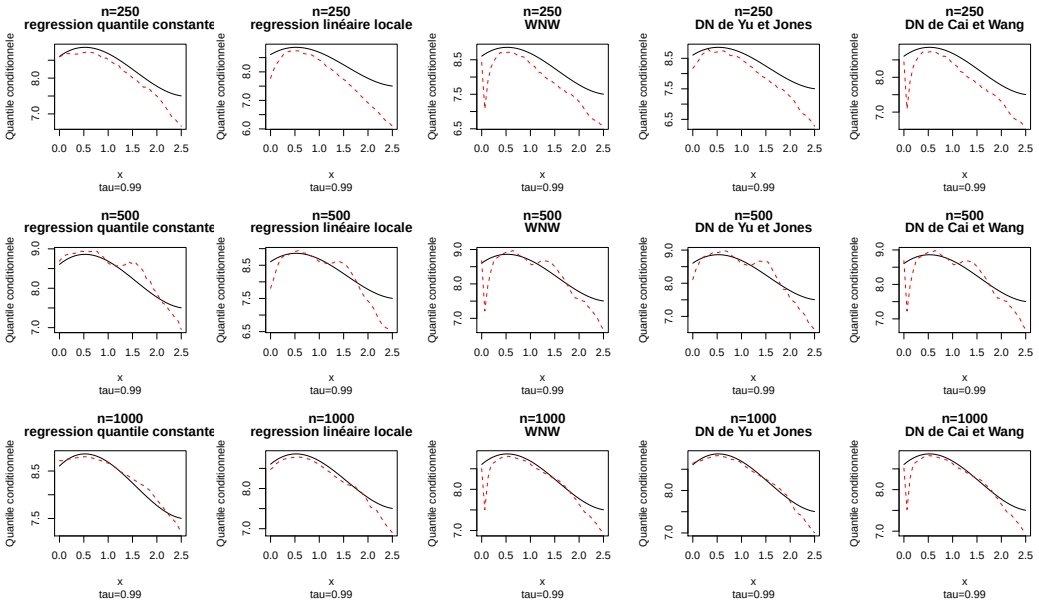


FIGURE 2.14 – Visualisation du quantile conditionnel (en noir) et son estimation (en rouge, pointillé), au niveau $\tau = 0.99$, pour les données simulées selon la loi de Weibull.

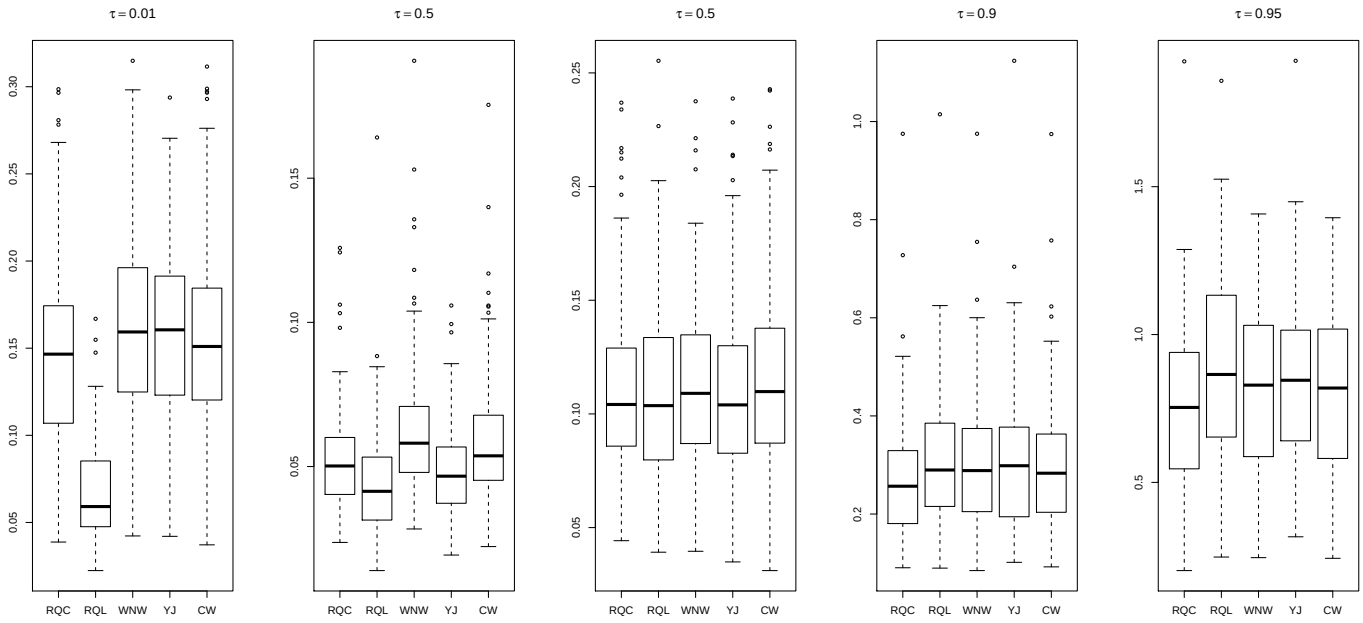


FIGURE 2.15 – Boxplots des erreurs MADE des cinq estimateurs, pour les données de Weibull, avec n=250.

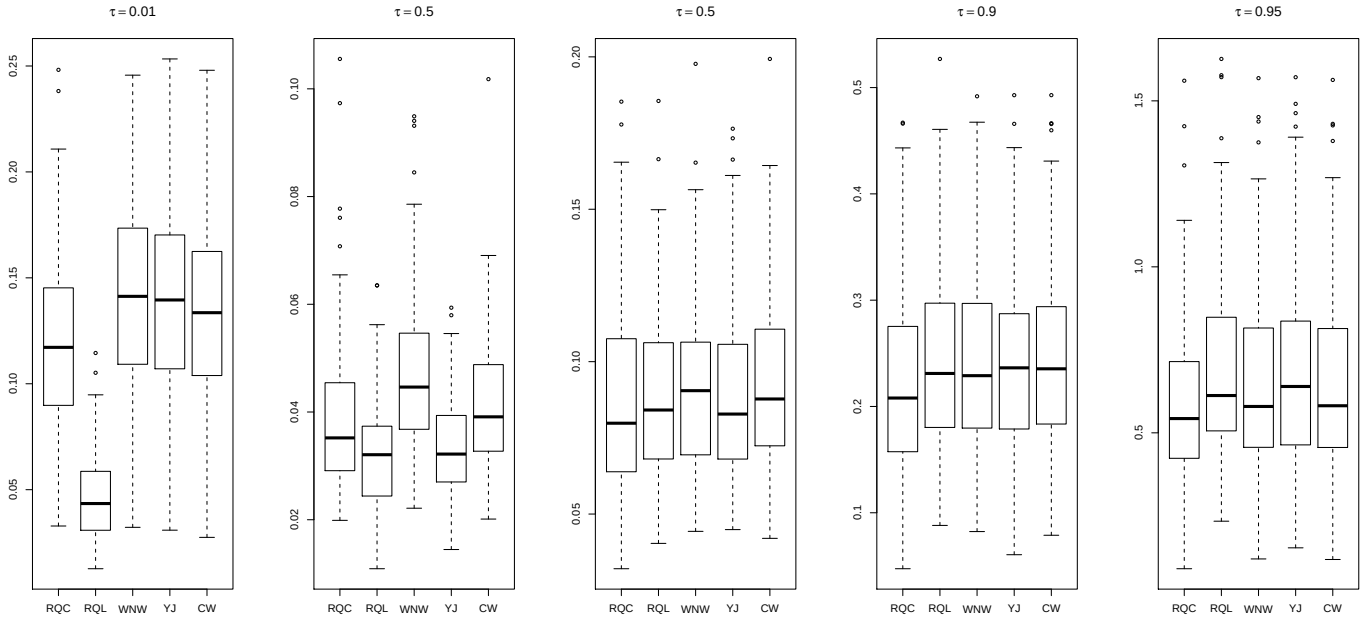


FIGURE 2.16 – Boxplots des erreurs MADE des cinq estimateurs, pour les données de Weibull, avec n=500.

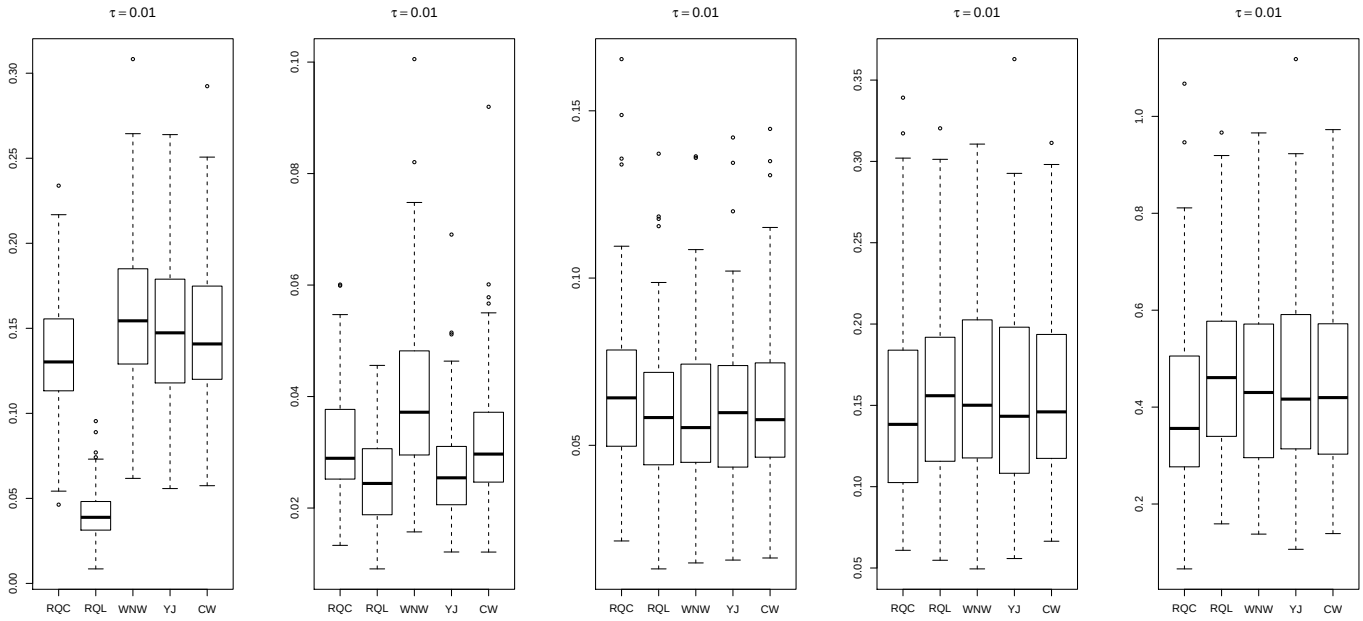


FIGURE 2.17 – Boxplots des erreurs MADE des cinq estimateurs, pour les données de Weibull, avec n=1000.

n	tau	RQC	RQL	WNW	YJ	CW
250	0.01	0.15	0.06	0.16	0.16	0.15
	0.10	0.05	0.04	0.06	0.05	0.05
	0.50	0.10	0.10	0.11	0.10	0.11
	0.90	0.26	0.29	0.29	0.30	0.28
	0.99	0.75	0.87	0.83	0.85	0.82
500	0.01	0.12	0.04	0.14	0.14	0.13
	0.10	0.04	0.03	0.04	0.03	0.04
	0.50	0.08	0.08	0.09	0.08	0.09
	0.90	0.21	0.23	0.23	0.24	0.24
	0.99	0.54	0.61	0.58	0.64	0.58
1000	0.01	0.13	0.04	0.15	0.15	0.14
	0.10	0.03	0.02	0.04	0.03	0.03
	0.50	0.06	0.06	0.06	0.06	0.06
	0.90	0.14	0.16	0.15	0.14	0.15
	0.99	0.36	0.46	0.43	0.42	0.42

TABLEAU 2.5 – Médiane des erreurs MADE des cinq estimateurs, pour les données de Weibull

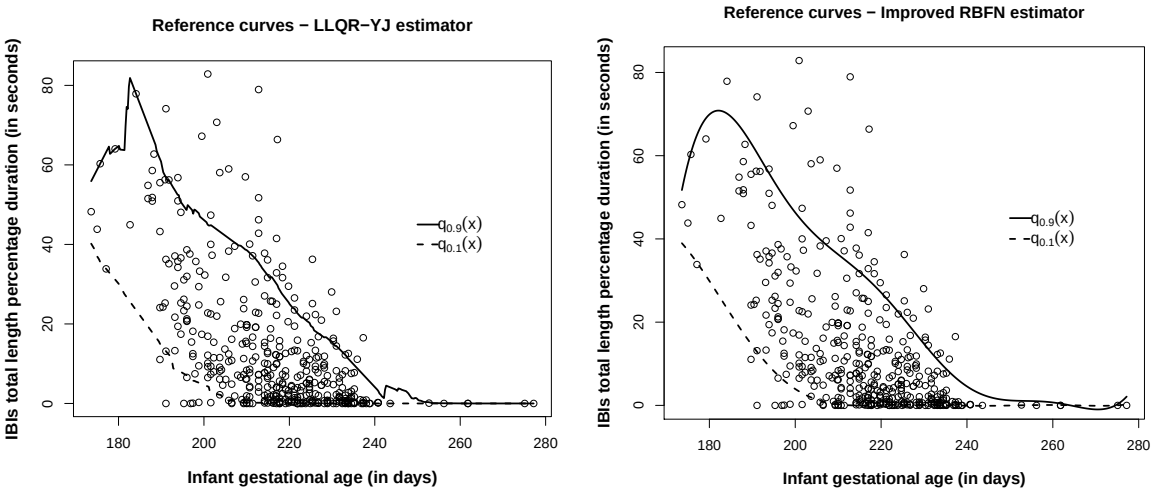


FIGURE 2.18 – Courbes de référence des données de maturation cérébrale

2.12 Références

ABBERGER, K. 1998, «Cross-validation in nonparametric quantile regression», *Allgemeines Statistisches Archiv*, vol. 82, nº 2, p. 149–161. 21, 37

ATTAR, H. E. E. 2008, *Bandwidth Selection for Local Linear Quantile Regression with Applications to Financial Market Data*, thèse de doctorat. 21, 36

DI BASILEA PER LA VIGILANZA BANCARIA, C. 2004, *International convergence of capital measurement and capital standards : A revised framework*, Bank for International Settlements. 20

CAI, Z. 2002, «Regression quantiles for time series», *Econometric Theory*, vol. 18, nº 01, p. 169–192. 21, 30, 32

CAI, Z. et R. C. TIWARI. 2000, «Application of a local linear autoregressive model to bod time series», *Environmetrics*, vol. 11, nº 3, p. 341–350. 21, 38

CAI, Z. et X. WANG. 2008, «Nonparametric estimation of conditional var and expected shortfall», *Journal of Econometrics*, vol. 147, nº 1, p. 120–130. 20, 21, 32, 33, 41

CAI, Z. et X. XU. 2008, «Nonparametric quantile estimations for dynamic smooth coefficient models», *Journal of the American Statistical Association*, vol. 103, nº 484. 20

FAN, J., T.-C. HU et Y. K. TRUONG. 1994, «Robust non-parametric function estimation», *Scandinavian Journal of Statistics*, p. 433–446. 20, 27

FAN, J., Q. YAO et H. TONG. 1996, «Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems», *Biometrika*, vol. 83, nº 1, p. 189–206. 21, 28, 36, 38

GASSER, T., A. KNEIP et W. KÖHLER. 1991, «A flexible and fast method for automatic smoothing», *Journal of the american statistical association*, vol. 86, nº 415, p. 643–652. 36

GRANGER, C. W., H. WHITE et M. KAMSTRA. 1989, «Interval forecasting : an analysis based upon arch-quantile estimators», *Journal of Econometrics*, vol. 40, nº 1, p. 87–96. 20

HALL, P., R. C. WOLFF et Q. YAO. 1999, «Methods for estimating a conditional distribution function», *Journal of the American Statistical Association*, vol. 94, nº 445, p. 154–163. 21, 30

JONES, M. et P. HALL. 1990, «Mean squared error properties of kernel estimates or regression quantiles», *Statistics & Probability Letters*, vol. 10, nº 4, p. 283–289. 25

KNEFATI, M., P. CHAUVET, A. OULIDI et M. DELECROIX. «Estimation du quantile conditionnel par les réseaux de neurones à fonction radiale de base», dans *44ièmes Journées de Statistique-JdS 2012*. 39

KNEFATI, M.-A., P. E. CHAUVET, S. N'GUYEN et B. DAYA. 2014, «Reference curves estimation using conditional quantile and radial basis function network with mass constraint», *Neural Processing Letters*, p. 1–14. 42

KOENKER, R. et G. BASSETT JR. 1978, «Regression quantiles», *Econometrica : journal of the Econometric Society*, p. 33–50. 20, 24, 25

SCAILLET, O. 2005, «Nonparametric estimation of conditional expected shortfall», *Insurance and Risk Management Journal*, vol. 74, nº 1, p. 639–660. 33

YAO, Q. et H. TONG. 1998, «Cross-validatory bandwidth selections for regression estimation based on dependent data», *Journal of Statistical Planning and Inference*, vol. 68, nº 2, p. 387–415. 38

YU, K. et M. JONES. 1997, «A comparison of local constant and local linear regression quantile estimators», *Computational statistics & data analysis*, vol. 25, nº 2, p. 159–166. 27

YU, K. et M. JONES. 1998, «Local linear quantile regression», *Journal of the American statistical Association*, vol. 93, nº 441, p. 228–237. 20, 21, 28, 29, 30, 35, 41

Chapitre 3

Estimation non-paramétrique à double noyau asymétrique

Sommaire

3.1 Introduction

L'objectif, dans ce chapitre, est d'utiliser un estimateur à double noyau, améliorant celui introduit par YU et JONES [1998], dans le cas où le support de X est positif.

Nous proposons et étudions un estimateur à double noyau, utilisant un noyau asymétrique en x. Plus précisément, nous utilisons un noyau *Beta* quand le support de la covariable X est compact et le noyau *Gamma* pour les supports bornés à gauche.

Le choix du noyau *Beta*, quand le support de la covariable est borné à gauche et à droite, est motivé par la possibilité de l'approcher uniformément, par un polynôme de Bernstein. L'approximation par des polynômes de Bernstein est équivalente à l'estimation à noyau, avec un paramètre de lissage d'ordre $n^{-\frac{1}{2}}$, (n étant la taille de l'échantillon) et un noyau binomial.

Ces estimateurs souffrent du problème de sous-lissage. Pour surmonter ce problème lorsqu'il s'agit de courbes de régression avec des points équidistants, BROWN et CHEN [1999] proposent un estimateur à noyau binomial. Ces estimateurs pallient au problème de biais aux bords. En outre, l'erreur quadratique moyenne intégrée associée est équivalente à celle d'estimateurs à noyau standard, lorsque la courbe est à support non borné.

Une extension de ce travail, au cas stochastique (*i.e.,* Les covariables sont aléatoires.) a été proposée par CHEN [2000a]. Un estimateur de densité de probabilité, basé sur les noyaux *Beta* et *Gamma* est proposé par CHEN [1999, 2000b]. Cet estimateur est à variance faible relativement au reste des estimateurs. CHEN [2002] propose, également d'utiliser les noyaux Beta et G$amma$ lors de l'estimation de la régression local linéaire, obtenant des estimateurs à variance finie et résistants aux données clairsemées.

Les estimateurs de quantile, proposés héritent de ces bonnes propriétés et l'erreur quadratique moyenne asymptotique associée est plus faible que celle des estimateurs basés sur les noyaux symétriques.

3.2 Estimateur à double noyau asymétrique en x

Soit (X,Y) un couple aléatoire, à valeurs dans $\mathbb{R}^2$, avec X à support compact, borné à gauche. Pour simplifier, nous supposerons que ce support est $[0,1]$ ou $[0,\infty)$. Notons $F(\cdot|x)$ la fonction de répartition conditionnelle de Y sachant $X = x$.

Supposons que $(X_1,Y_1),...,(X_n,Y_n)$ sont des observations indépendantes, de même loi que (X,Y). Pour estimer les quantiles $q_\tau(x)$, nous adoptons une approche indirecte. Tout d'abord, nous utilisons une méthode à double noyau pour estimer la fonction de répartition conditionnelle $F(\cdot|x)$; L'inverse de l'estimateur obtenu sera utilisé, pour estimer les quantiles conditionnels souhaités.

À cette fin, rappelons que la fonction de répartition conditionnelle $F(.|x)$ vérifie

$$F(y|x) = \arg\min_a \mathrm{E}\left[(\mathbb{1}(Y \leq y) - a)^2 | X = x\right],$$

où $\mathbb{1}(.)$ est la fonction indicatrice usuelle (cf. par exemple, FAN et YAO [2003], Page 455, pour plus de détails). Cette relation est la base de l'approche polynomiale locale classique.

Par exemple, les estimations, par régression locale linéaire, de $F(\cdot|x)$ et de sa dérivée en x, sont obtenues en minimisant selon a et b, le critère

$$\sum_{i=1}^{n}\left((\mathbb{1}(Y_i \le y) - a - b(x-X_i))^2 | X = x\right) K(x, h_1, X_i), \tag{3.1}$$

où h_1 est le paramètre de lissage en x et $K(., h_1, .)$ est une fonction noyau.

La solution $\hat{a}(x,y)$ peut être utilisée comme un estimateur non-paramétrique de $F(y|x)$. Cependant, pour tout x fixé, cet estimateur est discontinu en y. Ce défaut peut être atténué, comme proposé dans Cai et Xu [2008], en remplaçant la fonction indicatrice en (3.1) par ses versions lissées

$$(\mathbb{1} * W_{h_2})(y) = \int \mathbb{1}(Y_i \le t) W_{h_2}(y-t)\,dt = \int \mathbb{1}(t \le \frac{y-Y_i}{h_2}) W(t)\,dt := \Omega(\frac{y-Y_i}{h_2})$$

où h_2 est le paramètre de lissage en y, Ω est la fonction de répartition d'un noyau W et $W_{h_2}(.) = W(./h_2)/h_2$.

Le critère des moindres carrés pondérés en (3.1) est équivalent à

$$\sum_{i=1}^{n}\left(\left(\Omega(\frac{y-Y_i}{h_2}) - a - b(x-X_i)\right)^2 | X = x\right) K(x, h_1, X_i). \tag{3.2}$$

Nous utilisons la solution $\hat{a}$ comme estimateur de $F(\cdot|x)$; C'est l'estimateur à double noyau local linéaire et sera notée $\hat{F}(\cdot|x)$.
Les mêmes arguments que ceux dans Fan et Gijbels [1996] conduisent à l'expression

$$\hat{F}(y|x) = \sum_{i=1}^{n} \omega_i(x)\,\Omega(\frac{y-Y_i}{h_2}),$$

où

- $\omega_i(x) = \frac{S_2(x) - (x-X_i)S_1(x)}{S_2(x)S_0(x) - S_1^2(x)} K_{x,h_1}(X_i), \qquad i = 1, \ldots, n,$
- $S_l = \sum_{j=1}^{n} (x-X_j)^l K_{x,h_1}(X_i), \qquad l = 0, 1, 2,$
- $K_{x,h_1}(X_i) = K(x, h_1, X_i).$

Ensuite, pour les raisons discutées plus haut, nous faisons usage, dès maintenant, de noyaux asymétriques lors d'un lissage par rapport à x. Plus précisément, nous utilisons la densité de la loi Beta$(\frac{x}{h_1}+1, \frac{1-x}{h_1}+1)$, *i.e.*,

$$K_{x,h_1}(u) = \frac{u^{\frac{x}{h_1}}(1-u)^{\frac{1-x}{h_1}}}{B(\frac{x}{h_1}+1, \frac{1-x}{h_1}+1)}, \qquad u \in [0,1]$$

ou de la loi Gamma$(\frac{x}{h_1}+1, h_1)$, *i.e.*,

$$K_{x,h_1}(u) = \frac{u^{\frac{x}{h_1}} e^{-\frac{u}{h_1}}}{h_1^{\frac{x}{h_1}+1}\Gamma(\frac{x}{h_1}+1)}, \qquad u \in [0, \infty[$$

où B et Γ sont respectivement les fonctions *Beta* and *Gamma* classiques. Les noyaux des deux familles sont d'autant asymétriques que x est vers les bords.

Nous utilisons le noyau *Beta* si $[0,1]$ est le support de X et le noyau *Gamma* si $[0, \infty[$ est le support. Enfin, en utilisant la définition du quantile conditionnel (cf. équation 2.4), nous estimons $q_\tau(x)$, en inversant l'estimateur à la double noyau locale linéaire $\hat{F}(\cdot|x)$ *i.e.*,

$$\hat{q}_\tau(x) = \hat{F}^{-1}(\tau|x). \tag{3.3}$$

3.3 Propriétés asymptotiques

3.3.1 Développements asymptotiques de $\hat{\mathrm{F}}(\cdot|x)$

On s'intéresse, aux propriétés asymptotiques des estimateurs proposés. Plus préciment au développement asymptotique de l'erreur quadratique moyenne MSE, des estimateurs $\hat{\mathrm{F}}(\cdot|x)$ et $\hat{q}_\tau(x)$.

Notons $\mathrm{F}(x,y)$ et $f(x,y)$ la fonction de répartition et la densité de (X,Y). Soient $f(y|x)$ et $g(x)$ respectivement, la fonction de répartition conditionnelle de Y sachant X = x et la densité marginale de X . Considérons les hypothèses de régularité suivantes :

i.1) La fonction de répartition $\mathrm{F}(x,y)$ admet une densité continue bornée $f(x,y)$.

i.2) Les densités $g(.)$ et $f(.|.)$ sont bornées et strictement positive.

i.3) Pour tout $\tau \in (0,1)$ et tout x, Les quantiles conditionnels $q_\tau(x)$ sont uniques.

i.4) Le noyau W est une densité symétrique de moment de deuxième ordre fini.

i.5) Les fenêtres h_1 et h_2 satisfont $h_1 + (nh_1)^{-1} \to 0$ et $h_2 + (nh_2)^{-1} \to 0$ quand $n \to \infty$.

i.6) Pour n assez grand, $h_2 = o(h_1)$.

Aussi, rappelons que

$$\ell^{ab}(q_\tau(x)|x) = \frac{\partial^{ab}}{\partial y^a \partial x^b} l(y|x).$$

Le support de X sera partagé en deux parties : une région intérieure $\mathrm{S_I}$ et une région au bords $\mathrm{S_B}$. Pour un support compact $[0,1]$, ces régions peuvent être définies comme suit

$$\begin{aligned} \mathrm{S_I} &= \{x : x \in [0,\ 1], \frac{x}{h_1} \to \infty \text{ et } \frac{1-x}{h_1} \to \infty\}, \\ \mathrm{S_B} &= \{x : x \in [0,\ 1], \frac{x}{h_1} \to c \text{ ou } \frac{1-x}{h_1} \to c, \text{pour } c > 0\}. \end{aligned}$$

De même, pour le support $[0,\ \infty[$, nous utilisons

$$\begin{aligned} \mathrm{S_I} &= \{x : x \in [0,\ \infty], \frac{x}{h_1} \to \infty\}, \\ \mathrm{S_B} &= \{x : x \in [0,\ \infty], \frac{x}{h_1} \to c, \text{pour } c > 0\}. \end{aligned}$$

Le théorème à suivre, donne les développements asymptotiques du biais et de la variance de $\hat{\mathrm{F}}(\cdot|x)$.

Théorème 3.3.1 *Soit $\psi(x) = x(1-x)$ pour le noyau Beta et $\psi(x) = x$ pour le noyau Gamma. Sous les hypothèses i.1), i.2), i.3), i.4), i.5) et i.6),*

$$\mathrm{Bias}\{\hat{\mathrm{F}}(y|x)\} = \begin{cases} \frac{1}{2}\psi(x)\mathrm{F}^{02}(y|x)h_1 + \mathrm{O}(h_1^2) & si\ \ x \in \mathrm{S_I}; \\ \\ \frac{1}{2}(2+c)\mathrm{F}^{02}(y|x)h_1^2 + o(h_1^2) & si\ \ x \in \mathrm{S_B} \end{cases}$$

et

$$\mathrm{Var}\{\hat{\mathrm{F}}(y|x)\} = \begin{cases} \frac{\sigma^2(x,y)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)n\sqrt{h_1}} + o(\frac{1}{n\sqrt{h_1}}) & si\ \ x \in \mathrm{S_I}; \\ \\ \frac{\sigma^2(x,y)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)nh_1} + o(\frac{1}{nh_1}) & si\ \ x \in \mathrm{S_B}, \end{cases}$$

où $\sigma^2(x,y) = \mathrm{F}(y|x)(1-\mathrm{F}(y|x))$.

3.3.2 Propriétés asymptotiques de $\hat{q}_\tau(x)$

Là encore, des arguments similaires à ceux utilisés précédemment conduisent aux développements MSE de $\hat{q}_\tau(x)$.

Théorème 3.3.2 *Sous les hypothèses de Théorème 3.3.1,*

$$Bias\{\hat{q}_\tau(x)\} = \begin{cases} \frac{\frac{1}{2}\psi(x)\mathrm{F}^{02}(q_\tau(x)|x)h_1}{f(q_\tau(x)|x)} + \mathrm{O}(h_1^2) & si\ \ x \in \mathrm{S_I}; \\ \\ \frac{\frac{1}{2}(2+c)\mathrm{F}^{02}(q_\tau(x)|x)h_1^2}{f(q_\tau(x)|x)} + o(h_1^2) & si\ \ x \in \mathrm{S_B}, \end{cases}$$

et

$$Var\{\hat{q}_\tau(x)\} = \begin{cases} \frac{\tau(1-\tau)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)f^2(q_\tau(x)|x)n\sqrt{h_1}} + o(\frac{1}{n\sqrt{h_1}}) & si\ \ x \in \mathrm{S_I}; \\ \\ \frac{\tau(1-\tau)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)f^2(q_\tau(x)|x)nh_1} + o(\frac{1}{nh_1}) & si\ \ x \in \mathrm{S_B}. \end{cases}$$

Pour évaluer la performance de l'estimateur $\hat{q}_\tau(x)$, nous nous fondons sur l'erreur quadratique moyenne asymptotique AMSE. En effet, l'erreur AMSE de $\hat{q}_\tau(x)$ est donnée par

$$\mathrm{AMSE}(x) = \begin{cases} \frac{1}{4}\frac{\{\psi(x)\mathrm{F}^{02}(q_\tau(x)|x)h_1\}^2}{f^2(q_\tau(x)|x)} + \frac{\tau(1-\tau)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)f^2(q_\tau(x)|x)n\sqrt{h_1}} & \text{si}\ \ x \in \mathrm{S_I}; \\ \\ \frac{1}{4}\frac{\{(2+c)\mathrm{F}^{02}(q_\tau(x)|x)h_1^2\}^2}{f^2(q_\tau(x)|x)} + \frac{\{\tau(1-\tau)\}\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)f^2(q_\tau(x)|x)nh_1} & \text{si}\ \ x \in \mathrm{S_B}. \end{cases}$$

En minimisant $\mathrm{AMSE}(x)$, nous obtenons la fenêtre optimale

$$h_1^{opt}(x) = \begin{cases} \left(\frac{\tau(1-\tau)}{2\sqrt{\pi}\psi(x)^{5/2}\{\mathrm{F}^{02}(q_\tau(x)|x)\}^2 g(x)}\right)^{\frac{2}{5}} n^{-\frac{2}{5}} & \text{si}\ \ x \in \mathrm{S_I}; \\ \\ \left(\frac{\tau(1-\tau)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)(2+c)^2\{\mathrm{F}^{02}(q_\tau(x)|x)\}^2 g(x)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} & \text{si}\ \ x \in \mathrm{S_B}. \end{cases} \tag{3.4}$$

En substituant cette fenêtre optimale dans l'équation (3.3.2) et sous l'hypothèse $h_2 = o(h_1)$, nous obtenons l'erreur quadratique moyenne asymptotique optimale pour $\hat{q}_\tau(x)$ *i.e.,*

$$\mathrm{AMSE}^{opt}(x) = \begin{cases} \frac{5}{4}\left(\frac{\tau(1-\tau)}{2\sqrt{\pi}g(x)}\right)^{\frac{4}{5}}\{\mathrm{F}^{02}(q_\tau(x)|x)\}^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{si}\ \ x \in \mathrm{S_I}; \\ \\ \frac{5}{4}(2+c)^{\frac{1}{5}}\left(\frac{\tau(1-\tau)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)}\right)^{\frac{4}{5}}\{\mathrm{F}^{02}(q_\tau(x)|x)\}^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{si}\ \ x \in \mathrm{S_B}. \end{cases} \tag{3.5}$$

En imitant l'approche développée dans CHEN [2000a], le biais et la variance aux bords ont une contribution, à l'erreur quadratique moyenne asymptotique, négligeable avec un terme d'erreur de l'ordre $o(\frac{1}{n\sqrt{h_1}} + h_1^2)$. Il s'ensuit que

$$\mathrm{AMISE} = \frac{\tau(1-\tau)}{n\sqrt{h_1}}\int_0^b \frac{1}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)f^2(q_\tau(x)|x)}dx + \frac{h_1^2}{4}\int_0^b \psi(x)^2\left(\frac{\mathrm{F}^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right)^2 dx.$$

où $b = 1$ pour un noyau *Beta* et $b = \infty$ pour noyau *Gamma*.

En minimisant AMISE selon h_1, nous obtenons la fenêtre globale (selon notre classification), asymptotiquement optimale :

$$h_1^{opt} = \left\{ \frac{\tau(1-\tau)\int_0^b \frac{dx}{\sqrt{\psi(x)}g(x)f(q_\tau(x)|x)}}{2\sqrt{\pi}\int_0^b \psi(x)^2 \left(\frac{F^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right)^2 dx} \right\}^{\frac{2}{5}} n^{-\frac{2}{5}} \tag{3.6}$$

Les développements asymptotiques liés aux noyaux *Gamma* et *Beta* sont identiques, à l'exception du terme $\sqrt{1-x}$. Comme il a été remarqué par CHEN [2002], la variance liée au noyau *Gamma* diminue quand x augmente ($x^{-\frac{1}{2}}$ intervient au travers de ψ, au dénominateur de la variance asymptotique).

Cette propriété est fortement recommandée pour l'estimation des courbes de $\widehat{q}_\tau(x)$, dans les situations où les observations sont clairsemées à l'extrémité supérieure du support de g. Ce gain de variance pour de grandes valeurs x, conduit à l'augmentation du biais ; Cela est similaire à ce qui se produit, lorsqu'on utilise des grandes fenêtres, pour les régions clairsemées.

AMISE n'est pas affecté par le comportement des estimateurs, en $x \in S_B$; Son optimalité aux bords est perdue, lorsque la fenêtre globale h_1^{opt} est utilisée sur l'ensemble du support. En effet, la fenêtre h_1^{opt} est d'ordre $n^{-2/5}$, tandis que la fenêtre optimale locale est d'ordre $n^{-1/5}$ pour x dans S_B (voir (3.4)). Ainsi, en utilisant une seule fenêtre, à la fois sur S_I et sur S_B, cela donne lieu à des effets de bords.

3.4 Comparaison : Noyaux symétriques et noyaux asymétriques *Beta* et *Gamma*

3.4.1 MSE asymptotique

Selon YU et JONES [1998], l'erreur optimale asymptotique AMSE pour l'estimateur à double noyau, de quantiles conditionnels, est

$$\mathrm{AMSE}_{sym}^{opt}(x) = \begin{cases} \frac{5}{4}\mu_2(K_{sym})^{\frac{2}{5}} \left(R(K_{sym})\frac{\tau(1-\tau)}{g(x)}\right)^{\frac{4}{5}} \{F^{02}(q_\tau(x)|x)\}^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{si } x \in S_I; \\ \frac{5}{4}\alpha_c(K_{sym})^{\frac{2}{5}} \left(\frac{\tau(1-\tau)\beta_c(K_{sym})}{g(x)}\right)^{\frac{4}{5}} F^{02}(q_\tau(x)|x)^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{si } x \in S_B, \end{cases} \tag{3.7}$$

avec

- K_{sym} est un noyau symétrique,
- $\alpha_c(K_{sym}) = \frac{a_2^2 - a_1 a_3}{a_0 a_2 - a_1^2}$,
- $\beta_c(K_{sym}) = \frac{\int_\infty^c \{a_2 - a_1 u\}^2 K_{sym}(u)du}{\{a_0 a_2 - a_1^2\}^2}$,
- $a_l = \int_0^c u^l K_{sym}(u)du$, pour l=0,1,2,3.

L'erreur AMSE des deux types d'estimateurs (à noyau symétrique et à noyau asymétrique) est de même ordre de grandeur, soit $n^{-\frac{4}{5}}$.

La différence réside dans les constantes multiplicatives données dans le tableau (3.1), ci-dessous ; $C(K_{sym})$ est la constante multiplicative de AMSE de l'estimateur à noyau symétrique et $C(K_{asym})$, lorsque le noyau est asymétrique.
Nous n'utilisons que le noyau d'Epanechnikov, puisque choix optimal, pour les noyaux

TABLEAU 3.1 – Comparaison : noyau symétrique et noyau asymétrique

	$C(K_{sym})$	$C(K_{asym})$
Interior	$\mu_2(K_{sym})^{2/5}R(K_{sym})^{4/5}$	$\left(\frac{1}{2\sqrt{\pi}}\right)^{4/5}$
Boundary	$\alpha_c(K_{sym})^{2/5}\beta_c(K_{sym})^{4/5}$	$(2+c)^{1/5}\left(\frac{\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)}\right)^{4/5}$

symétriques, comme le prouve FAN et *al.* [1997]. On constate que pour x appartenant à la région intérieure S_I, $C(K_{sym}) \approx 0.34$ et $C(K_{asym}) = 0.36$. Cela montre simplement que la différence entre ces deux familles de noyaux est négligeable pour $x \in S_I$.
En revanche si $x \in S_B$ la différence est importante. Le tableau 3.2 ci-dessous présente plusieurs valeurs de ces constantes et démontre la supériorité potentielle des noyaux asymétriques quand le support de X est compact ou borné à gauche.

TABLEAU 3.2 – Valeurs de $C(K_{sym})$, et $C(K_{asym})$ aux bords de x, pour différentes valeurs de $c \in]0,1[$.

c	$C(K_{sym})$	$C(K_{asym})$
0.01	27.42	0.652
0.25	2.192	0.542
0.5	1.39	0.48
0.75	0.7	0.44
0.99	0.524	0.414

3.4.2 Variance finie

SEIFERT et GASSER [1996a] montre que l'estimateur de la régression locale linéaire, peut avoir une variance inconditionnelle infinie lorsque les noyaux compacts sont utilisés ; Le dénominateur $S_2(x)S_0(x) - S_1^2(x)$ peut être nul. Pour pallier à ce problème, FAN [1993] ajoute n^{-2} au dénominateur, tandis que SEIFERT et GASSER [1996b] proposent la *ridge regression* qui ajoute le scalaire c_1 à $S_2(x)$.

Une solution alternative repose sur l'utilisation des noyaux *Beta* ou *Gamma*, exactement, comme indiqué par CHEN [2002]. L'estimateur associé à une variance finie avec probabilité 1.

3.5 Applications empiriques

3.5.1 Choix de la fenêtre

Le choix de h_2 est moins important que celui de h_1 (voir YU et JONES [1998], par exemple) ; Il doit juste vérifier l'hypothèse de régularité $h_2 = o(h_1)$. Ainsi, nous prenons $h_2 = h_1 n^{-\epsilon}$, avec $\epsilon > 0$ (dans nos simulations, $\epsilon = 0.5$).

Concernant h_1, nous utilisons la méthode *leave-one-out* de validation croisée, adoptée pour la régression quantile, par ABBERGER [1998] *i.e.,* Ainsi, h_1 minimise le critère

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \rho_\tau(\mathrm{Y}_i - \hat{q}_\tau^{(-i)}(\mathrm{X}_i)),$$

où $\rho_\tau(u) = u\tau\mathbb{1}(u \geq 0) + u(1-\tau)\mathbb{1}(u < 0)$ et $\hat{q}^{(-i)}$ est l'estimateur du quantile conditionnel, construit à partir de toutes les observations de l'échantillon, exceptée la $i^{\text{ème}}$.

Dans le but de comparer et évaluer les estimations proposées, nous utilisons le critère de validation croisée, pour déterminer la fenêtre pour des noyaux symétriques, comme asymétriques ; Nous utilisons la fenêtre théorique donnée par l'équation (3.6), lorsque les données sont simulées.

3.5.2 Données simulées

Pour l'étude des performances des noyaux asymétriques, nous considérons deux jeux de données simulées ; Les observations sont générées de manière à être clairesemées, à l'extrémité droite du support.

Afin d'éviter le problème de dégénérescence de la variance, nous utilisons la *ridge regression*. Nous évaluons la performance des estimateurs à double noyau et comparons les noyaux symétriques (gaussien et Epanechnikov) aux noyaux asymétriques (*Gamma* et *Beta*).

Pour sélectionner le paramètre de lissage h_1, nous utilisons la fenêtre optimale globale donnée en (3.6) pour les noyaux *Beta* et *Gamma*, tandis que pour les noyaux symétriques, nous utilisons comme fenêtre optimale globale (voir, YU et JONES [1998]),

$$h_{1,sym}^{opt} = \left\{ \frac{\mu_0(\mathrm{K}^2)\tau(1-\tau)\int_0^\infty \frac{dx}{g(x)f(q_\tau(x)|x)}}{\mu_2^2(\mathrm{K})\int_0^\infty \left(\frac{\mathrm{F}^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right)^2 dx} \right\}^{\frac{1}{5}} n^{-\frac{1}{5}},$$

$h_2 = h_1 n^{-0.5}$ et W est le noyau gaussien.

Pour chaque modèle, nous considérons trois tailles d'échantillon : $n = 50,\ 100,\ 200$; Pour chacune des tailles d'échantillon, on génère 100 échantillons. Nous calculons l'estimateur du quantile conditionnel, selon les trois niveaux $\tau = 0.05, 0.5$ et 0.95.

Nous utilisons l'erreur moyenne absolue `MADE`

$$\mathtt{MADE} = \frac{1}{n^*}\sum_{j=1}^{n^*} |\hat{q}_\tau(x_j) - q_\tau(x_j)|$$

où x_j, $j = 1, \ldots, n^*$ sont les points d'une grille régulière.

Modèle générateur 1 :
Comme première application, nous considérons le modèle paramétrique

$$Y_i = (X_i - 0.5)^2 + \epsilon_i,$$

où les termes ϵ_i sont indépendants et de loi $N(0, 0.05^2)$ et les X_i sont indépendants de loi $N(0, 0.3^2)$, tronquée sur $[0,\ 1]$. Pour ce modèle, le noyau *Beta* est utilisé, pour lisser par rapport à x.

La figure 3.1 relate les résultats, pour les échantillons simulés de tailles respectives 50, 100, et 200. Elle montre la vraie fonction quantile conditionnel, et trois estimateurs à double noyau (gaussien, Epanechnikov et Beta).
La figure 3.2 présente le *boxplot* des valeurs, pour les trois estimateurs du quantile conditionnel, considérés . Il y apparaît un sérieux biais au voisinage de $x = 1$; Ce biais est dû à la densité de X qui est monotone décroissante dans $[0, 1]$. La figure 3.2 montre clairement que les estimateurs basés sur le noyau *Beta* sont plus performants que les deux autres. La différence semble être moins importante lorsque $\tau = 0.5$.

Modèle générateur 2 :
Comme deuxième application, considérons le modèle, à support borné à gauche,

$$Y_i = exp(-X_i) + exp(-4(X_i - 1)^2) + \epsilon_i$$

où les termes ϵ_i sont indépendants, de loi $N(0, 0.05^2)$ et les X_i sont indépendants de loi $N(0, 1)$, tronquée à gauche sur $[0, \infty)$. Comme le support de X est borné à gauche, nous utilisons le noyau *Gamma* pour lisser par rapport à X. Les résultats de la simulation sont résumés par la figure 3.3 et le *boxplot* des 100 valeurs de `MADE` est donné, dans la figure 3.4. Là encore, les estimateurs basés sur le noyau *Gamma* sont plus performants que leurs concurrents (les estimateurs basés respectivement, sur le noyau gaussien et le noyau d'Epanechnikov).

3.5.3 Données réelles : *geyser*

Comme dernier exemple, considérons les données des éruptions du geyser du *Yellowstone National Park*, consistant en 299 observations représentant le temps écoulé entre les éruptions (Y) et la durée de l'éruption (X). Ces données, sont obtenues du package `MASS`, du logiciel `R`.

Dans la figure 3.5 nous avons tracé des courbes de référence qui sont des estimateurs du quantile conditionnel (au niveau 0.1, 0.25, 0.5, 0.75, et 0.9) en utilisant l'estimateur à double noyau, basé sur les noyaux gaussien et *Gamma*, pour le lissage en x. Les valeurs des fenêtres sont sélectionnées comme expliqué dans (3.5.1).

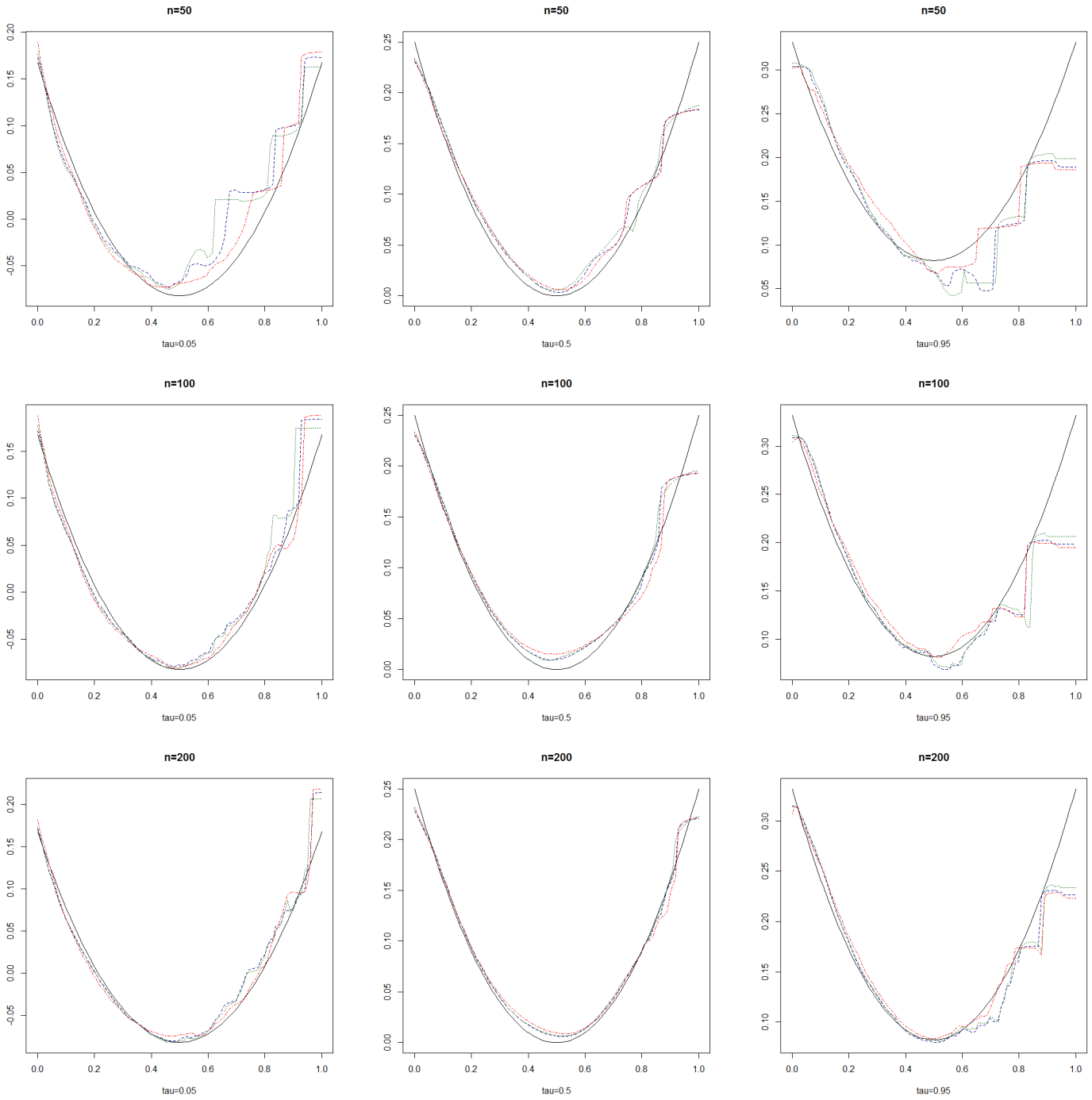


FIGURE 3.1 – Estimation du quantile conditionnel. – Vrai quantile, – Estimation par noyau gaussien, – Estimation par noyau Epanechnikov, – Estimation par noyau *Beta*

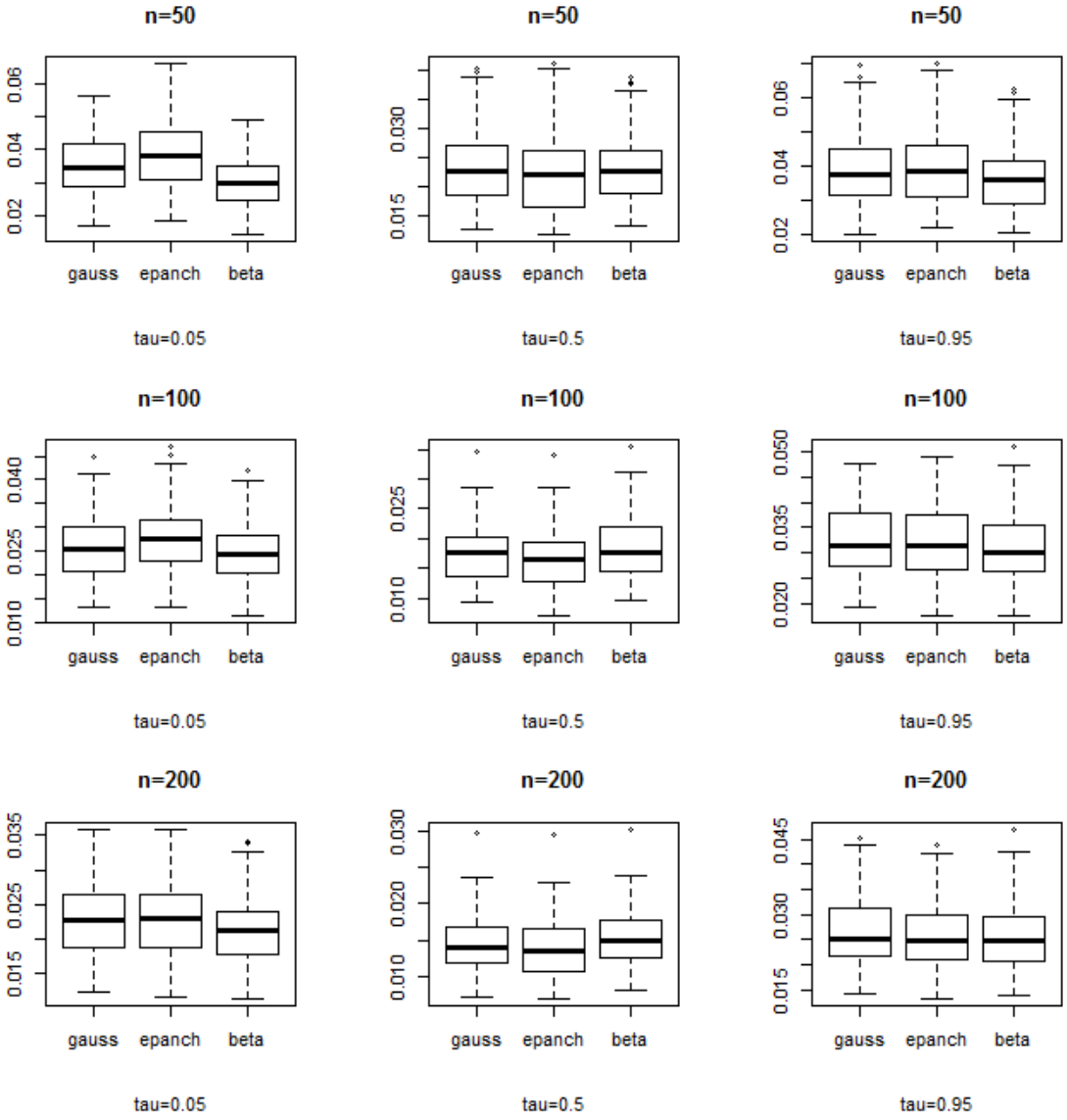


FIGURE 3.2 – *Boxplots* des 100 valeurs de *MADE* en modèle 1.

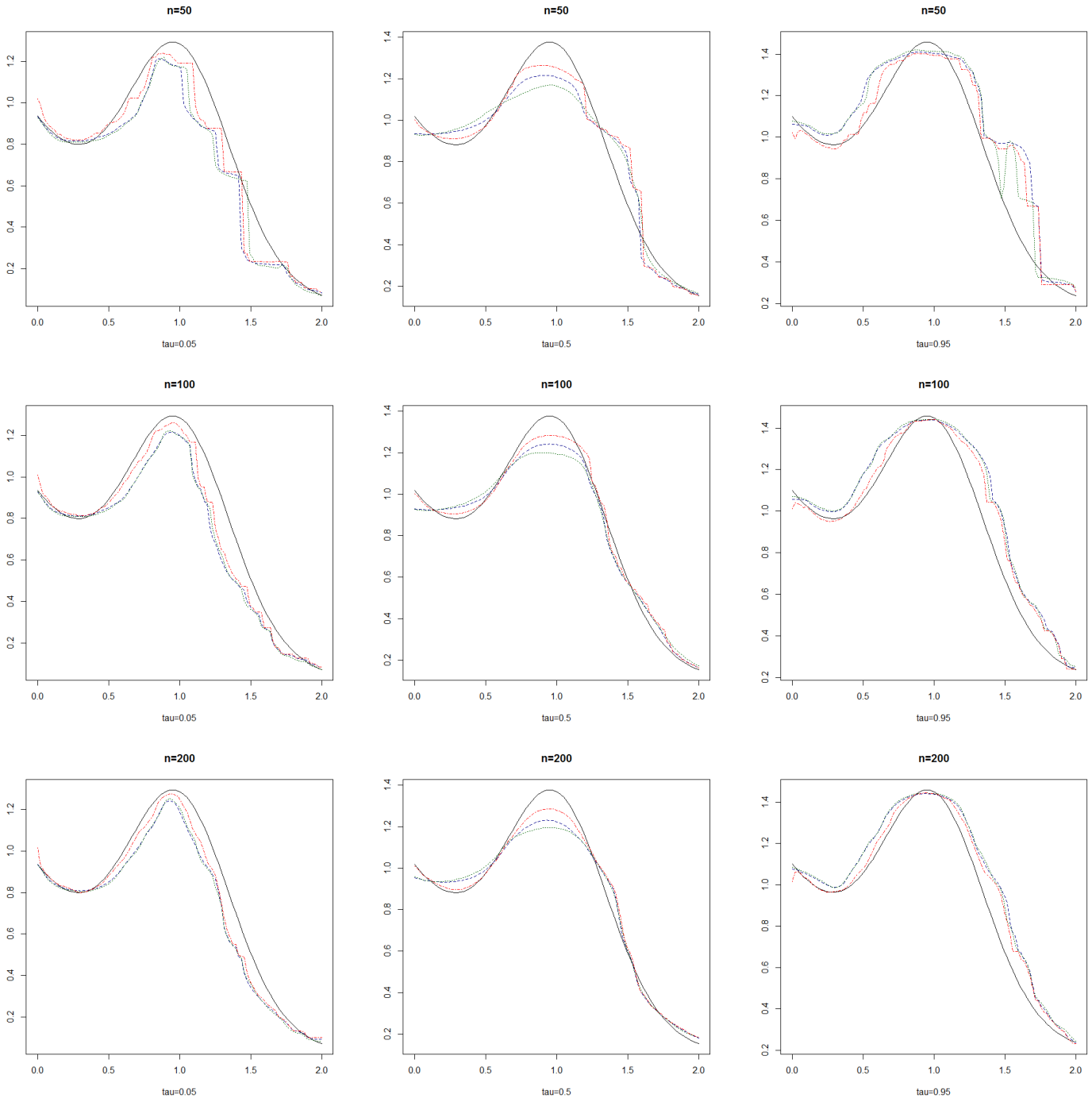


FIGURE 3.3 – Estimation du quantile conditionnel. – Vrai quantile, – Estimation par noyau gaussien, – Estimation par noyau Epanechnikov, – Estimation par noyau *Gamma*

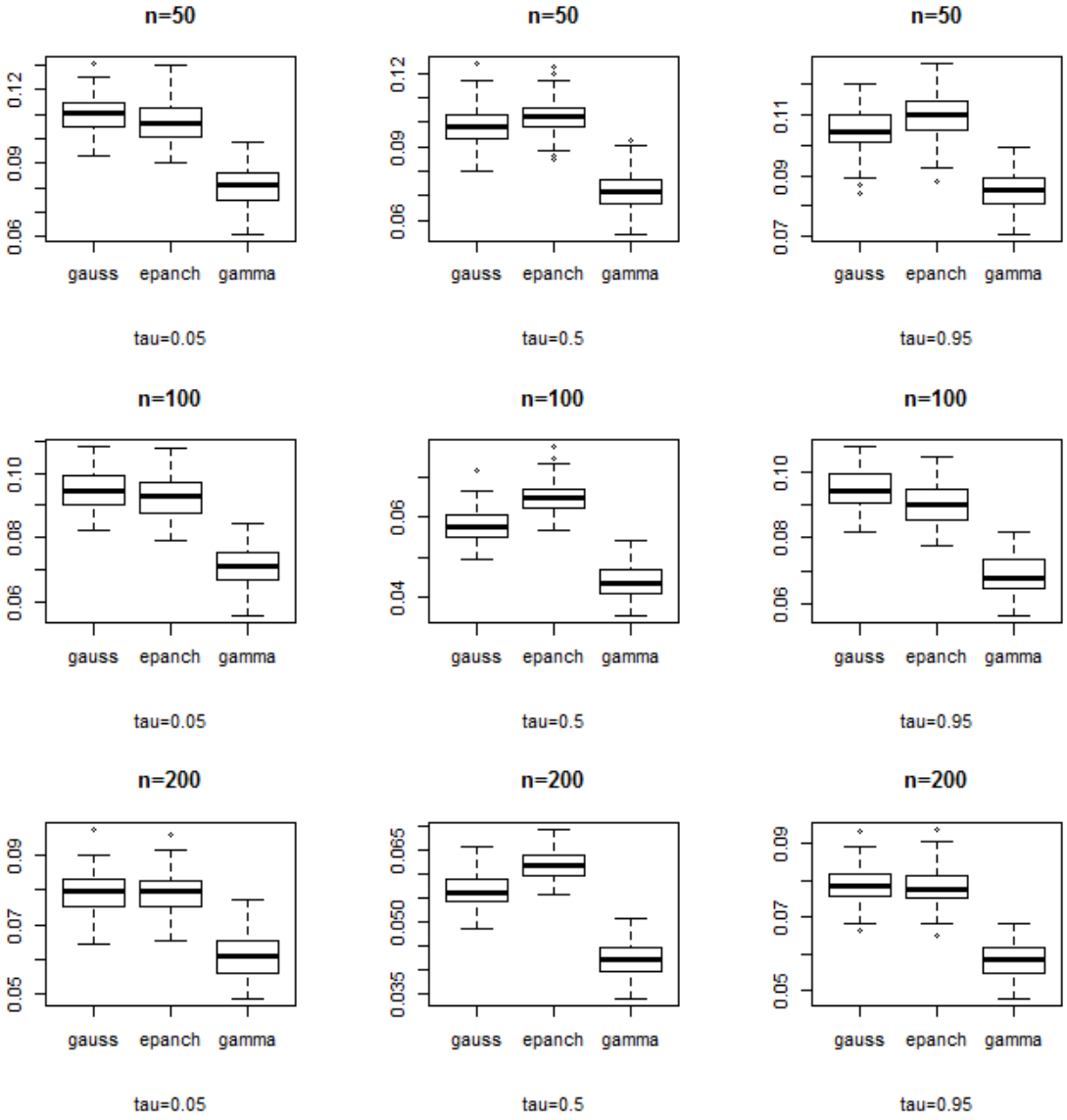


FIGURE 3.4 – *Boxplots* des 100 valeurs de *MADE* en modèle 2.

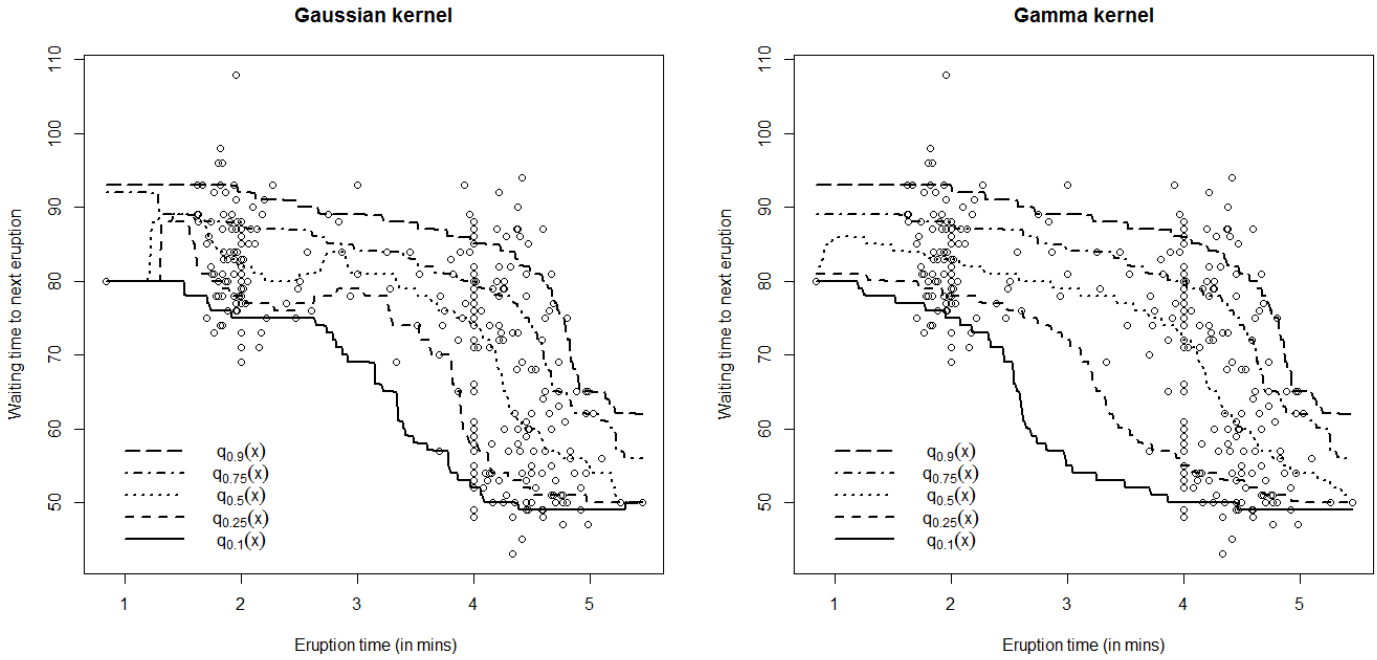


FIGURE 3.5 – Courbes de référence avec le noyau Gaussian/*Gamma*.

3.6 Références

ABBERGER, K. 1998, «Cross-validation in nonparametric quantile regression», *Allgemeines Statistisches Archiv*, vol. 82, nº 2, p. 149–161. 62

BROWN, B. M. et S. X. CHEN. 1999, «Beta-bernstein smoothing for regression curves with compact support», *Scandinavian Journal of Statistics*, vol. 26, nº 1, p. 47–59. 56

CAI, Z. et X. XU. 2008, «Nonparametric quantile estimations for dynamic smooth coefficient models», *Journal of the American Statistical Association*, vol. 103, nº 484. 57

CHEN, S. X. 1999, «Beta kernel estimators for density functions», *Computational Statistics & Data Analysis*, vol. 31, nº 2, p. 131–145. 56

CHEN, S. X. 2000a, «Beta kernel smoothers for regression curves», *Statistica Sinica*, vol. 10, nº 1, p. 73–92. 56, 59

CHEN, S. X. 2000b, «Probability density function estimation using gamma kernels», *Annals of the Institute of Statistical Mathematics*, vol. 52, nº 3, p. 471–480. 56

CHEN, S. X. 2002, «Local linear smoothers using asymmetric kernels», *Annals of the Institute of Statistical Mathematics*, vol. 54, nº 2, p. 312–323. 56, 60, 61

FAN, J. 1993, «Local linear regression smoothers and their minimax efficiencies», *The Annals of Statistics*, p. 196–216. 61

FAN, J., T. GASSER, I. GIJBELS, M. BROCKMANN et J. ENGEL. 1997, «Local polynomial regression : optimal kernels and asymptotic minimax efficiency», *Annals of the Institute of Statistical Mathematics*, vol. 49, nº 1, p. 79–99. 61

FAN, J. et I. GIJBELS. 1996, «Local polynomial modelling and its applications», . 57

FAN, J. et Q. YAO. 2003, *Nonlinear time series : nonparametric and parametric methods*, Springer Science & Business Media. 56

SEIFERT, B. et T. GASSER. 1996a, «Finite-sample variance of local polynomials : analysis and solutions», *Journal of the American Statistical Association*, vol. 91, nº 433, p. 267–275. 61

SEIFERT, B. et T. GASSER. 1996b, «Variance properties of local polynomials and ensuing modifications», dans *Statistical Theory and Computational Aspects of Smoothing*, Springer, p. 50–79. 61

YU, K. et M. JONES. 1998, «Local linear quantile regression», *Journal of the American statistical Association*, vol. 93, nº 441, p. 228–237. 56, 60, 62

Deuxième partie

Transfert d'un modèle statistique, en classification supervisée

Chapitre 4

Introduction à l'apprentissage statistique

Sommaire

4.1 Introduction

L'apprentissage statistique est un vaste champ de recherche, intéressant différents domaines d'application, comme l'écologie, la médecine, la biologie, la banque, l'actuariat et assurance, l'informatique,...

Les recherches et travaux portent notamment, sur la construction de méthodes de prédiction à but décisionnel et leur amélioration. Nous nous limitons, dans cette thèse, à la prédiction du groupe d'appartenance de l'individu statistique, à partir de sa description suivant des variables (ou covariables) fixées par avance. On se limite donc, à la classification supervisée ou discrimination statistique à but décisionnel (*Machine learning, statistical learning,...*).

Étant données

- une population statistique, partition exacte en un nombre fini de classes ou groupes prédéfinis,
- des variables de description, de cette population (On dit aussi, variables explicatives, covariables, descripteurs,...),

le problème est d'estimer, en se servant d'un échantillon observé, une règle optimale d'affectation aux classes, à partir de la seule description. L'optimalité de la règle signifie une erreur d'affectation de coût minimal.

On distingue trois grandes catégories de méthodes de classification, dépendant de la donnée ou pas d'une variable "groupe d'appartenance" ou variable réponse :

- **La Classification supervisée** : La description et la classe d'appartenance sont connues pour chaque individu de l'échantillon ;
- **La Classification non-supervisée** : La description des individus de l'échantillon est connue mais leur classe d'appartenance est inconnue. Bien mieux, il arrive qu'on ne connaisse pas de partition (ni le nombre de classes, ni la définition de celles-ci.), *i.e.*, il n'y a pas de variable "groupe d'appartenance" ;
- **La Classification semi-supervisée** : La description est connue pour l'ensemble des individus de l'échantillon ; la classe d'appartenance est connue partiellement *i.e.*, uniquement, pour une partie des individus de l'échantillon.

Nous nous plaçons, ici, en classification supervisée. Le problème est d'estimer un lien simple entre les covariables et l'appartenance. L'utilité de ce lien est l'étude de l'effet de chaque covariable sur l'appartenance au groupe, autrement dit l'étude du pouvoir de chaque covariable à séparer les groupes ; On dit aussi pouvoir explicatif.
Bien mieux, si les covariables sont explicatives, parfois ce lien permet de prédire efficacement le groupe d'appartenance de l'individu statistique.

4.2 Notations et vocabulaire

Soit

- $\mathcal{U}$ un univers ou population statistique, partition exacte des C classes ou groupes $G_1, G_2, \ldots, G_C$,

- $\mathbf{X}$ ($\mathcal{U} \to \mathbb{R}^p$) un vecteur de covariables, et $\mathcal{X} \subset \mathbb{R}^p$ son domaine de réalisations ou des valeurs possibles,
- $\mathbf{Z}$ ($\mathcal{U} \to \{0,1\}^C$) le label (ou groupe d'appartenance) et $\mathcal{Z} \subset \{0,1\}^C$ son domaine des valeurs possibles. Dans le cas binaire, $\mathbf{Z} = (1-Y, Y)$ et $Y \in \{0,1\}$. Plus généralement, dans le cas multinomial, $Z = e_Y$ avec $Y \in \{1,...,C\}$ et e_Y le vecteur de la base canonique correspondant (la $Y^{ème}$ composante vaut 1 et toutes les autres, zéro.),
- $\mathcal{S}$ l'observation d'un échantillon de n individus de $\mathcal{U}$ *i.e.*, $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{z}_1), ..., (\mathbf{x}_n, \mathbf{z}_n)\}$. Cet échantillon est supposé indépendant et identiquement distribué.

Dans la suite, on distinguera l'échantillon $\mathcal{S}$, de taille n, et l'échantillon d'apprentissage $\mathcal{S}_T \subseteq \mathcal{S}$, de taille $n_T \leq n$, qu'on utilise pour apprendre la règle ou l'estimer.

Le problème principal, en classification supervisée, est de prédire, pour un individu dont le groupe est inconnu, le groupe d'appartenance Y à partir de la description $\mathbf{X}$, en se servant d'observations de l'échantillon $\mathcal{S}$.

Exemples introductifs :

— Biomètrie : $\mathcal{U}$ est une population d'oiseaux d'une même espèce, $\mathbf{X}$ est le vecteur des descripteurs morphométriques (envergure des ailes ; longueur des tarses ;...), $\mathbf{Z}$ indique le sexe et $\mathcal{S}$ est un échantillon observé quant au couple $(\mathbf{X}, \mathbf{Z})$.
Le problème est de prédire le sexe (mâle ou femelle), à partir de la description $\mathbf{X}$, en se servant d'observations de l'échantillon $\mathcal{S}$. On trouve de telles données, dans BIERNACKI et *al.* [2002].

— *Credit-scoring* : $\mathcal{U}$ est une population de prétendants à un prêt, d'une même référence (Prêt immobilier ou bien prêt à la consommation,...), $\mathbf{X}$ est le vecteur de descripteurs socio-économiques de l'emprunteur, $\mathbf{Z}$ indique le comportement, en remboursement (0 incident ; 1 incident ; 2 incidents et plus.) et $\mathcal{S}$ est un échantillon observé.
Le problème pour cet exemple, est de prédire le comportement en remboursement $\mathbf{Z}$, à partir de la description $\mathbf{X}$, en se servant d'observations de l'échantillon $\mathcal{S}$. Ici, l'échantillon d'apprentissage est constitué d'emprunteurs dont le dossier est clos.
Nous traiterons, comme indiqué en section 6.6, dévolue aux expérimentations numériques, les données bancaires disponibles au lien `http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kreditvar_e.html`.

— *Marketing* : $\mathcal{U}$ est une population de consommateurs avec carte de fidélité, $\mathbf{X}$ est le vecteur de descripteurs (information extraite du formulaire d'adhésion) et une information additive sur la consommation, $\mathbf{Z}$ est la tranche de pouvoir d'achat et $\mathcal{S}$ est un échantillon observé, quant aux habitudes de consommation.
Le problème est de prédire la tranche de pouvoir d'achat $\mathbf{Z}$, à partir des descripteurs $\mathbf{X}$, en se servant des observations de l'échantillon $\mathcal{S}$.

4.3 Les catégories de méthodes d'apprentissage supervisé

Les méthodes d'appentissage supervisé, sont très nombreuses ; Nous les regroupons, selon le type de critère optimisé, pour affecter aux groupes d'appartenance.

Les règles probabilistes ou bayesiennes : Elles sont basées sur la maximisation de la probabilité *a posteriori*, d'appartenance au groupe. C'est-à-dire, l'affectation d'un profil $\mathbf{x}$ à une classe, se base sur l'estimation des probabilités *a posteriori* associées $P(Y=\ell|\mathbf{X}=\mathbf{x})$, $\ell=1,\ldots,C$: On affecte à la classe $\arg\max_{\ell\in\{1,\ldots,C\}}\hat{P}(Y=\ell|\mathbf{X}=\mathbf{x})$.
Ces règles d'affectation sont quelques fois appelées règles MAP (Maximisation A Posteriori).
Nous substituons, au sigle MAP, dans ce cas précis et dans un but de clarification, le sigle MAPP (Maximisation A Postériori Probabiliste) ; Le sigle MAP sera réservé à la Maximisation A Postériori, qu'elle soit probabiliste ou autre.
Parmi les règles MAPP, citons celles dérivées de la discrimination gaussienne, de la discrimination logistique, de la discrimination non-paramétrique (Noyau uniforme, noyau gaussien, noyau d'Epanechnikov,...).

Les règles basées sur la minimisation de la dissimilarité aux groupes : Nous les appelons règles mDAP (minimisation de la Dissimilarité A Posteriori).
L'affectation à une classe se base sur la proximité à cette classe ou à ses représentants : Après avoir défini et estimé la dissimilarité $d(\mathbf{x},G_\ell),\ell=1...,C$ où G_ℓ est le $\ell^{\text{ème}}$ groupe, on affecte a la classe $\arg\min_{\ell\in\{1,\ldots,C\}}\hat{d}(\mathbf{x},G_\ell)$.
Parmi ces règles, citons l'analyse discriminante de Fisher où la dissimilarité aux groupes est définie par la mètrique de Mahalanobis (voir en section 5.1.1.), la méthode KNN (*K Nearest Neighbors* ou des K-plus proches voisins)...

Les règles découpant l'espace de description : Ces règles découpent l'espace de manière à séparer, au mieux, les observations de l'échantillon d'apprentissage, selon leur classe d'appartenance. L'espace est découpé en autant de zones (souvent des hyperplans) que classes. Parmi ces règles, citons les hyperplans séparateurs (SVM ou *Support Vectors Machines* ; Les hyperplans selon l'algorithme de Rosenblatt ;...), les arbres de décision...
On affecte un nouvel individu, à la classe représentée majoritairement par les individus de l'échantillon d'apprentissage, dans la zone de l'espace de description, la plus proche .

Notons que sans optimiser, de façon explicite, un critère de qualité du découpage, l'approche MAPP et l'approche mDAP aboutissent, elles aussi, à un découpage de l'espace de description $\mathcal{X}$. Aussi, il est possible d'établir le caractère MAPP d'une règle mDAP et inversement. Le choix de l'une ou l'autre des deux approches, est basé sur les seules commodités calculatoires et de formulation.

4.4 Notion de classifieur

Definition : Un classifieur MAP est une fonction $\phi(\mathcal{X}\to\mathbb{R}^C)$ ou

$$\mathbf{x}\to\phi(\mathbf{x})=(\phi_1(\mathbf{x}),...,\phi_C(\mathbf{x}))$$

associé à la règle d'affectation $D_\phi:\mathcal{X}\to\{1,...,C\}$ telle que

$$\forall\mathbf{x}\in\mathcal{X}\qquad D_\phi(\mathbf{x})=\arg\max_{\ell\in\{1,\ldots,C\}}\phi_\ell(\mathbf{x}).$$

Etant donné un profil individuel $\mathbf{x}$, $\phi_\ell(\mathbf{x})$ est le soutien apporté à la classe G_ℓ, par le classifieur ϕ.

Ici, $\phi_\ell(\mathbf{x})$ peut être

- une valeur de proximité ou de similarité (variant à inverse de la dissimilarité) de l'observation $\mathbf{x}$ au groupe G_ℓ,
- une valeur de probabilité *a posteriori* d'appartenance à la classe G_ℓ,
- un score ou note d'appartenance à G_ℓ, appelé, parfois, score d'Anderson.

On distingues trois types de classifieur `MAP` :

Les Classifieurs binaires : le classifieur binaire est un classifieur dont la sortie est un vecteur de composants 0 ou 1 indiquant directement l'appartenance *i.e.*,

$$\forall \mathbf{x} \in \mathcal{X} \qquad \phi_\ell(\mathbf{x}) \in \{0,1\} \qquad \text{et} \qquad \sum_{j=1}^{C} \phi_j(x) = 1.$$

Les Classifieurs probabilistes : le classifieur probabiliste est une fonction associant à la description, le vecteur de probabilité d'appartenance *i.e.*,

$$\phi_\ell(\mathbf{x}) = P(Y = \ell|\mathbf{x}), \quad \text{et donc} \qquad \sum_{j=1}^{C} \phi_j(\mathbf{x}) = 1.$$

Les classifieurs binaires constituent un cas particulier de classifieurs probabilistes.

Les Classifieurs possibilistes : Ici, on ne met pas de contraintes sur ce type de classifieurs *i.e.*,

$$\phi_\ell(\mathbf{x}) \in \mathbb{R}.$$

Ainsi, les classifieurs possibilistes constituent le cas le plus général. Mais l'on peut ramener ces classifieurs à des classifieurs probabilistes, en utilisant par exemple la transformation *softmax*, utilisée par certains spécialistes (voir DUDA et *al.* [2012])

$$\phi^*_\ell(\mathbf{x}) = \frac{\exp(\phi_\ell(\mathbf{x}))}{\sum_{j=1}^{C} \exp(\phi_j(\mathbf{x}))}.$$

Le choix de la fonction exponentielle est pour augmenter la variabilité, permettant une meilleure combinaison, dans le cas de plusieurs classifieurs.

Les trois types de classifieurs *i.e.*, binaires, probabilistes et possibilistes, peuvent être obtenus, à partir de règles `MAPP`, comme de règles `mAPG`.

4.5 Méthodes de classification supervisée

Les méthodes conventionnelles de classification supervisée ont, souvent, à estimer un unique classifieur. Cela est le cas pour

- l'analyse discriminante linéaire de Fisher,
- la discrimination gaussienne,
- la régression logistique,
- la méthode `KNN`,
- la méthodes des noyaux,

- les méthodes SIM semi-paramétriques.

Ces méthodes sont fondées sur des hypothèses, quant aux données et présentent, donc, des limites. Parmi ces limites, on retrouve la nature des données (Le type des covariables *i.e* continues ou nominales,...), la taille des données (nombre de covariables élevé, comme en génomique), le fait que les observations soient non identiquement distribuées, comme pour les modèles mélanges, les données stratifiées,...

La mise en oeuvre de ces méthodes, alors que les hypothèses ne sont pas vérifiées, conduit à des problèmes calculatoires (dégénérescence d'algorithme,...) et (ou) de validité des résultats. Pour pallier à cela, des méthodes, présentées comme plus adaptées, sont proposées.

Certaines méthodes visent à améliorer le seul classifieur existant, comme

- les *Support Vectors Machines* (SVM),
- la discrimination basée sur un mélange de modèles, du type MIXMOD,
- Le transfert de modèles ...

Les méthodes utilisant un seul classifieur, n'exploitent qu'une partie des covariables (ou une partie de l'information disponible). Or, les covariables non exploitées ou peu pondérées peuvent tout à fait bien prédire l'appartenance d'une partie des individus.

Ce constat motive l'introduction de méthodes combinant plusieurs "règles d'affectation" et pour obtenir une unique décision. Cela est, par exemple, le cas de la combinaison de classifieurs (voir KUNCHEVA [2004]) :

- Diviser l'espace de description des observations en sous-espaces, et sélectionner un classifieur local *i.e.*, le meilleur sur chacun de ces sous-espaces. Cette approche est appelée *sélection de classifieurs*. L'idée de cette méthode a été proposée par DASARATHY et SHEELA [1979] puis approfondie par RASTRIGIN et ERENSTEIN [1981] qui ont introduit la méthodologie actuellement utilisée ;

- Appliquer une règle d'agrégation combinant tous les classifieurs. Ici, on a trois types d'agrégation (voir KUNCHEVA [2004]) :
  - Agrégation sur classifieurs binaires, comme par exemple, le vote à la majorité, la règle de décision "naïve bayesienne" et la règle de décision bayesienne ;
  - Agrégation par classes séparées, comme par exemple : La règle du minimum, celles du maximum, de la moyenne, du produit et celle basée sur les intégrales floues ;
  - Agrégation Indifférente à la classe, comme par exemple la combinaison de Dempster-Schaffer (BEYNON et *al.* [2000]).

4.6 Références

BEYNON, M., B. CURRY et P. MORGAN. 2000, «The dempster–shafer theory of evidence : an alternative approach to multicriteria decision modelling», *Omega*, vol. 28, n° 1, p. 37–50. 78

BIERNACKI, C., F. BENINEL et V. BRETAGNOLLE. 2002, «A generalized discriminant rule when training population and test population differ on their descriptive parameters», *Biometrics*, vol. 58, nº 2, p. 387–397. 75

DASARATHY, B. V. et B. V. SHEELA. 1979, «A composite classifier system design : concepts and methodology», *Proceedings of the IEEE*, vol. 67, nº 5, p. 708–713. 78

DUDA, R. O., P. E. HART et D. G. STORK. 2012, *Pattern classification*, John Wiley & Sons. 77

KUNCHEVA, L. I. 2004, *Combining pattern classifiers : methods and algorithms*, John Wiley & Sons. 78

RASTRIGIN, L. et R. ERENSTEIN. 1981, «Method of collective recognition», *Energoizdat, Moscow*. 78

Chapitre 5

Méthodes conventionnelles de classification supervisée

Sommaire

Dans ce chapitre, on se limite à la présentation des méthodes conventionnelles de classification supervisée ou les plus rencontrées dans la littérature.
Selon les hypothèses sur la fonction $\mathbb{E}(Y|\mathbf{X}=x)$, on peut classifier les méthodes conventionnelles d'apprentissage, selon trois catégories : Les méthodes paramétriques, Les méthodes non-paramétriques et les méthodes semi-paramétriques.

5.1 Méthodes paramétriques

Ces méthodes sont basées sur l'hypothèse que la fonction $\mathbb{E}(Y|X=\mathbf{x})$ est déterminée par un ensemble de paramètres que l'on estime à partir des données de l'échantillon. Plusieurs méthodes paramétriques ont été proposées dans la littérature (Voir, par exemple, HASTIE *et al.* [2005] et BERGER *et al.* [2002]). L'analyse discriminante linéaire de Fisher, l'analyse discriminante gaussienne et la régression logistique comptent parmi les méthodes paramétriques les plus rencontrées dans la littérature et les plus utilisées par les praticiens.

5.1.1 Analyse discriminante de Fisher (1936)

L'idée de cette méthode est de choisir, parmi les combinaisons linéaires des covariables, celles qui maximisent l'homogénéité des classes. Le principe de l'analyse discriminante de Fisher est de projeter le vecteur de covariables $\mathbf{X}$ (de dimension p) sur un espace de dimension $C-1$ (C est supposé inférieur ou égal p). Les p covariables initiales (ou variables explicatives) sont remplacées par $C-1$ combinaisons linéaires des covaiables appellées fonctions discriminantes (ou scores)

$$g_\ell(\mathbf{x}) = \mathbf{w}_\ell^t \mathbf{x}, \qquad \ell = 1, ..., (C-1), \qquad \mathbf{w}_\ell \in \mathbb{R}^p.$$

Soit $\mathbf{W} = [\mathbf{w}_1|\mathbf{w}_2|...|\mathbf{w}_{C-1}]$, et convenons que $|.|$ désigne le déterminant ; l'analyse discriminante de Fisher revient à maximiser, selon la matrice $\mathbf{W}$, le critère

$$J(\mathbf{W}) = \frac{|\mathbf{W}^t S_B \mathbf{W}|}{|\mathbf{W}^t S_W \mathbf{W}|} \tag{5.1}$$

où $\mathbf{S}_B$ est la matrice de dispersion inter-classes (B comme *Between*) *i.e.,*

$$\mathbf{S}_B = \sum_{\ell=1}^{C} N_\ell (\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})^t,$$

et où S_W est la matrice de dispersion intra-classes (W comme *Within*) *i.e.,*

$$\mathbf{S}_W = \sum_{\ell=1}^{C} \sum_{i=1}^{n} \mathbb{1}_{\{y_i=\ell\}} (\mathbf{x}_i - \bar{\mathbf{x}}_\ell)(\mathbf{x}_i - \bar{\mathbf{x}}_\ell)^t,$$

avec

$$\bar{\mathbf{x}}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{n_T} \mathbf{x}_i \mathbb{1}_{\{Y_i=\ell\}}, \tag{5.2}$$

$$\bar{\mathbf{x}} = \frac{1}{n_T} \sum_{\ell=1}^{C} N_\ell \bar{\mathbf{x}}_\ell, \tag{5.3}$$

et N_ℓ le nombre d'individus dans la classe ℓ, parmi les n_T individus de l'échantillon d'apprentissage.

Cas de deux classes (C = 2) **:**

Dans le cas de deux classes, l'analyse discriminante de Fisher revient à projeter le vecteur de covariables $\mathbf{X}$ (de dimension p) en une unique combinaison $\mathbf{w}^t\mathbf{x}$ et le critère (5.1), à maximiser, devient

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}, \tag{5.4}$$

où $\mathbf{w}$ est un vecteur de $\mathbb{R}^p$. Ce quotient est connu, comme quotient de Rayleigh.
La solution, maximisant le quotient, est donnée, dans le cas ou $\mathbf{S}_W$ est inversible, par $\mathbf{w} = \mathbf{S}_W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$; plus généralement $\mathbf{w} = \mathbf{S}_W^{+}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ où $\mathbf{S}_W^{+}$ est l'inverse généralisé.

Cas de plus de deux classes (C > 2) **:**

Dans ce cas, la solution est obtenue par décomposition spectrale de $S_{\mathbf{W}}^{+}S_B$; $S_{\mathbf{W}}^{+}$ étant un inverse généralisé (Par exemple, celui de Moore-Penrose).

Remarquons que la solution $\mathbf{W}$ n'est pas unique , elle est déterminée à une homothétie près.

5.1.2 Discrimination gaussienne

Pour cette méthode de classification, on suppose que la loi de $\mathbf{X}$, conditionnellement à Y, est gaussienne *i.e.,* $\mathbf{X}|_{Y=\ell} \backsim \mathcal{N}_p(\mu_\ell, \boldsymbol{\Sigma}_\ell)$. La densité est donnée par

$$f_\ell(\mathbf{x}) = \frac{1}{(2\pi)^p |\boldsymbol{\Sigma}_\ell|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{x}-\mu_\ell)^t \boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\mu_\ell)),$$

avec $\mu_\ell = \mathbb{E}(\mathbf{X}|Y=\ell)$ et $\boldsymbol{\Sigma}_\ell = \mathbb{V}(\mathbf{X}|Y=\ell)$.
Par la formule de Bayes,

$$P(Y=\ell|\mathbf{X}=\mathbf{x}) = \frac{\pi_\ell f_\ell(\mathbf{x})}{\sum_{j=1}^{C} \pi_j f_j(\mathbf{x})}, \qquad \ell = 1,...,C,$$

et $\quad \pi_\ell = P(Y=\ell)$.

Pour comparer la vraisemblance de l'appartenance, à la classe G_l, et l'appartenance à la classe de référence G_C, il suffit du log-rapport

$$\log\frac{P(Y=l|\mathbf{X}=\mathbf{x})}{P(Y=C|\mathbf{X}=\mathbf{x})} = \log\frac{\pi_\ell}{\pi_C} - \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_\ell|}{|\boldsymbol{\Sigma}_C|} - \frac{1}{2}(\mu_\ell^t \boldsymbol{\Sigma}_\ell^{-1}\mu_\ell - \mu_C^t \boldsymbol{\Sigma}_C^{-1}\mu_C) + \mathbf{x}^t(\mu_\ell^t \boldsymbol{\Sigma}_\ell^{-1} - \mu_C^t \boldsymbol{\Sigma}_C^{-1}) + \\ + \frac{1}{2}\mathbf{x}^t(\boldsymbol{\Sigma}_C^{-1} - \boldsymbol{\Sigma}_\ell^{-1})\mathbf{x}. \tag{5.5}$$

L'individu décrit par $\mathbf{x}$ est affecté au groupe G_{l^*} maximisant le log-rapport ou *logit*. La fonction de discrimination ϕ_l, associée au groupe G_l, est donnée par

$$\phi_\ell(\mathbf{x}) = \beta_{0\ell} + \boldsymbol{\beta}_{\mathbf{1}\boldsymbol{\ell}}^{\boldsymbol{t}}\mathbf{x} + \mathbf{x}^t\boldsymbol{\beta}_{\mathbf{2}\boldsymbol{l}}\mathbf{x}, \qquad l = 1,\ldots,C, \tag{5.6}$$

où

$$\begin{aligned}
\beta_{0\ell} &= \log\frac{\pi_\ell}{\pi_C} - \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_\ell|}{|\boldsymbol{\Sigma}_C|} - \frac{1}{2}(\mu_\ell^t\boldsymbol{\Sigma}_\ell^{-1}\mu_\ell - \mu_C^t\boldsymbol{\Sigma}_C^{-1}\mu_C), && (5.7)\\
\boldsymbol{\beta_{1\ell}} &= \boldsymbol{\Sigma}_\ell^{-1}\mu_\ell - \boldsymbol{\Sigma}_C^{-1}\mu_C, && (5.8)\\
\boldsymbol{\beta_{2l}} &= \frac{1}{2}(\boldsymbol{\Sigma}_C^{-1} - \boldsymbol{\Sigma}_\ell^{-1}). && (5.9)
\end{aligned}$$

Notons que $\phi_C(\mathbf{x}) = 0$, par définition de la fonction *logit*.
Cette fonction de discrimination est une fonction quadratique en $\mathbf{x}$, c'est pourquoi on parle, dans ce cas, d'analyse discriminante quadratique (QDA).
Nous pouvons substituer, au vecteur de covariables $\mathbf{X} = (X^1, ..., X^p)$, le vecteur $\tilde{\mathbf{X}}$ ayant pour composantes les covariables de $\mathbf{X}$, et des covariables construites à partir de $\mathbf{X}$, comme le carré des covariables et leurs produits deux à deux ; Ce qui augmente la dimension de l'espace de description.

Ainsi, l'égalité (5.5) peut être mise sous la forme

$$\log\frac{P(Y = l|\mathbf{X} = \mathbf{x})}{P(Y = C|\mathbf{X} = \mathbf{x})} = \beta_{0\ell} + \beta_{1\ell}^t\tilde{\mathbf{x}}. \qquad (5.10)$$

Cela signifie que la discrimination QDA peut être vue comme une discrimination linéaire, mais avec comme vecteur de description $\tilde{\mathbf{X}}$.
En réalité, les paramètres des loi gaussiennes, en présence, ne sont pas connus. On peut les estimer, selon la méthode du maximum de vraisemblance, par

$$\begin{aligned}
\hat{\pi}_l &= N_\ell / n_T,\\
\hat{\mu}_\ell &= \frac{1}{N_\ell}\sum_{i=1}^{n_T}\mathbf{x}_i\mathbb{1}_{\{y_i=l\}},\\
\hat{\boldsymbol{\Sigma}}_\ell &= \frac{1}{n-C}\sum_{i=1}^{n_T}(\mathbf{x}_i - \hat{\mu}_\ell)(\mathbf{x}_i - \hat{\mu}_\ell)^t\mathbb{1}_{\{y_i=l\}}.
\end{aligned}$$

où N_ℓ est le nombre d'individus dans la classe G_l parmi les n_T individus de l'échantillon d'apprentissage. Ces estimateurs ne sont valables que si l'échantillon d'apprentissage est en randomisation totale ; Ils ne sont pas adaptés au cas d'un échantillon rétrospectif, par exemple.

L'analyse discriminante gaussienne, dans le cas homoscédastique *i.e.*, $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$ pour tout $\ell = 1, ..., C$, se ramène au cas linéaire ou LGDA, pour faire la différence avec l'analyse discriminante de Fisher ou LDA.

Ainsi,

$$\log\frac{P(Y = l|\mathbf{X} = \mathbf{x})}{P(Y = C|\mathbf{X} = \mathbf{x})} = \log\frac{\pi_\ell}{\pi_C} - \frac{1}{2}(\boldsymbol{\mu}_\ell + \boldsymbol{\mu}_C)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_C) + \mathbf{x}^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_C).$$

Les fonctions de discrimination, de l'analyse LGDA, sont données par

$$\begin{aligned}
\phi_\ell(\mathbf{x}) &= \beta_{0\ell} + \boldsymbol{\beta_{1\ell}^t}\mathbf{x}\\
\text{avec } \beta_{0\ell} &= log\frac{\pi_\ell}{\pi_C} - \frac{1}{2}(\mu_\ell + \mu_C)^t\boldsymbol{\Sigma}^{-1}(\mu_\ell - \mu_C)\\
\text{et } \boldsymbol{\beta_{1\ell}} &= (\mu_\ell - \mu_C)^t\Sigma^{-1}
\end{aligned}$$

On estime $\boldsymbol{\Sigma}$ par $\hat{\boldsymbol{\Sigma}} = \frac{1}{C}\sum_{j=1}^{C}\hat{\boldsymbol{\Sigma}}_j$.
L'affectation pour un individu, décrit par $\mathbf{x}^*$, dans la discrimination gaussienne est donnée par $\hat{y} = \arg\max_\ell \phi_\ell(\mathbf{x}^*)$.

5.1.3 Régression logistique

La régression logistique postule l'hypothèse, que la loi de $\mathbf{X}$ est telle que $\phi_\ell(\mathbf{x}) = \log\dfrac{P(Y=\ell|\mathbf{X}=\mathbf{x})}{P(Y=C|\mathbf{X}=\mathbf{x})}$ est linéaire en $\mathbf{x}$ *i.e.,*

$$\phi_\ell(\mathbf{x}) = \beta_{0\ell} + \boldsymbol{\beta}_{1\ell}^t\mathbf{x}, \quad \beta_{ol} \in \mathbb{R}, \quad \boldsymbol{\beta}_{1\boldsymbol{\ell}} \in \mathbb{R}^p \text{ et } \qquad \ell = 1,...,C-1. \tag{5.11}$$

La discrimination gaussienne peut être vue comme cas particulier de la régression logistique ; Plus généralement, cela est le cas lorsque la loi de $\mathbf{X}$, conditionnellement à Y, est de la famille exponentielle. ANDERSON [1982] et VENDITTI [1998]) caractérisent les modèles conditionnels, vérifiant l'équation (5.11).

Soit $\boldsymbol{\beta}_\ell \simeq (\beta_{0\ell}||\boldsymbol{\beta}_{1\ell})$ le vecteur (p+1)-dimensionnel obtenu par concaténation de $\beta_{0\ell}$ et du p-vecteur $\boldsymbol{\beta}_{1\ell}$.

De (5.11), on a

$$P(Y=l|\mathbf{X}=\mathbf{x}) = P(Y=C|\mathbf{X}=\mathbf{x}).exp(\beta_{o\ell} + \boldsymbol{\beta}_\ell^t\mathbf{x}), \text{ pour } \ell = 1,...,C-1. \tag{5.12}$$

Comme

$$P(Y=C|\mathbf{X}=\mathbf{x}) + \sum_{\ell=1}^{C-1} P(Y=\ell|\mathbf{X}=\mathbf{x}) = 1. \tag{5.13}$$

Alors, de (5.12) et (5.13), on obtient

$$P(Y=\ell|\mathbf{X}=\mathbf{x}) = \frac{exp(\beta_{o\ell} + \boldsymbol{\beta}_\ell^t\mathbf{x})}{1+\sum_{j=1}^{C-1} exp(\beta_{o\ell} + \boldsymbol{\beta}_j^t\mathbf{x})}, \tag{5.14}$$

$$P(Y=C|\mathbf{X}=\mathbf{x}) = \frac{1}{1+\sum_{j=1}^{C-1} exp(\beta_{o\ell} + \boldsymbol{\beta}_j^t\mathbf{x})}. \tag{5.15}$$

Cas de deux classes (C = 2) **:**

Dans ce cas, on a un seul vecteur de paramètres *i.e.,* $\boldsymbol{\beta}_\ell \simeq (\beta_0\|\boldsymbol{\beta}) \in \mathbb{R}^{p+1}$.

$$\begin{aligned} P(Y=1|\mathbf{X}=\mathbf{x}) &= \frac{exp(\beta_0 + \boldsymbol{\beta}^t\mathbf{x})}{1+exp(\beta_0 + \boldsymbol{\beta}^t\mathbf{x})}; \\ P(Y=0|\mathbf{X}=\mathbf{x}) &= \frac{1}{1+exp(\beta_0 + \boldsymbol{\beta}^t\mathbf{x})}. \end{aligned}$$

On estime β_0 et $\boldsymbol{\beta}$, en résolvant le problème de programmation non-linéaire consistant en la maximisation de la vraisemblance conditionnelle

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i \log P(y_i=1|\mathbf{x}_i) + (1-y_i)\log(1-P(y_i=0|\mathbf{x}_i)))); \\ &= \sum_{i=1}^{n}(y_i\boldsymbol{\beta}^t\mathbf{x}_i - \log(1+exp(\beta_0 + \boldsymbol{\beta}^t\mathbf{x})). \end{aligned}$$

On utilise l'algorithme de Newton ou l'une de ses variantes. L'affectation, d'un individu décrit par $\mathbf{x}^*$, est donnée par $\hat{y} = \mathbb{1}_{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^t \mathbf{x}^* > s}$, où s est le seuil fixé par le praticien ou utilisateur, selon qu'il veuille optimiser la sensibilité ou la spécificité.

Ce qui signifie qu'on affecte à la classe G_1 si le score $\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^t \mathbf{x}^*$ est strictement supérieur à s, ou à la classe G_2 si ce score est strictement inférieur à s. Ainsi, en classification logistique binaire, une seule fonction score détermine la règle d'affectation.

Cas de plus de deux classe ($C > 2$) **:**

Soientt $\hat{\beta}_0$ et $\hat{\boldsymbol{\beta}}_\ell$ respectivement, les estimateurs de β_0 et $\boldsymbol{\beta}_\ell$, en (5.15), obtenus en maximisant la vraisemblance conditionnelle. Pour un individu décrit par $\mathbf{x}^*$, on a $C-1$ scores $\hat{\beta}_0 + \boldsymbol{\beta}_1^t \mathbf{x}^*, ..., \hat{\beta}_0 + \boldsymbol{\beta}_{C-1}^t \mathbf{x}^*$. On affecte à la $C^{\text{ème}}$ classe si tous les scores sont strictement négatifs ; Sinon on l'affecte à la $\ell^{\text{ème}}$ classe, correspondant au score le plus élevé.

5.2 Méthodes non-paramétriques

Ces méthodes sont basées sur l'hypothèse que l'espérance conditionnelle $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ est une fonction lisse et aucune autre hypothèse, quant à sa forme, n'est postulée. Plusieurs méthodes non-paramétriques ont été proposées, dans le contexte de la classification, comme par exemple, la méthode des K-*plus proches voisins*, méthodes des noyaux, les réseaux de neurones, les SVM (*Support Vectors Machine*),... On peut, par exemple, se référer à HASTIE et *al.* [2005], pour une revue synthétique et plus exhaustive.
Les méthodes non paramétriques peuvent s'avérer pertinentes, en l'absence d'hypothèses strictes, sur les données utilisées. Cependant, elles souffrent de défauts sévères, parmi lesquels, le fait que la précision de l'estimation diminue rapidement, quand la dimension du vecteur de covariables augmente (*curse of dimensionality* ou fléau des grandes dimensions). Un autre défaut sévère, est que ces méthodes ne donnent pas un estimateur de $f_{|Y}(x)$, pour les points x qui se trouvent en dehors du support du vecteur de covariables $\mathbf{X}$.

5.2.1 Méthode des *K plus proches voisins*

La méthode des K-*plus proches voisins* se base sur la répartition des K voisins, dans les classes. Les K-*plus proches voisins* sont déterminés en fonction du choix de la mesure de dissimilarité (Distance Euclidienne, Distance L_1, ...). Dans le cas binaire,

$$\hat{f}_1(x) = \frac{1}{K} \sum_{x_i \in V_K(x)} y_i \qquad \text{et} \qquad \hat{f}_0(x) = 1 - \hat{f}_1(x) \tag{5.16}$$

où $V_K(x)$ est le voisinage de x défini par les K plus proches points x_i de notre échantillon d'apprentissage.
Dans le cas multinomial,

$$f_\ell(x) = \frac{1}{K} \#\{X_i \in V_K(x) : y_i = l\}, \qquad \sum_{\ell=1}^{K} f_\ell(x) = 1.$$

D'après la formule (5.16), il est évident que cette méthode dépend d'un seul paramètre qui est le nombre des voisins K. Une petite valeur de K réduit le biais de l'estimateur mais augmente sa variance ; En revanche, une grande valeur réduit sa variance mais augmente

son biais. C'est pourquoi, il faut choisir K permettant un compromis biais-variance ; Pour cela, on peut utiliser, par exemple, la méthode de validation croisée pour choisir K.
La règle d'affectation pour un nouvel individu x^*, est donnée par : $\hat{y} = \arg\max_\ell \hat{f}_\ell(x)$.

5.2.2 Méthodes des noyaux

D'après la formule (5.16) de la méthode des *k−plus proches voisins*, il est clair que cet estimateur est une fonction discontinue en x car la moyenne dans la formule change d'une manière discrète. Cette propriété indésirable conduit à modifier la formule (5.16) de la façon suivante : Au lieu de donner, à tous les voisins de x, des poids égaux, on leur attribue des poids qui s'affaiblissent doucement avec la distance du point cible. De tels poids sont des noyaux, en général, des densités de probabilité symétriques.
La méthode la plus utilisée est la méthode de Nadaraya-Watson :

$$\hat{f}(x) = \sum_{i=1}^{n} W_i(x) y_i \tag{5.17}$$

où $W_i(x) = \frac{K_h(x_i - x)}{\sum_{j=1}^{n} K_h(x_i - x)}$ est la fonction de poids, $K_h(.) = K(./h)$ avec K un noyau symétrique, h est un paramètre de lissage.

D'une façon plus générale, l'estimateur de la régression locale linéaire $\hat{f}_\theta(x)$ est la solution du problème d'optimisation suivant :

$$\min_\theta \sum_{i=1}^{n} \big(y_i - f_\theta(x_i)\big)^2 K_h(x_i - x) \tag{5.18}$$

où $f_\theta(x)$ est une fonction connue, dependant du paramètre θ. Les deux cas, les plus utilisés dans la littérature, sont

— $f_\theta(\mathbf{x}) = \theta$, la fonction constante ($\theta \in \mathbb{R}$). Là, on obtient l'estimateur de Nadaraya-Watson qui est déjà expliqué(la régression constante),

— $f_{\boldsymbol{\theta}}(\mathbf{x}) = \theta_0 + \boldsymbol{\theta}_1^t \mathbf{x}$ ($\theta = (\theta_0 || \boldsymbol{\theta}_1)$, avec $\theta_0 \in \mathbb{R}$ et $\boldsymbol{\theta}_1 \in \mathbb{R}^p$) qui correspond à la régression locale linéaire.

Le paramètre le plus important, déterminant la qualité de l'estimateur *plug in* $\hat{f} = f_{\hat{\theta}}$, est le paramètre de lissage h, appelé aussi largeur de la fenêtre. Cela, parce qu'il permet le contrôle du nombre d'observations, utilisées pour construire l'estimateur : Si n_T est le nombre d'observations (au total), alors, seulement, hn_T observations sont utilisées pour construire l'estimateur. Plusieurs méthodes ont été proposées pour estimer la fenêtre optimale h, voir par exemple LI et RACINE [2007], pour une revue détaillée. Dans la pratique, la méthode des noyaux n'est utilisée que si la dimension de la variable explicative $\mathbf{X}$ est inférieur ou égale 2. Dans le cas d'une dimensionalité plus élevée, la précision de l'estimateur $\hat{f}(x)$ devient trop faible et donc l'estimateur est inefficient, dans ce cas.

5.3 Méthodes semi-paramétriques

Les méthodes paramétriques sont basées sur des hypothèses quant à la fonction $G(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, comme en régression linéaire, sur la loi conditionnelle de $\mathbf{X}$ comme en discrimination gaussienne ou sur le rapport des probabilités d'appartenance entre deux groupes,

comme en discrimination logistique. En réalité, ces hypothèses peuvent être non valides.

En revanche, les méthodes non-paramétriques ne postulant pas ces hypothèses sont attractives quand on n'a pas suffisamment d'informations sur les données. Cependant, la précision de l'estimateur s'altère rapidement quand la dimension de $\mathbf{X}$, augmente.

Une possibilité, pour surmonter les défauts des deux précédentes approches, est de proposer une approche intermédiaire. Plusieurs tentatives ont été proposées, dans ce sens. On mentionne, par exemple, les modèles linéaires partiels (*Varying coefficient models*) et les modèles à indice unique (*Single Index Models*) ou modèles `SIM`. Ces méthodes réduisent le risque de mauvaise spécification de méthodes paramétriques et évitent les défauts mentionnés. Pour une revue plus détaillée, on renvoie, par exemple, à LI et RACINE [2007].

Dans ce chapitre, on s'intéresse aux modèles à score unique ou modèles `SIM` qui sont les plus rencontrés dans la littérature. Ces modèles se caractérisent par

$$\mathbb{E}(\mathrm{Y}|\mathbf{X}=\mathrm{x}) = \mathrm{G}(\mathrm{x}^t\boldsymbol{\beta}) \tag{5.19}$$

où Y est la variable réponse, $\mathbf{X} \in \mathbb{R}^d$ le vecteur de covariables, G une fonction lien inconnue et $\boldsymbol{\beta} \in \mathbb{R}^p$ le vecteur de coefficients associés aux covariables.

Dans le cas des méthodes semi-paramétriques, on se place dans la situation où G est inconnue. G et $\boldsymbol{\beta}$ peuvent être estimés à partir des observations de S_T, pour $(\mathbf{X}, \mathrm{Y})$. Le modèle (5.19) est semi-paramétrique car d'une part, la forme de l'indice est spécifiée et d'autre part, la forme de la fonction G ne l'est pas.

Dans le cas où la variable réponse Y est binaire *i.e.*, $\mathrm{Y} \in \{0,1\}$, le lien avec les modèle `SIM` est plus explicite : On suppose qu'on a une variable latente Y^* reliée à $\mathbf{X}$ par la relation

$$\mathrm{Y}^* = \boldsymbol{\beta}^t\mathbf{X} + u \tag{5.20}$$

où le terme d'erreur u est indépendant de $\mathbf{X}$. La variable Y^* est construite de manière que $\mathrm{Y} = \mathbb{1}_{\{\mathrm{Y}^*>0\}}$. Alors

$$\begin{aligned}
\mathbb{E}(\mathrm{Y}|\mathbf{X}=\mathrm{x}) &= \mathrm{P}(\mathrm{Y}=1|\mathbf{X}=\mathrm{x}) && (5.21)\\
&= \mathrm{P}(\mathrm{Y}^* \geq 0|\mathbf{X}=\mathrm{x}) && (5.22)\\
&= \mathrm{P}(u \geq -\boldsymbol{\beta}^t\mathbf{X}|\mathbf{X}=\mathrm{x})\\
&= 1-\mathrm{F}(-\boldsymbol{\beta}^t\mathbf{x})\\
&= \mathrm{G}(\boldsymbol{\beta}^t\mathrm{x}) && (5.23)
\end{aligned}$$

où F est la fonction de répartition de u.
Lorsque la loi de u est symétrique, $\mathrm{F}(\beta^t\mathrm{x}) = 1-\mathrm{F}(-\boldsymbol{\beta}^t\mathrm{x})$ et dans ce cas $\mathrm{G}(.) = \mathrm{F}(.)$, comme dans le modèle *Probit* où la loi de u est la loi gaussienne standard *i.e.*, $\mathbf{u} \sim \mathcal{N}(0,1)$.

5.4 Références

ANDERSON, J. 1982, «Logistic regression», *Handbook of Statistics. North-Holland, New York*, p. 169–191. 85

BERGER, A. N., W. S. FRAME et N. H. MILLER. 2002, «Credit scoring and the availability, price, and risk of small business credit», . 82

HASTIE, T., R. TIBSHIRANI, J. FRIEDMAN et J. FRANKLIN. 2005, «The elements of statistical learning : data mining, inference and prediction», *The Mathematical Intelligencer*, vol. 27, nº 2, p. 83–85. 82, 86

LI, Q. et J. S. RACINE. 2007, *Nonparametric econometrics : theory and practice*, Princeton University Press. 87, 88

VENDITTI, V. 1998, «Aspects du principe de maximum d'entropie en modélisation statistique», *Thèse de doctorat, Université Grenoble 1*. 85

Chapitre 6

Transfert semi-paramétrique d'un modèle SIM

Sommaire

6.1 Introduction

Dans ce chapitre, nous nous intéressons à l'adaptation ou la mise à jour de règles de classification binaire pour une structure de données, particulière : L'échantillon d'apprentissage provient d'une certaine sous-population et l'échantillon de prédiction, d'une autre.

Les modèles semi-paramétriques *single index models* ou `SIM` résument les effets des covariables, composant $\mathbf{X}$, en une seule variable appelée indice ou score pour certains spécialistes. Dans ces modèles, la fonction moyenne conditionnelle, a la forme

$$\mathbb{E}(Y|\mathbf{X}=\mathbf{x}) = G(\boldsymbol{\beta}^t\mathbf{x}) \,, \tag{6.1}$$

où $\boldsymbol{\beta}$ est un vecteur de dimension p, de paramètres réels et $G(\mathbb{R}\to\mathbb{R})$ une fonction réelle. Ces modèles impliquent que toute l'information portée par $\mathbf{X}$, est résumée par une combinaison linéaire des composantes de $\mathbf{X}$, à savoir le score ou indice $\boldsymbol{\beta}^t\mathbf{x}$.

Étant données les estimations de $\boldsymbol{\beta}$ et $G(.)$, l'estimation de la moyenne conditionnelle s'obtient, en utilisant l'équation (6.1).

Dans ce chapitre, la variable réponse Y est binaire. Le but est de prédire la valeur de Y pour un nouvel individu pour lequel l'appartenance à l'un ou l'autre des deux groupes $\mathbf{x}=(x^1,\ldots,x^p)^t$ est connu.
Plusieurs exemples d'un tel problème sont disponibles, comme en *credit-scoring* , où il s'agit prédire le comportement des emprunteurs, à rembourser un prêt, en utilisant des informations historiques, relatives à des clients-emprunteurs ayant fini de rembourser ou en défaut. Un autre exemple, en médecine, où il s'agit de prédire le risque de récidive du cancer du poumon, chez un patient préalablement traité, sur la base d'informations relatives au traitement utilisé pour la première occurrence du cancer, les mesures cliniques et épidémiologiques.

Le problème, en apprentissage supervisé, est que tout individu à prédire est supposé provenir de la même population statistique que l'échantillon d'apprentissage. Ces hypothèses ne sont pas réalistes, en général. Par exemple, en *credit-scoring*, pour prédire le comportement des non-clients, on utilise un échantillon d'apprentissage formé de clients seulement. Aussi, en médecine, le risque de récidive du cancer du poumon est appris, à partir de patients européens et sera estimé pour des patients asiatiques.

Afin de pallier au problème de la taille faible des données de l'échantillon d'apprentissage disponible, nous préconisons une méthode qui vise à transférer les connaissances acquises sur sous-population source, à la sous-population cible (sous-population contenant les individus à prédire).

Cette idée a été initroduite par BIERNACKI et *al.* [2002] et developpée dans le contexte gaussien : la distribution de $\mathbf{X}$ sur la sous-population source (des oiseaux d'une certaine sous-espèce, de la famille des petrels ou *Procellariidae*) et la distribution sur la sous-population cible (des pétrels d'une autre sous-espèce) sont gaussiennes.

Un lien stochastique de type linéaire, entre les deux restrictions, est postulé. Ainsi, la règle d'affectation résultant d'un modèle de transfert (à appliquer sur l'échantillon non classé) est obtenue par l'estimation des paramètres de cette liaison linéaire, en utilisant plusieurs

cas de contraintes sur cette liaison.

Cette technique donne lieu à des meilleures performances que les méthodes classiques. **?** étend cette approche à la classification logistique multinomiale, avec des modèles de lien supplémentaires.
BENINEL et *al.* [2012a], dans leur ouvrage, utilisent le transfert d'apprentissage dans les domaines de la classification probabiliste, la régression linéaire (y compris le mélange de régressions) et le clustering à base de modèle gaussien et de Student.

Le modèle semi-paramétrique *Single Index* (`SIM`) en classification a une supériorité potentiel par rapport aux méthodes classiques de classification. Nous développons l'idée du transfert d'apprentissage en modèle `SIM`, comme dans le travail de BENINEL et *al.* [2012b].

Les données se composent de deux échantillons : Le premier $\mathcal{S} = \{(\mathrm{Y}_1, \mathrm{X}_1), ..., (\mathrm{Y}_n, \mathbf{X}_n)\}$ de n points tirés d'une sous-population source $\mathcal{U}$ et le deuxième est $\mathcal{S}^* = \{(\mathrm{Y}_1^*, \mathbf{X}_1^*), ..., (\mathrm{Y}_{n^*}^*, \mathbf{X}_{n^*}^*)\}$ de n^* points tirés d'une sous-population cible $\mathcal{U}^*$. Ici, l'idée du transfert est d'affecter les individus de la sous-population cible, en utilisant deux échantillons $\mathcal{S}$ et $\mathcal{S}^*$.

6.2 Transfert gaussien

On suppose que la loi de $\mathbf{X}|_{\mathrm{Y}=k}$ est gaussienne de moyenne μ_k et de matrice de variance Σ_k *i.e.*, $\mathbf{X}|_{\mathrm{Y}=k} \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k)$, $k = 1, ..., \mathrm{C}$ et Y distribué selon la loi multinomiale $\mathcal{M}(1, p_1, ..., p_\mathrm{C})$, avec $p_k = \mathrm{P}(\mathrm{Y} = k)$ la probabilité *a priori*, d'appartenance au groupe k.

Dans le contexte gaussien, BIERNACKI et *al.* [2002] proposent une méthode de transfert du modèle appris sur la population $\mathcal{U}$ à la population $\mathcal{U}^*$, basée sur les hypothèses ci-après :

$$\mathbf{X}^*|_{\mathrm{Y}^*=k} \sim \mathcal{N}_p(\boldsymbol{\mu}_k^*, \Sigma_k^*), \quad \mathrm{Y}^* \sim \mathcal{M}_\mathrm{K}(1, p_1^*, ..., p_\mathrm{C}^*),$$

$$\text{et} \quad \mathbf{X}^*|_{\mathrm{Y}^*=k} \sim \Lambda_k \mathbf{X}|_{\mathrm{Y}=k}.$$

Ce lien donne la relation entre paramètres de la population source $\mathcal{U}$ et ceux de la population cible $\mathcal{U}^*$, ci-après :

$$\boldsymbol{\mu}_k^* = \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k, \tag{6.2}$$
$$\boldsymbol{\Sigma}_k^* = \boldsymbol{\Lambda}_k \Sigma_k \Lambda_k, \text{ avec } k = 1, ..., \mathrm{C}, \tag{6.3}$$

où $\boldsymbol{\Lambda}_k$ est une matrice diagonale de $\mathbb{M}_p(\mathbb{R})$.

On étudie, les cinq scénarios, ci-après :

M_0 : $\boldsymbol{\Lambda}_k = \mathbf{I}_p$: Les deux sous-populations sont distribuées à l'identique *i.e.*, il n'y a pas de transfert.

M_1 : $\boldsymbol{\Lambda}_k = \alpha \mathbf{I}_p$, $\alpha \in \mathbb{R}$: La transformation est indépendante des groupes et des covariables.

M_2 : $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}$: La transformation est indépendante des groupes, mais pas des covariables ; La transformation est identique pour toutes les covariables.

M_3 : $\boldsymbol{\Lambda}_k = \lambda_k \mathbf{I}_p$, $\lambda_k \in \mathbb{R}$: La transformation est dépendante des groupes, mais pas des covariables.

M_4 : C'est le cas le plus général *i.e.*, aucune contrainte sur les matrices $\boldsymbol{\Lambda}_k$.

L'estimation des paramètres est réalisée selon que les probabilités *a priori* $p_1, ..., p_C$ sont connues ou à estimer. La méthode d'estimation utilisée, valable pour les cinq modèles, est la méthode du maximum de vraisemblance. La méthode des moindres carrées, pourrait être utilsée pour estimer les paramètres de liaison, lorsque celle-ci est indépendante des groupes, comme dans le cas des modèles M_1 et M_2.

6.3 Transfert logistique

BENINEL et BIERNACKI [2009] et BENINEL et *al.* [2012b] étendent la méthodologie de transfert, en classification logistique s'inspirant de travail de BIERNACKI et *al.* [2002].
On suppose que la variable réponse dans les deux sous-populations $\mathcal{U}$ et $\mathcal{U}^*$ est binaire et que le modèle statistique dans la population source $\mathcal{U}$ est logistique, de vecteur de paramètres $\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}$ tandis que le modèle dans la population cible est logistique, de vecteur de paramètres $\begin{pmatrix} \beta_0^* \\ \boldsymbol{\beta}_1^* \end{pmatrix}$.

BENINEL et BIERNACKI [2007] montre l'existence d'un lien entre les paramètres du modèle gaussien et celui de modèle logistique.
A titre illustratif, dans le cas binaire gaussien homoscédastique *i.e.*, $\mathbf{X}|_{Y=k} \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k)$, $\mathbf{X}^*|_{Y^*=k} \sim \mathcal{N}_p(\boldsymbol{\mu}_k^*, \Sigma_k^*)$ $(k = 1, 2.)$ où $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ et $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_1^* = \boldsymbol{\Sigma}_2^*$, on obtient alors les liens suivants, entre les paramètres du modèle logistique et les paramètres du modèle gaussien, pour les deux sous-populations $\mathcal{U}$ et $\mathcal{U}^*$:

$$\beta_0 = \frac{1}{2}(\boldsymbol{\mu}_2^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1) \quad , \quad \boldsymbol{\beta}_1 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \; ; \tag{6.4}$$

$$\beta_0^* = \frac{1}{2}(\boldsymbol{\mu}_2^{*\,t} \boldsymbol{\Sigma}^{*-1} \boldsymbol{\mu}_2^* - \boldsymbol{\mu}_1^{*\,t} \boldsymbol{\Sigma}^{*-1} \boldsymbol{\mu}_1^*) \quad , \quad \boldsymbol{\beta}_1^* = \boldsymbol{\Sigma}^{*-1}(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*). \tag{6.5}$$

En substituant les expressions de $\boldsymbol{\mu}_k^*$ et $\boldsymbol{\Sigma}_k^*$ données en (6.2) et (6.3), on obtient alors le lien de transfert *via* les paramètres de $\mathcal{U}$ et ceux de $\mathcal{U}^*$:

$$\beta_0^* = \alpha + \beta_0 \quad \text{et} \quad \boldsymbol{\beta}_1^* = \boldsymbol{\Lambda}\boldsymbol{\beta}_1,$$

où α est un réel et $\boldsymbol{\Lambda}$ est une matrice carrée diagonale de taille p.
On distingue les cas de contraintes, sur α et $\boldsymbol{\Lambda}$, ci-après :

- M_0 : Pas de paramètre à estimer : $\alpha = 0$ et $\boldsymbol{\Lambda} = \mathbf{I}_p$, où $\mathbf{I}_p$ la matrice d'identité.
- M_1 : Ici, nous n'avons que α à estimer, $\boldsymbol{\Lambda} = \mathbf{I}_p$.
- M_2 : $\alpha = 0$ et $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p$, où $\lambda \in \mathbb{R}$.
- M_3 : α est libre et $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p$ *i.e.*, deux paramètres sont à estimer.
- M_4 : $\alpha = 0$ et $\Lambda = \mathbf{Diag}(\lambda_1, ..., \lambda_p)$, où $\lambda_1, ..., \lambda_p \in \mathbb{R}$ sont à estimer.
- M_5 : Le modèle le plus complexe : On estime α et $\boldsymbol{\Lambda} = \mathbf{Diag}(\lambda_1, ..., \lambda_p)$.

Dans tous ces cas, les paramètres peuvent être estimés en utilisant la méthode du maximum de vraisemblance.

6.4 Modèle SIM semi-paramétrique

Le modèle SIM semi-paramétrique a la forme

$$Y = G(\boldsymbol{\beta}^t X) + \epsilon, \qquad \boldsymbol{\beta} \in \mathbb{R}^p \tag{6.6}$$

avec Y la variable réponse et ϵ l'erreur telle que $\mathbb{E}(\epsilon|\mathbf{X}) = 0$.
L'estimation de la fonction G, sera réalisée selon une approche non-paramétrique.

Dans un but d'identifiabilité des estimateurs de $\boldsymbol{\beta}$ et G, on suppose que la fonction G n'est pas constante et que $\mathbf{X}$ ne contient pas de covariable constante *i.e.*, $\boldsymbol{\beta}$ ne contient pas d'*intercept*.

Aussi, on suppose que $\mathbf{X}$ contient au moins une covariable continue, dont le coefficient, composante de $\boldsymbol{\beta}$ associée, est non nul. Et enfin, on fixe le coefficient d'une des covariables de $\mathbf{X}$, comme étant égale à un (la covariable est dite *offset*).
Ce problème d'identifiabilité a été étudié, pour l'estimation par maximum de vraisemblance, par MANSKI [1988] et KLEIN et SPADY [1993], dans la cas de la classification binaire.
ICHIMURA [1993] a traité ce problème, pour l'estimation par la méthode des moindres carrés, pour une variable Y quelconque *i.e.,* continue ou discrète.
DELECROIX et *al.* [2003] étendent l'idée de KLEIN et SPADY [1993], dans le contexte d'estimation par maximum de vraisemblance et pour une variable réponse Y quelconque.
DELECROIX et *al.* [2006] traitent une grande classe de modèles SIM semi-paramétriques, selon une approche de M-estimation.
PATILEA [2007] introduit l'utilisation du modèle SIM semi-paramétrique, comme alternative aux modèles paramétriques classiques, en classification supervisée.
Récemment, KOMAROVA [2013] étudie ce problème, dans le cas de la classification binaire, pour le cas où toutes les covariables sont discrètes.

Ici, la variable réponse Y est binaire. On utilise l'approche M-estimation, pour déterminer l'estimateur de $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n_T} \psi(Y_i, \hat{G}(\boldsymbol{\beta}^t X_i; \boldsymbol{\beta}))\tau_{n_T}(X_i) \tag{6.7}$$

où

— $\hat{G}(t;\boldsymbol{\beta})$ est un estimateur non-paramétrique de la fonction de régression $\mathbb{E}(Y|\boldsymbol{\beta}^t X = t)$. On choisit l'estimateur de Nadaraya-Watson

$$\hat{G}(t;\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Y_i K(\frac{t-\boldsymbol{\beta}^t X_i}{h})}{\sum_{j=1}^{n} K(\frac{t-\boldsymbol{\beta}^t X_j}{h})} \tag{6.8}$$

où K(.) est un noyau symétrique et h le paramètre de lissage ou largeur de la fenêtre. Afin d'éviter le problème de dégénérescence, on substitue à l'estimateur (6.7), l'estimateur

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n_T} \psi(Y_i, \hat{G}^{(-i)}(\boldsymbol{\beta}^t X_i; \boldsymbol{\beta}))\tau_{n_T}(X_i), \tag{6.9}$$

où

$$\hat{G}^{(-i)}(\boldsymbol{\beta}^t X_i; \boldsymbol{\beta}) = \sum_{k\neq i} \frac{Y_k K(\frac{\boldsymbol{\beta}^t X_i - \boldsymbol{\beta}^t X_k}{h})}{\sum_{j\neq i} K(\frac{\boldsymbol{\beta}^t X_i - \boldsymbol{\beta}^t X_j}{h})},$$

— ψ est la fonction coût : Pour la méthode `MMC`, $\psi(y,r) = (y-r)^2$; Pour la méthode `MV`, $\psi(y,r) = -y\log(r) - (1-y)\log(1-r)$.
ICHIMURA [1993] utilise la méthode `MMC`,dans contexte de régression et KLEIN et SPADY [1993] utilise la méthode `MV`, dans un contexte de classification binaire.

— $\tau_{n_T}(.)$ est la fonction de cadrage, pour éviter les petites valeurs du dénominateur de l'estimateur $\hat{G}(.)$.

L'estimateur de $\boldsymbol{\beta}$ est très sensible au choix de h. ICHIMURA [1993] montre qu'on peut estimer, simultanément, $\boldsymbol{\beta}$ et h de (6.9).
KONG et XIA [2007] introduisent la méthode appelée *separated crossvalidation method*, plus efficace que la méthode d'Ichimura.

6.5 L'Algorithme de Transfert, en apprentissage supervisé

6.5.1 Méthodologie

A partir de l'échantillon d'apprentissage $\mathcal{S}$, on estime les paramètres du modèle semi-paramétrique `SIM` et la fonction lien *i.e.*, $\hat{\boldsymbol{\beta}}$ et $\hat{G}$.
L'affectation des individus de $\mathcal{S}^*$, se fait au moyen de la probabilité *a posteriori* $\hat{P}(Y_j^* = 1|\mathbf{X}_j^*)$, dans le cas binaire, égale à

$$\hat{\mathbb{E}}(Y_j^*|\mathbf{X}_j^*) = \hat{G}(L(\mathbf{X}_j^*)), \qquad j = 1,...,n^*, \tag{6.10}$$

où $L(\mathbf{X}_j^*) = c + \hat{\boldsymbol{\beta}}^t \boldsymbol{\Lambda}\mathbf{X}_j^*$, $\quad c \in \mathbb{R}$ et $\boldsymbol{\Lambda} = \mathbf{Diag}(\lambda_1,...,\lambda_p)$.

Afin d'estimer les paramètres $(c, \boldsymbol{\Lambda})$, on utilise l'approche `MV`, avec comme fonction à maximiser, la vraisemblance conditionnelle

$$\ell(c,\boldsymbol{\Lambda}) = \sum_{j=1}^{n_T^*} Y_j^* \log\left(\hat{P}(Y_j^* = 1|\mathbf{X}_j^*)\right) + (1-Y_j^*)\log\left(1-\hat{P}(Y_j^*|\mathbf{X}_j^*)\right). \tag{6.11}$$

n_T^* étant la taille de l'échantillon d'apprentissage, pris dans $\mathcal{S}^*$. Dans ce qui suit, on présente les cas (ou hypothèses) sur $(\alpha, \boldsymbol{\Lambda})$, chacun des cas correspond à un type de lien entre modèle et modèle transféré.

6.5.2 Relations entre modèle et modèle transféré

L'estimation des paramètres c et $\boldsymbol{\Lambda}$ est selon différents cas de domaine de valeurs admissibles :

— M_0^* : Pas de paramètre à estimer : $c = 0$ et $\boldsymbol{\Lambda} = \mathbf{I}_p$, où $\mathbf{I}_p$ la matrice d'identité.

— M_1^* : Ici, c est à estimer et $\boldsymbol{\Lambda} = \mathbf{I}_p$.

— M_2^* : $c = 0$ et $\boldsymbol{\Lambda} = \lambda I_p$, où $\lambda \in \mathbb{R}$.

— M_3^* : c est libre et $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p$ *i.e.*, deux paramètres sont à estimer.

— M_4^* : $c = 0$ et $\boldsymbol{\Lambda} = \mathbf{Diag}(\lambda_1,...,\lambda_p)$, où $\lambda_1,...,\lambda_p \in \mathbb{R}$ sont à estimer.

— M_5^* : Le modèle le plus complexe : c est libre et $\boldsymbol{\Lambda} = \mathbf{Diag}(\lambda_1,...,\lambda_p)$, où $\lambda_j \in \mathbb{R}$, $j = 1,...,p$.

6.6 Expériences numériques

6.6.1 Algorithme de transfert semi-paramétrique

L'échantillon $\mathcal{S}^*$ est scindé en deux échantillons : Un échantillon d'apprentissage $\mathcal{S}_{\mathrm{T}}^*$ et un échantillon de prédiction (test) $\mathcal{S}_{\mathrm{P}}^*$. Les données, en entrée, de cet algorithme sont $\mathcal{S}_{\mathrm{T}}^*$, $\mathcal{S}_{\mathrm{P}}^*$ et $\hat{\boldsymbol{\beta}}$. Les principales étapes de l'algorithme sont les suivantes :

1. Calculer
$$\mathrm{L}(\mathrm{X}_j^*) = c + \hat{\boldsymbol{\beta}}^t \boldsymbol{\Lambda} \mathrm{X}_j^*, \qquad j = 1, ..., n_{\mathrm{T}}^* \tag{6.12}$$
2. Estimer les paramètres de transfert $c \in \mathbb{R}$ et $\boldsymbol{\Lambda} \in \mathbb{R}^p$ selon le lien stochastique choisi en maximisant la fonction de vraisemblance empirique donnée en (6.11) par rapport à c et $\boldsymbol{\Lambda}$.
3. Remplacer les paramètres de transfert estimés en (6.12) afin d'obtenir $\mathrm{L}(\mathbf{X}^*)$ et puis remplacer $\mathrm{L}(\mathbf{X}^*)$ en (6.10) pour affecter les individus de $\mathcal{S}^*$.
4. Nous prédisons l'échantillon de test $\mathrm{S}_{\mathrm{P}}^* = \mathrm{S}^* \setminus \mathrm{S}_{\mathrm{T}}^*$ en utilisant l'estimateur obtenu.
5. Pour mesurer la qualité de la méthode proposée, nous calculons le taux d'erreur (en anglais : error rate) : $e = \frac{1}{n_{\mathrm{P}}^*} \sum_{j=1}^{n_{\mathrm{P}}^*} \mathbb{1}_{\mathrm{Y}_j^* \neq \hat{\mathrm{Y}}_j^*}$, où n_{P}^* est la taille de $\mathrm{S}_{\mathrm{P}}^*$ et $\hat{\mathrm{Y}}_j^*$ la prédiction de l'affectation Y_j^*.

Pour montrer l'utilité du transfert de l'apprentissage, nous comparons les différents modèles de transfert à la méthode classique $\mathtt{M}_6$ utilisant un échantillon d'apprentissage, formé à partir de $\mathcal{S}$ et $\mathcal{S}^*$.

6.6.2 Données Morphométriques

L'application, présentée ici, se basera sur les données biologiques traités dans BIERNACKI et *al.* [2002], par l'analyse discriminante généralisée.

Les données considérées consistent en trois échantillons de petrels, provenant chacun de l'une parmi trois sous-espèces :

- Le premier est constitué d'oiseaux *Borealis* composé de 93 femelles (SEX= 2) et 113 mâles (SEX= 1) ;
- le second est constitué d'oiseaux *Diomedea* composé de 22 femelles et 16 mâles ;
- le troisième est constitué d'oiseaux de l'espèce *Edwardsii* composé de 44 mâles et 48 femelles.

Les trois sous-espèces se distinguent par leur répartition géographique (THIBAULT et BRETAGNOLLE [1998]). Le problème, ici, est de prédire le sexe (Variable SEX), à partir de cinq variables consistant en des mesures morphométriques :

BECH et BECL : Deux mesures relatives au bec (hauteur et longueur) ;

TARSE : Longueur du tarse ;

AILE : Envergure des ailes ;

QUEUE : Longueur de la queue.

La figure (6.1) visualise la densité de chacune des covariables, et par sous-espèce : *Borealis, Diomedea* et *Edwardsii*. Cette figure montre bien que les densités, par sous-espèce,

sont différentes. Ce qui signifie que nous ne pouvons utiliser un échantillon d'une sous-espèce comme un échantillon d'apprentissage et un autre comme échantillon de prédiction. Il est donc nécessaire de chercher une méthode de classification plus adaptée ; D'où l'idée du transfert de l'apprentissage.

Dans un premier temps, on s'intéresse à la prédiction du sexe d'oiseaux de la sous-espèce *Diomedea*, en se servant, aussi, de l'échantillon des *Borealis*.

L'échantillon S^* sera composé de *Diomedea* et sera décomposé en échantillon (S_T^*) pour l'apprentissage et $S_P^* = S^* \setminus S_T^*$ pour les tests, sur la qualité de prédiction. L'échantillon S est constitué de l'ensemble d'oiseaux *Borealis*. On tire au hasard 50 sous-échantillons S_T^* *Diomedea*, pour chacune des tailles parmi $10, 14, ..., 34$ et on estime les six modèles à partir de chacun de ces échantillons, combiné avec l'échantillon *Borealis*. La sous-figure, à gauche de la figure (6.2), visualise les différents résultats obtenus.

Dans un deuxième temps, le même type de simulation est réalisé pour la prédiction du sexe des individus de la sous espèce *Edwardsii*, en considérant aussi les oiseaux *Borealis*. On tire au hasard 50 sous-échantillons S_T^* *Edwardsii* de taille $10, 15, ..., 90$. La sous-figure à droite de la figure (6.2) montre les différents résultats obtenus.

Les résultats obtenus dans les deux sous-figures montrent la supériorité des modèles faisant intervenir les deux échantillons, en présence *i.e.,* les modèles de transfert.

6.6.3 Données de *Credit-scoring*

Nous considérons un exemple de données réelles, de *credit scoring*, sur les prêts privés d'une banque allemande. L'ensemble de données et la description des covariables, sont disponibles au lien
`http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kreditvar_e.html`,
voir aussi, pour une description plus détaillée, FAHRMEIR et *al.* [1994].

Ces données décrivent 1000 consommateurs. Pour chaque consommateur, la variable réponse binaire `creditability`, est disponible (`Kredit` = 1, si bon rembourseur ; `Kredit` = 0, sinon.), ainsi que 20 covariables pouvant influencer le comportement, en remboursement. Une phase de traitement primaire des données conduit à ne retenir les six covariables suivantes :

`laufkont` *Balance of current account with the following four categories :*

1 : `no running account;`

2 : `no balance or debit;`

3 : `medium running account (less than 200 Deutsche Mark (DM));`

4 : `large running account (greater or equal to 200 DM or checking account for at least one year)`

`laufzeit` : *Duration of credit in months*;

`sparkont` : *Value of savings or stocks*;

`moral` : *Payment of previous credits*;

`beszeit` : *Duration of employment* with five categories :

1 : `unemployed;`

2 : less than one year;

3 : more than one year and less than four years;

4 : more than four years and less than seven years;

5 : more than seven years.

weitkred : *Further running credits.*

Les 1000 observations se répartissent en deux strates : 700 observations pour lesquelles Kredit = 1 et 300 pour Kredit = 0. Pour cette expérience, nous étudions le comportement de emprunteurs non-clients. La variable Laufkont est utilisée pour séparer l'ensemble de ces observations en deux échantillons : L'échantillon des clients $\mathcal{S}$ (Laufkont>1) de 726 observations et l'échantillon des non-clients $\mathcal{S}^*$ (Laufkont=1) de 274 observations.

On tire, au hasard, l'échantillon d'apprentissage S_T^*. Plusieurs tailles sont expérimentées $n^* = 50, 100, 150, 200$, à partir de l'échantillon *non-clients* S^*.
Les résultats obtenus sont synthétisés par la figure (6.3) et montrent la supériorité manifeste des modèles faisant intervenir les deux échantillons en présence (clients et non-clients).

Afin d'illustrer l'importance des modèles de transfert, on a mis dans le tableau 6.1 et pour trois exemples, le temps écoulé et de le taux d'erreur pour chaque modèle de transfert et également pour le modèle classique M_6(sans transfert). Il est clair que le temps écoulé et le taux d'erreur pour la majorité des modèles de transfert sont moindres que ceux du modèle sans transfert, ce qui signifie que la méthode de transfert proposée donne des meilleurs résultats que la méthode classique.

	Temps écoulé et taux d'erreur					
S	Borealis				Clients	
S^*	Diomedea		Edwardsii		Non-clients	
	temps	erreur	temps	erreur	temps	erreur
M_1	26.27	10.52	26.40	11.95	45.35	43.06
M_2	26.24	13.15	26.36	9.78	45.55	41.97
M_3	26.65	10.52	26.74	9.78	47.55	39.78
M_4	29.30	10.52	27.93	9.78	56.33	40.51
M_5	29.38	13.15	29.89	9.78	56.30	39.41
M_6	34.23	15.78	28.08	9.78	261.97	43.79

TABLEAU 6.1 – Performance de l'algorithme : Temps de calcul (en secondes) et taux d'erreur de classement.

6.7 Conclusion

De ces expériences numériques, nous retenons les conclusions ci-après :

En premier lieu, l'approche consistant à apprendre un modèle d'une première sous-population, de prédire le groupe d'appartenance d'individus d'une autre sous-population (sans aucune adaptation de la règle d'allocation *i.e.*, modèle $\mathtt{M}_0$, conduit à un taux d'erreur élevé. Ceci est en contradiction avec les hypothèses, à la base des méthodes classiques de classification supervisée. Cela justifie l'idée d'utiliser un maximum d'individus, pour adapter ou mettre à jour la règle d'affectation.

Deuxièmement, l'approche consistant à l'apprentissage sans exploiter le premier ensemble de données n'est pas satisfaisante ; Les résultats souffrent de la petite taille du deuxième échantillon. Le problème de la taille du deuxième ensemble de données est ce qui motive l'idée du transfert de modèle.

Troisièmement, le modèle $\mathtt{M}_6$ qui amène à utiliser les deux échantillon pour prédire les individus du deuxième échantillon est coûteuse et donne un taux d'erreur assez élevé par rapport aux modèles de transfert (surtout $\mathtt{M}_2$ et $\mathtt{M}_3$).

Quatrièmement, le modèle $\mathtt{M}_1$ correspond à la pratique actuelle et très connue parmi les biologistes (Voir VAN FRANEKER et TER BRAAK [1993]) et des spécialistes de *credit scoring*. Cette pratique consiste à changer, de façon empirique et sans justification théorique, la valeur du seuil d'affectation aux classes (ou de manière équivalente, l'intercept de la fonction de score linéaire ou score Anderson).

Enfin, il est assez clair que les meilleurs modèles (au sens de l'erreur empirique), à savoir, $\mathtt{M}_2$, $\mathtt{M}_3$ et $\mathtt{M}_4$, sont ceux qui exploitent de façon séquentielle les deux ensembles de données. Le premier est utilisée pour estimer une première règle d'affectation (approprié à la prédiction des individus de la première sous-population) et le second pour adapter cette règle, pour prédire les individus de la seconde sous-population.

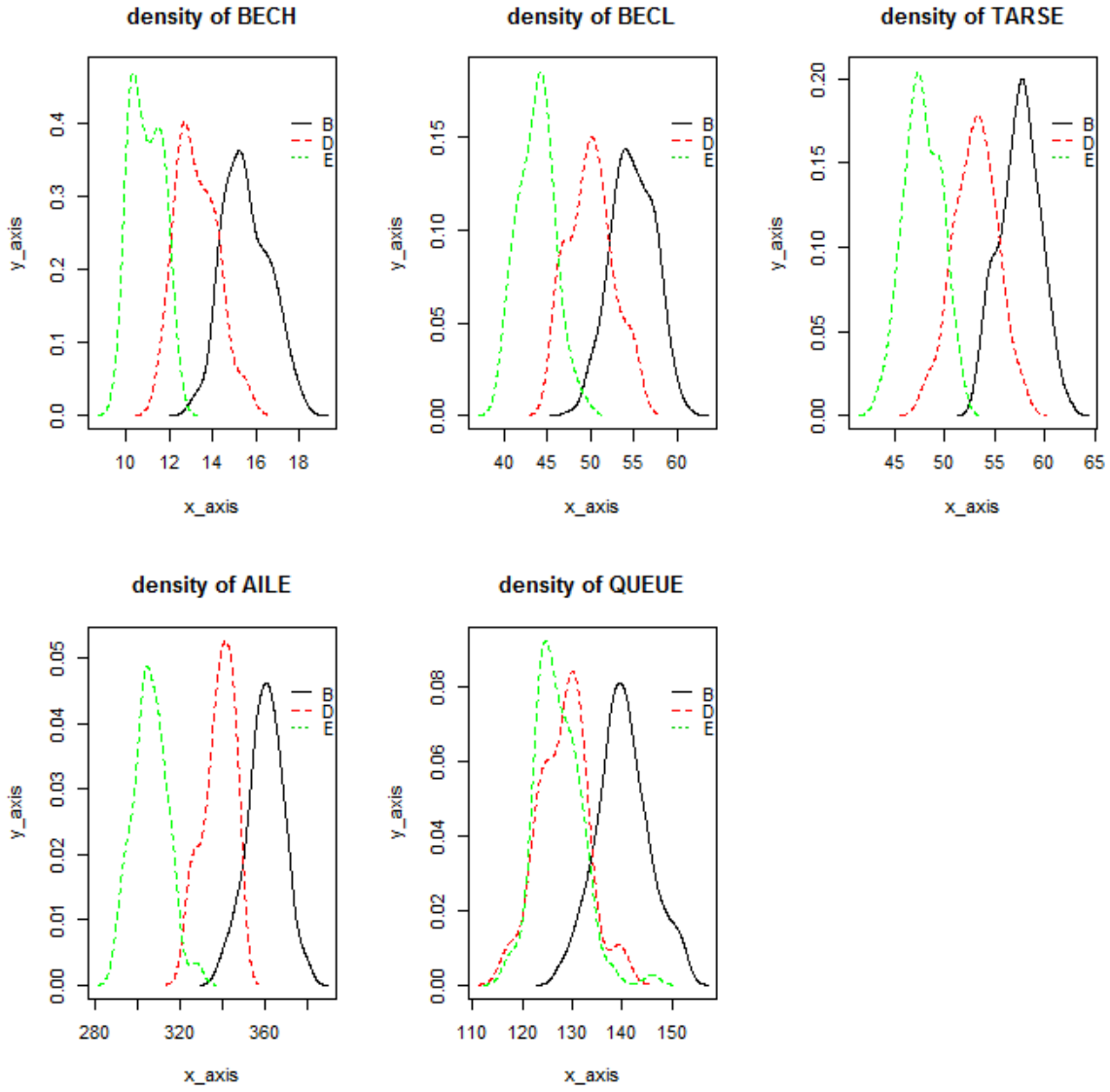


FIGURE 6.1 – Densité des différentes covariables des échantillons *Borealis*, *Diomedea* et *Edwardsii*.

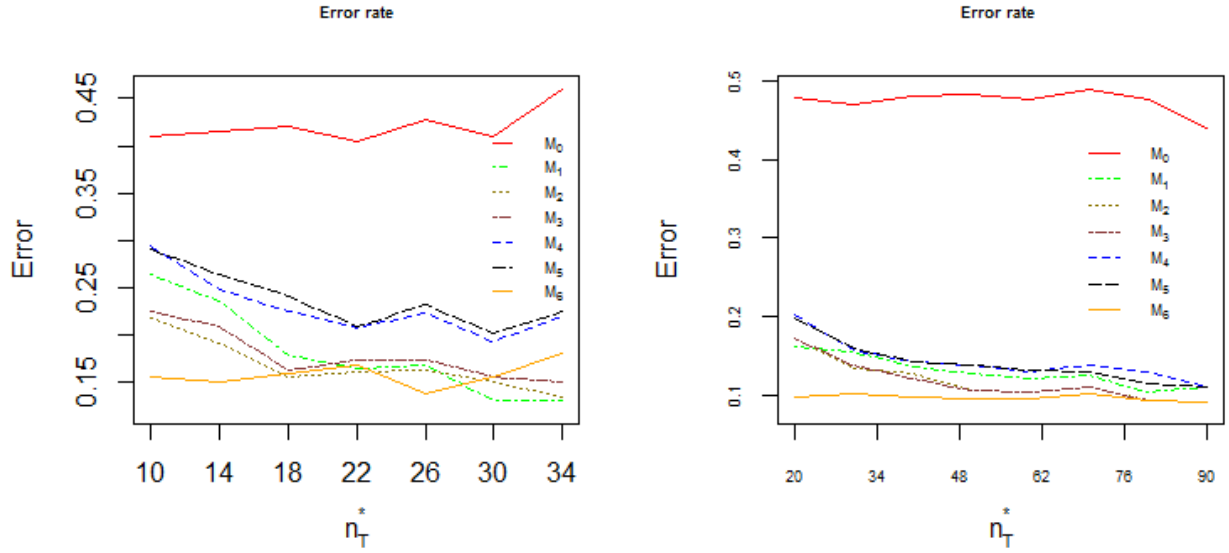


FIGURE 6.2 – La sous-espèce *Borealis* forme échantillon d'apprentissage. Dans la sous-figure à gauche, la sous-espèce *Diomedea* forme l'échantillon de prédiction ; Dans la sous-figure à droite, la sous-espèce *Edwardsii* forme l'échantillon de prédiction.

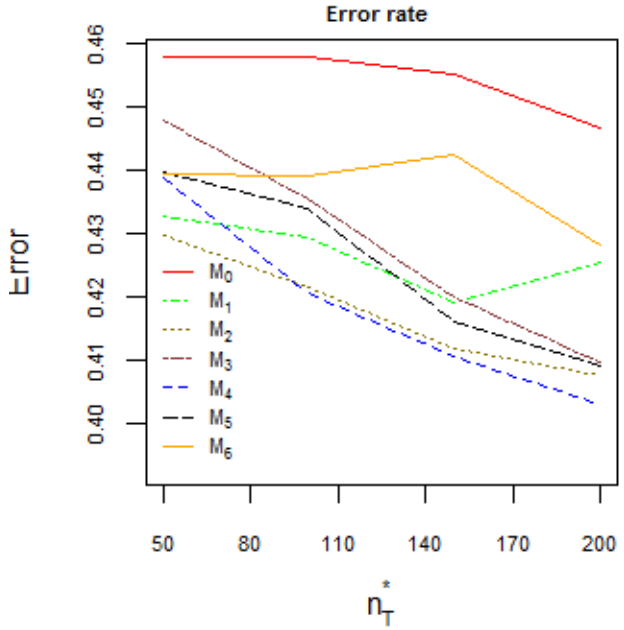


FIGURE 6.3 – Les clients forment l'échantillon d'apprentissage et les non-clients l'échantillon de prédiction.

6.8 Références

BENINEL, F. et C. BIERNACKI. 2007, «Relaxations de la régression logistique : modèles pour l'apprentissage sur une sous-population et la prédiction sur une autre», *RNTI A1, Data Mining et Apprentissage Statistique : application en assurance, banque et marketing*, p. 200–212. 94

BENINEL, F. et C. BIERNACKI. 2009, «Updating a logistic discrimination rule - comparing some logistic submodels in credit-scoring», *Proceedings of the International Conference on Agents and Artificial Intelligence*, p. 267–274. 94

BENINEL, F., C. BIERNACKI, C. BOUVEYRON, J. JACQUES et A. LOURME. 2012a, «Parametric link models for knowledge transfer in statistical learning», *Knowledge Transfer : Practices, Types and Challenges*. 93

BENINEL, F., W. BOUAGUEL et G. B. MUFTI. 2012b, «Transfer learning using logistic regression in credit scoring», *CoRR*, vol. abs/1212.6167. URL http://arxiv.org/abs/1212.6167. 93, 94

BIERNACKI, C., F. BENINEL et V. BRETAGNOLLE. 2002, «A generalized discriminant rule when training population and test population differ on their descriptive parameters», *Biometrics*, vol. 58, nº 2, p. 387–397. 92, 93, 94, 97

DELECROIX, M., W. HÄRDLE et M. HRISTACHE. 2003, «Efficient estimation in conditional single-index regression», *Journal of Multivariate Analysis*, vol. 86, nº 2, p. 213–226. 95

DELECROIX, M., M. HRISTACHE et V. PATILEA. 2006, «On semiparametric m-estimation in single-index regression», *Journal of Statistical Planning and Inference*, vol. 136, nº 3, p. 730–769. 95

FAHRMEIR, L., G. TUTZ, W. HENNEVOGL et E. SALEM. 1994, *Multivariate statistical modelling based on generalized linear models*, vol. 2, Springer New York. 98

ICHIMURA, H. 1993, «Semiparametric least squares (sls) and weighted sls estimation of single-index models», *Journal of Econometrics*, vol. 58, nº 1, p. 71–120. 95, 96

KLEIN, R. W. et R. H. SPADY. 1993, «An efficient semiparametric estimator for binary response models», *Econometrica : Journal of the Econometric Society*, p. 387–421. 95, 96

KOMAROVA, T. 2013, «Binary choice models with discrete regressors : Identification and misspecification», *Journal of Econometrics*, vol. 177, nº 1, p. 14–33. 95

KONG, E. et Y. XIA. 2007, «Variable selection for the single-index model», *Biometrika*, vol. 94, nº 1, p. 217–229. 96

MANSKI, C. F. 1988, «Identification of binary response models», *Journal of the American Statistical Association*, vol. 83, nº 403, p. 729–738. 95

PATILEA, V. 2007, «Semiparametric regression models with applications to scoring : a review», *Communications in Statistics—Theory and Methods*, vol. 36, nº 14, p. 2641–2653. 95

THIBAULT, J.-C. et V. BRETAGNOLLE. 1998, «A mediterranean breeding colony of cory's shearwater calonectris diomedea in which individuals show behavioural and biometric characters of the atlantic subspecies», *Ibis*, vol. 140, nº 3, p. 523–528. 97

VAN FRANEKER, J. et C. TER BRAAK. 1993, «A generalized discriminant for sexing fulmarine petrels from external measurements», *The Auk*, p. 492–502. 100

Conclusion

Nous avons abordé, dans ce rapport de thèse, deux parties : Une partie consacrée à l'estimation des quantiles conditionnels et une autre, à l'apprentissage supervisé.
Dans la première partie, nous avons comparé cinq méthodes d'estimation du quantile conditionnel : La régression quantile constante(1990), la régression quantile linéaire locale(1994), l'estimateur à double noyau proposé par Yu et Jones(1998), l'estimateur de Nadaraya-Watson pondéré(2002) et l'estimateur à double noyau de Cai et Wang (2008).
Nous recommandons la régression quantile linéaire locale comme une approche directe, et l'estimateur à double noyau de Cai et Wang (2008) comme une approche indirecte.
Nous avons remarqué que ces estimateurs peuvent se mettre sous une forme unifiée.
Nous avons aussi, traité le problème d'estimation des paramètres de lissage et pour cela, nous avons abordé cinq méthodes : Règle *rule of thumb* de Yu et Jones (1998), la méthode *plug-in* itérative proposée par Attar (2002), la validation croisée classique, la validation croisée généralisée et critère de Cai proposé par Cai et Tiwari (2000).
Parmi ces méthodes, nous recommandons la règle de Yu et Jones, car simple à implémenter et fournissant des résultats satisfaisants.
Nous avons proposé, en fin de cette partie, d'estimer le quantile conditionnel par un estimateur à double noyau, en utilisant un noyau asymétrique en x.
Nous avons montré, que sous certaines hypothèses faibles, cet estimateur a la supériorité potentielle sur celui qui utilise un noyau symétrique en x. Cette supériorité vient du fait que les noyaux asymétriques ne mettent pas de poids en dehors du support de la courbe.
L'estimateur obtenu possède un AMSE faible, variance finie et il est résistant aux données clairsemés.
La performance de notre estimateur est montrée par des expériences numériques sur des données simulées et des données réelles.
Dans la deuxième partie, et après avoir présenté les méthodes conventionnelles de classification supervisée (les plus rencontrées dans la littérature et les plus connues des spécialistes), nous proposons une méthode de transfert d'un modèle d'apprentissage semi-paramétrique. Cette méthode de transfert, est basée sur un ensemble de paramètres ; différentes contraintes sur ces paramètres, sont expérimentées et comparées.
Nous avons montré la performance de notre méthode au moyen des expériences numériques sur des données morphométriques et des données de *credit-scoring*.

Articles publiés et article sous press :

1. Knefati, M.A., Oulidi, A., Abdous, M. Local linear double and asymmetric kernel estimation of conditional quantiles. Communications in Statistics : Theory and Methods (accepté le 28 janvier 2014, sous press).
2. Knefati, M. A., Chauvet, P. E., N'Guyen, S., Daya, B. (2014). Reference Curves Estimation Using Conditional Quantile and Radial Basis Function Network with Mass Constraint. Neural Processing Letters, 1-14.
3. Knefati, M-A., Beninel, F.(2014). Transfer of semiparametric single index model in binary classification. COMPSTAT2014, (Eds. Gili,M., Gonzalez-Rodriguez, G., Nieto-Reyes, A.), 609-616.

Local linear double and asymmetric kernel estimation of conditional quantiles

Muhammad Anas Knefati[1]*Abderrahim Oulidi[2] and Belkacem Abdous[3]

[1]*Department of Mathematics, Faculty of sciences, Poitiers University, France*;

[2]*Rabat International University, Rabat-Salé, Morroco;*

[3]*Department of Preventive and Social Medicine, Laval University, Québec, Canada*

January 27, 2014

In this work, we propose and investigate a family of nonparametric quantile regression estimates. The proposed estimates combine local linear fitting and double kernel approches. More precisely, we use a Beta kernel when covariate's support is compact and Gamma kernel for left bounded supports. Finite sample properties together with asymptotic behavior of the proposed estimators are presented. It is also shown that these estimates enjoy the property of having finite variance and resistance to sparse design.

Keywords: Quantile Regression; Double Kernel Conditional Quantile Estimation; Asymmetric Kernels; Beta Kernels; Gamma Kernels.

1 Introduction

In many applications, one is interested in quantile of a response variable conditioned on covariates rather than in the usual mean regression. A classical exemple is the value at risk(VaR) which is widely used as a market risk measure.

Nonparametric estimation of conditional Quantiles has been tackled by several authors. The proposed approaches might be cast into two classes: direct and indirect methods.

Direct methods consist mainly in minimizing a sum of asymmetrically weighted absolute residuals that means simply giving differing weights to positive and negative residuals by using the check function introduced by Koenker and Basset (1978) given by $\rho_\tau(u) = u\tau\mathbb{1}(u \geq 0) + u(1-\tau)\mathbb{1}(u < 0)$, where $\mathbb{1}$ is the indicator function. See, for instance, Fan et al. (1994),Yu and Jones (1998), and Goh (2012) for a local linear estimate of conditional quantiles that is a direct method.

Indirect methods are performed in two steps: the estimation of the conditional distribution is performed first, then the inverse of the obtained estimator is used to estimate

*Address correspondence to Muhammad Anas Knefati, Department of Mathematics, Faculty of sciences, Poitiers University, France; Email: muhammad.anas.knefati@univ-poitiers.fr

the conditional quantile. Examples of the indirect methods are provided by Yu and Jones (1998), Cai (2002), and Cai and Wang (2008).

This work deals with conditional quantiles estimation problem by adapting the local linear double kernel technique of Yu and Jones (1998). More precisely, we assume that the predictor variable's support has at least one finite endpoint and make use of either a Beta or a Gamma kernel family.

The choice of Beta kernels when covariate's support has finite endpoints can be motivated by the well known result which states that any continuous function on $[0, 1]$ might be uniformly approximated by a Bernstein polynomial. In fact, approximation by Bernstein polynomials is equivalent to kernel estimation with a smoothing parameter of order $n^{-1/2}$(n being the sample size) and a binomial probability function as a kernel. That said, these estimators suffer from the under-smoothing problem. To overcome this problem when dealing with regression curves with equally spaced design points, Brown and Chen (1999) proposed a kernel estimator based on a binomial density function as a kernel. These estimators avoid boundary bias problem since they do not put any weight outside the curve's support. Moreover, the associated optimal mean integrated squared error is equivalent to that of standard kernel estimators when the curve has an unbounded support. An extension of this work to arbitrary designs has been proposed by Chen (2000a). Also, a probability density estimator based on Beta and gamma kernels has been proposed by Chen (1999, 2000b). This estimator has smaller finite-sample variance compared to that of some existing estimators. Chen (2002) also proposed to use the Beta and Gamma kernels when estimating regression functions by local linear smoother. He showed that these estimators have finite variance and resistance to sparse designs. As it will be shown below, the proposed quantile estimators inherit these desirable properties and its asymptotic mean square error is better than the one based on symmetric kernels.

This work is organized as follows: Section 2 is devoted to the construction of conditional quantiles local linear double kernel estimators. Their asymptotic properties are investigated in Section 3. A comparison between the use of symmetric and asymmetric(Gamma/beta) kernels for the covariate variable is presented in Section 4. In Section 5, the performance of the proposed estimates is assessed by means of a simulation study and a real example is also used with an easily method for selecting the bandwidth based on the cross validation. Proofs are postponed to the last Section.

2 Local linear double kernel smoothing using asymmetric kernel

Let (X, Y) be a bivariate random variable, where Y is a real-valued random variable while the covariate X has either a compact or a left-bounded support. For simplicity, we will assume that this support is $[0, 1]$ or $[0, \infty)$. Denote by $F(y|x)$ the conditional distribution function of Y given $X = x$. For $\tau \in (0, 1)$, define τth conditional quantile of Y by

$$q_\tau(x) = \inf\{y \in \mathbb{R} : F(y|x) \geq \tau\} \equiv F^{-1}(\tau|x). \tag{1}$$

Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are a set of independent observations distributed as (X, Y). To estimate quantiles $q_\tau(x)$, we adopt an indirect approach. First, we will use a double kernel method to estimate the conditional distribution $F(y|x)$, then the inverse of the obtained estimator will be used to estimate the desired conditonal quantiles. To this end, let us first recall that the conditional distribution function $F(y|x)$ fulfills

$$F(y|x) = \arg\min_a E\left[(\mathbb{1}(Y \leq y) - a)^2 | X = x\right],$$

where $\mathbb{1}(.)$ is the usual indicator function. See, for instance, Fan and Yao (2003)[1] for more details. This relationship is the foundation of the classical local polynomial fitting approach. Indeed, for instance, local linear fitting estimates of $F(y|x)$ and its derivative are obtained by minimizing in a and b the following criterion

$$\sum_{i=1}^{n} \left((\mathbb{1}(Y_i \leq y) - a - b(x - X_i))^2 | X = x\right) K(x, h_1, X_i) \tag{2}$$

where h_1 is a given smoothing parameter and $K(., h_1, .)$ is an arbitrary kernel function to be specified. The solution $\hat{a}(x, y)$ might be used as a nonparametric estimator of $F(y|x)$. However, for any fixed x, this estimate is a discontinuous function of y. This drawback may be alleviated, as proposed in Cai and Wang (2008), by replacing the indicator functions in (2) by their smoothed versions:

$$(\mathbb{1} * W_{h_2})(y) = \int \mathbb{1}(Y_i \leq t) W_{h_2}(y - t)dt = \int \mathbb{1}(t \leq \frac{y - Y_i}{h_2}) W(t)dt := \Omega(\frac{y - Y_i}{h_2})$$

where h_2 is a a given smoothing parameter, Ω is the cumulative distribution function of an arbitrary kernel density function W and $W_{h_2}(.) = W(./h_2)/h_2$. Then, the weighted least squares problem in (2) becomes

$$\sum_{i=1}^{n} \left(\left(\Omega(\frac{y - Y_i}{h_2}) - a - b(x - X_i) \right)^2 | X = x \right) K(x, h_1, X_i) \tag{3}$$

In the sequel we will use the solution $\hat{a}$ as an estimator of $F(y|x)$. This solution will be referred as the double kernel local linear estimator and will be denoted by $\hat{F}(y|x)$. Hence, the same arguments as those in Fan and Gijbels (1996) lead to the following expression

$$\hat{F}(y|x) = \sum_{i=1}^{n} \omega_i(x) \Omega(\frac{y - Y_i}{h_2}),$$

where

$$\omega_i(x) = \frac{S_2(x) - (x - X_i)S_1(x)}{S_2(x)S_0(x) - S_1^2(x)} K_{x,h_1}(X_i), \qquad i = 1, ..., n, \qquad \text{with}$$

[1] Page 455

$$S_l = \sum_{j=1}^{n} (x - X_j)^l K_{x,h_1}(X_i), \qquad l = 0, 1, 2$$

where $K_{x,h_1}(X_i) = K(x, h_1, X_i)$.

Next, for the reasons discussed earlier, henceforth we will make use of asymmetric kernels when smoothing with respect to x. More precisely, we will use either the probability density function of a Beta$(\frac{x}{h_1} + 1, \frac{1-x}{h_1} + 1)$ distribution, *i.e.*

$$K_{x,h_1}(u) = \frac{u^{\frac{x}{h_1}} (1-u)^{\frac{1-x}{h_1}}}{B(\frac{x}{h_1} + 1, \frac{1-x}{h_1} + 1)}, \qquad u \in [0, 1]$$

or the Gamma$(\frac{x}{h_1} + 1, h_1)$ distribution, i.e.

$$K_{x,h_1}(u) = \frac{u^{\frac{x}{h_1}} e^{-\frac{u}{h_1}}}{h_1^{\frac{x}{h_1}+1} \Gamma(\frac{x}{h_1} + 1)}, \qquad u \in [0, \infty)$$

where B and Γ are the classical *Beta* and *Gamma* functions respectively. Both kernels families have varying kernel shapes and become more asymmetric as x moves towards the boundary. We use a *Beta* kernel if X has the support $[0, 1]$ and a *Gamma* kernel if X has the support $[0, \infty)$. Finally, by using the definition of conditional quantiles (1), we estimate $q_\tau(x)$ by inverting the double kernel local linear estimator $\hat{F}(y|x)$, *i.e.*

$$\hat{q}_\tau(x) = \hat{F}^{-1}(\tau|x). \tag{4}$$

The next section presents some asymptotic properties of the proposed estimators. Namely, we focus on the asymptotic expansions of the mean squared error (MSE) of the estimators $\hat{F}(y|x)$ and $\hat{q}_\tau(x)$ introduced above.

3 Asymptotic properties

3.1 Asymptotic expansions for $\hat{F}(y|x)$

To begin with, let us state some notations and regularity assumptions. Denote by $F(x, y)$ and $f(x, y)$ the cdf and pdf of (X, Y). Let $f(y|x)$ and $g(x)$ stand for the conditional pdf of Y given $X = x$ and the marginal pdf of X respectively. Consider the following regularity assumptions:

1. The cdf $F(x, y)$ admits a continuous and bounded density $f(x, y)$.
2. The densities $g(.)$ and $f(.|.)$ are bounded and strictly positive.
3. For any $\tau \in (0, 1)$ and any x, the conditional quantiles $q_\tau(x)$ are unique.
4. The kernel W is a symmetric pdf with finite second order moment.

5. The bandwidths h_1 and h_2 satisfy $h_1 + (nh_1)^{-1} \to 0$ and $h_2 + (nh_2)^{-1} \to 0$ as $n \to \infty$.

6. For n large enough, $h_2 = o(h_1)$.

Also, we will use the notations $\mu_2(K) = \int u^2 K(u)du$, $R(K) = \int K^2(u)du$, and

$$\ell^{ab}(q_\tau(x)|x) = \frac{\partial^{ab}}{\partial y^a \partial x^b} l(y|x), \qquad \text{for any function } \ell.$$

The X's support will be split into two parts: an interior region S_I and a boundary region S_B. For a compact support $[0,1]$, these regions might be defined as follows

$$\begin{aligned} S_I &= \{x : x \in [0,1], x/h_1 \to \infty \text{ and } (1-x)/h_1 \to \infty\}, \\ S_B &= \{x : x \in [0,1], x/h_1 \to c \text{ or } (1-x)/h_1 \to c, \text{for some } \ c > 0\}. \end{aligned}$$

Similarly, for a left bounded support $[0, \infty)$, we will use

$$\begin{aligned} S_I &= \{x : x \in [0,\infty], x/h_1 \to \infty\}, \\ S_B &= \{x : x \in [0,\infty], x/h_1 \to c, \text{for some } \ c > 0\}. \end{aligned}$$

The following theorem gives the asymptotic expansions of the bias and the variance of $\hat{F}(y|x)$ for X with a compact support.

Theorem 1. *Let $\psi(x) = x(1-x)$ for Beta kernel and $\psi(x) = x$ for Gamma kernel. Then under assumptions 1-6, one has*

$$Bias\{\hat{F}(y|x)\} = \begin{cases} \frac{1}{2}\psi(x)F^{02}(y|x)h_1 + O(h_1^2) & \text{if } \ x \in S_I; \\ \\ \frac{1}{2}(2+c)F^{02}(y|x)h_1^2 + o(h_1^2) & \text{if } \ x \in S_B \end{cases}$$

and

$$Var\{\hat{F}(y|x)\} = \begin{cases} \frac{\sigma^2(x,y)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)n\sqrt{h_1}} + o(\frac{1}{n\sqrt{h_1}}) & \text{if } \ x \in S_I; \\ \\ \frac{\sigma^2(x,y)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)nh_1} + o(\frac{1}{nh_1}) & \text{if } \ x \in S_B, \end{cases}$$

where $\sigma^2(x,y) = F(y|x)(1-F(y|x))$.

3.2 Asymptotic expansions for $\hat{q}_\tau(x)$

Here again, similar arguments to those used above lead to the next expansions MSE of $\hat{q}_\tau(x)$.

Theorem 2. *Under assumptions of Theorem 1, we get*

$$Bias\{\hat{q}_\tau(x)\} = \begin{cases} \frac{\frac{1}{2}\psi(x)F^{02}(q_\tau(x)|x)h_1}{f(q_\tau(x)|x)} + O(h_1^2) & if \;\; x \in S_I; \\ \\ \frac{\frac{1}{2}(2+c)F^{02}(q_\tau(x)|x)h_1^2}{f(q_\tau(x)|x)} + o(h_1^2) & if \;\; x \in S_B \end{cases}$$

and

$$Var\{\hat{q}_\tau(x)\} = \begin{cases} \frac{\tau(1-\tau)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)f^2(q_\tau(x)|x)n\sqrt{h_1}} + o(\frac{1}{n\sqrt{h_1}}) & if \;\; x \in S_I; \\ \\ \frac{\tau(1-\tau)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)f^2(q_\tau(x)|x)nh_1} + o(\frac{1}{nh_1}) & if \;\; x \in S_B. \end{cases}$$

To assess the performance of the estimator $\hat{q}_\tau(x)$, we will use its asymptotic mean squared error ($AMSE$). Indeed, on the basis of the foregoing, the $AMSE$ of $\hat{q}_\tau(x)$ is given by

$$AMSE(x) = \begin{cases} \frac{1}{4}\frac{\{\psi(x)F^{02}(q_\tau(x)|x)h_1\}^2}{f^2(q_\tau(x)|x)} + \frac{\tau(1-\tau)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)f^2(q_\tau(x)|x)n\sqrt{h_1}} & \text{if} \quad x \in S_I; \\ \\ \frac{1}{4}\frac{\{(2+c)F^{02}(q_\tau(x)|x)h_1^2\}^2}{f^2(q_\tau(x)|x)} + \frac{\{\tau(1-\tau)\}\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)f^2(q_\tau(x)|x)nh_1} & \text{if} \quad x \in S_B. \end{cases}$$

By minimizing $AMSE(x)$, we obtain the optimal bandwidth

$$h_1^{opt}(x) = \begin{cases} \left(\frac{\tau(1-\tau)}{2\sqrt{\pi}\psi(x)^{5/2}\{F^{02}(q_\tau(x)|x)\}^2 g(x)}\right)^{\frac{2}{5}} n^{-\frac{2}{5}} & \text{if} \;\; x \in S_I; \\ \\ \left(\frac{\tau(1-\tau)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)(2+c)^2\{F^{02}(q_\tau(x)|x)\}^2 g(x)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} & \text{if} \;\; x \in S_B. \end{cases} \tag{5}$$

Substituting this optimal bandwidth in (3.2) and under the assumption $h_2 = o(h_1)$, we obtain the optimal asymptotic mean square error for double kernel local linear estimation:

$$AMSE^{opt}(x) = \begin{cases} \frac{5}{4}\left(\frac{\tau(1-\tau)}{2\sqrt{\pi}g(x)}\right)^{\frac{4}{5}} \{F^{02}(q_\tau(x)|x)\}^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{if} \;\; x \in S_I; \\ \\ \frac{5}{4}(2+c)^{\frac{1}{5}}\left(\frac{\tau(1-\tau)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)}\right)^{\frac{4}{5}} \{F^{02}(q_\tau(x)|x)\}^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{if} \;\; x \in S_B. \end{cases} \tag{6}$$

By mimicking the approach in Chen (2000a), it can be shown that the bias and the variance in the boundary areas have negligible contribution to the asymptotic mean integrated square error with an error term of $o(\frac{1}{n\sqrt{h_1}} + h_1^2)$. It follows that

$$AMISE = \frac{\tau(1-\tau)}{n\sqrt{h_1}} \int_0^b \frac{1}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)f^2(q_\tau(x)|x)} dx + \frac{h_1^2}{4}\int_0^b \psi(x)^2 \left(\frac{F^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right)^2 dx.$$

where $b = 1$ for Beta kernel and $b = \infty$ for Gamma kernel.

By minimizing $AMISE$ with respect to h_1, we obtain the asymptotically optimal global bandwidth

$$h_1^{opt} = \left\{ \frac{\tau(1-\tau)\int_0^b \frac{dx}{\sqrt{\psi(x)}g(x)f(q_\tau(x)|x)}}{2\sqrt{\pi}\int_0^b \psi(x)^2 \left(\frac{F^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right)^2 dx} \right\}^{\frac{2}{5}} n^{-\frac{2}{5}} \tag{7}$$

The asymptotic expansions associated to the *Gamma* and the *Beta* kernels are identical up to the term $\sqrt{1-x}$. As it has been remarked by Chen (2002), the variance with *Gamma* kernel decrease when x increase as $x^{-\frac{1}{2}}$ appeared in the leading term of the asymptotic variance, and this property is highly recommended when estimating curves with sparse regions in the upper tail of the design density g. This gain in variance for large x is at the price of increasing the bias. This is equivalent to the technique of using lager bandwidth values in areas where the design is sparse.

Furthermore, note that even though AMISE is unaffected by the behavior of the proposed estimators in S_B, optimality will be destroyed at the boundary region if the global bandwidth h_1^{opt} is used on the whole support. Indeed, the bandwidth h_1^{opt} is of order $n^{-2/5}$, while the optimal local bandwidth is of order $n^{-1/5}$ for x in S_B (see (5)). Thus, using a single bandwidth on both S_I and S_B will likely give rise to boundary problems.

4 Comparison of symmetric and asymmetric (Beta/Gamma) kernels

4.1 Asymptotic mean squared errors

According to Yu and Jones (1998), the optimal asymptotic MSE for the double kernel local linear estimator might be expressed as follows

$$AMSE_{sym}^{opt}(x) = \begin{cases} \frac{5}{4}\mu_2(K_{sym})^{\frac{2}{5}} \left(R(K_{sym})\frac{\tau(1-\tau)}{g(x)}\right)^{\frac{4}{5}} \{F^{02}(q_\tau(x)|x)\}^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{for } x \in S_I; \\ \frac{5}{4}\alpha_c(K_{sym})^{\frac{2}{5}} \left(\frac{\tau(1-\tau)\beta_c(K_{sym})}{g(x)}\right)^{\frac{4}{5}} F^{02}(q_\tau(x)|x)^{\frac{2}{5}} n^{-\frac{4}{5}}, & \text{for } x \in S_B. \end{cases} \tag{8}$$

where K_{sym} is a symmetric kernel, and

$$\alpha_c(K_{sym}) = \frac{a_2^2 - a_1 a_3}{a_0 a_2 - a_1^2}, \quad \beta_c(K_{sym}) = \frac{\int_\infty^c \{a_2 - a_1 u\}^2 K_{sym}(u)du}{\{a_0 a_2 - a_1^2\}^2}$$

with $a_l = \int_0^c u^l K_{sym}(u)du$, for l=0,1,2,3.

Clearly, the $AMSE$ of both symmetric and asymmetric kernel estimators have the same magnitude order $n^{-\frac{4}{5}}$. The only difference lies in multiplicative constants illustrated in Table (1) below, where $C(K_{sym})$ denotes the $AMSE$ multiplicative constant

for the estimator using a symmetric kernel and $C(K_{asym})$ denotes the $AMSE$ multiplicative constant for the one using an asymmetric kernel.

Table 1: Comparison between symmetric and asymmetric kernels

	$C(K_{sym})$	$C(K_{asym})$
Interior	$\mu_2(K_{sym})^{2/5} R(K_{sym})^{4/5}$	$\left(\frac{1}{2\sqrt{\pi}}\right)^{4/5}$
Boundary	$\alpha_c(K_{sym})^{2/5} \beta_c(K_{sym})^{4/5}$	$(2+c)^{1/5} \left(\frac{\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)}\right)^{4/5}$

We just take the classical Epanechnikov kernel, the optimal choice for symmetric kernels as proved by Fan et al. (1997), then we can easily see that for x belonging to the interior region, $C(K_{sym}) \approx 0.34$ whereas $C(K_{asym}) = 0.36$. This simply shows that the difference between these two kernel families is negligible for $x \in S_I$. While, for $x \in S_B$ the difference between them is quite different. Table 2 below exhibits several values of these constants and demonstrate the potential superiority of asymmetric kernels over symmetric ones when X's support is either compact or left bounded.

Table 2: Values of criteria $C(K_{sym})$, and $C(K_{asym})$ in boundary of x, for several values of $c \in]0, 1[$.

c	$C(K_{sym})$	$C(K_{asym})$
0.01	27.42	0.652
0.25	2.192	0.542
0.5	1.39	0.48
0.75	0.7	0.44
0.99	0.524	0.414

4.2 Finite Sample Variance

Seifert and Gasser (1996a) showed that the local linear estimate of a regression function might have an infinite unconditional variance when optimal local weights (compact kernels) are used. This is due to the fact that the denominator $S_2(x)S_0(x) - S_1^2(x)$ has a positive probability of being zero or being arbitrary small. To avoid this problem, Fan (1993) added n^{-2} in the denominator whereas Seifert and Gasser (1996b) proposed the local linear ridge estimation which adds a scalar c_1 to $S_2(x)$. An alternative solution relies on using beta or Gamma kernels, indeed as shown by Chen (2002) the associated local linear estimate has finite variance with probability 1.

5 Emprical applications

5.1 Bandwidth selection

The previous kernel-based estimates require the bandwidth parameter selection. This is a critical issue since the precision and the performance of the proposed estimates , it plays an important role

With the basic model at hand, one must address the important bandwidth selection issue, as the quality of the curve estimates depends sensitively on the choice of the bandwidth. According to Fan et al. (1996) and Yu and Jones (1998), the choice of h_2 is less critical than that of h_1. In fact, we just need to select a value that fulfills the regularity assumption $h_2 = o(h_1)$. Therefore we will put $h_2 = h_1 n^{-\epsilon}$ with $\epsilon > 0$ (in our simulations we took $\epsilon = 0.5$)

Now to select h_1, we use the leave-one-out cross validation method that has been adopted for quantile regression by Abberger (1998), i.e. we select h_1 that minimizes the following criterion :

$$CV(h) = \sum_{i=1}^{n} \rho_\tau(Y_i - \hat{q}_\tau^{(-i)}(X_i))$$

where $\rho_\tau(u) = u\tau \mathbb{1}(u \geq 0) + u(1-\tau)\mathbb{1}(u < 0)$ and $\hat{q}^{(-i)}$ is the conditional quantile estimator constructed from the sample with the ith observation omitted.

For comparison purposes and empirical assessment of the proposed estimates, we will use the cross-validation criterion to select the bandwidth for real data sets and use the optimal and theoretical bandwidth defined by (7) for simulated data.

5.2 Simulated examples

To illustrate the performance of asymmetric kernels, we consider two simulated examples where in both models the design density is relatively sparse towards the right end of the support. For comparison purposes and in order to avoid the finite sample variance problem in areas of sparse design, we will use the local linear ridge estimation proposed by Seifert and Gasser (1996b) who added a scalar c_1 to $S_2(x)$. We will assess the performance of the proposed local linear double kernel estimates by comparing symmetric (Gaussian or Epmechnikov) with asymmetric kernels (Gamma and Beta).

To select the smoothing parameter h_1, we will use the optimal global bandwidth given in (7) for Beta or Gamma kernel, while for symmetric kernels, we will use the following optimal bandwidth (see Yu and Jones (1998))

$$h_{1,sym}^{opt} = \left\{ \frac{R(K)\tau(1-\tau)\int_0^\infty \frac{dx}{g(x)f(q_\tau(x)|x)}}{\mu_2^2(K)\int_0^\infty \left(\frac{F^{02}(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right)^2 dx} \right\}^{\frac{1}{5}} n^{-\frac{1}{5}}$$

As discussed earlier, the bandwidth h_2 is less crucial, we will merely take $h_2 = h_1 n^{-0.5}$ and W to be a gaussian kernel. For each example, we consider three sample sizes: n=50,

n=100, and n=200. For each n, 100 replications of sample are performed. we compute the conditional quantile estimator for three values: $\tau = 0.05, 0.5$ and 0.95.

We will use the mean absolute deviation error ($MADE$) to assess the performance of the proposed estimates

$$MADE = \frac{1}{k}\sum_{j=1}^{k} |\hat{q}_\tau(x_j) - q_\tau(x_j)|$$

where x_j, $j = 1, ..., k$ are the regular grid points.

Example 1

As a first application, we consider the following parametric model

$$Y_i = (X_i - 0.5)^2 + \epsilon_i$$

where ϵ_i are independent $N(0, 0.05^2)$ and X_i are independent $N(0, 0.3^2)$ random variables truncated on $[0, 1]$. For this model, the Beta kernel is used to smooth with respect to x. Figure 1 depicts the results for a simulated samples of sizes: 50, 100, and 200. It shows the true conditional quantile function, and three local linear double kernel quantile regression estimates with Gaussian, Epanechnikov and Beta kernels. To assess the performances of the proposed estimates and their variations, we performed a simulation study with various sample sizes and 100 replications for each. Figure 2 presents boxplots of the MADE values of three local linear double kernel quantile estimates (Gaussian, Epanechnikov and Beta kernels). It is shown from this figure, that we have a serious bias near $x = 1$, this bias is due to the design density that is monotonic decreasing within $[0, 1]$. Figure 2 shows clearly that the Beta based estimates outperform the two other estimates. The difference seems to be less important when $\tau = 0.5$.

Example 2

As a second application, let us consider the following model that has a left bounded support:

$$Y_i = exp(-X_i) + exp(-4(X_i - 1)^2) + \epsilon_i$$

where ϵ_i are independent $N(0, 0.05^2)$ and X_i are left truncated $N(0, 1)$ on $[0, \infty)$. Since the support is left bounded, we will use the Gamma kernel to smooth with respect to x. The simulation results of this example are summarized in Figure 3 and the boxplot of the 100 $MADE$ values are also displayed in Figure 4. Here again, note that the Gamma-based estimate outperforms its competitors : the Gassian and Epanechnikov-based estimates.

5.3 Real data example

As a final example, let us consider the classical *Old Faithful Geyser Data* that consists of 299 observations representing waiting time between eruptions (Y) and the duration of the eruption (X) for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. This data set can be obtained easily from package `MASS` in `R`. In figure 5 we plotted a collection of estimated conditional quantiles (at levels: 0.1, 0.25, 0.5, 0.75, and 0.9) by using the local linear double kernel quantile regression with the use of gaussian or gamma kernel for the covariate variable. The bandwidth values are selected as explained in (5.1)

6 Proof of theorems

We will prove the theorems for Beta kernels only because the proof for Gamma kernels is quite similar. Recall that for Beta kernels $\psi(x) = x(1-x)$. Let us put

$$\Omega(\frac{y-Y_i}{h_2}) = m(X_i, y) + \epsilon_i(y) \tag{9}$$

where $m(x,y) = E(\Omega(\frac{y-Y_i}{h_2})\,|X_i = x)$, while $E(\epsilon_i(y)\,|X_i = x) = 0$.

Lemma 1. *Under the assumptions 1-6, we have*

$$m(x,y) = F(y|x) + \frac{h_2^2}{2}\mu_2(W)f^{10}(y|x) + o(h_2^2)$$

Moreover, if we put $\hat{F}(y|x)$ for the local linear estimator of $m(x,y)$, then, under conditions 1-5, as $n \to \infty$, we have

$$E(\hat{F}(y|x) - m(x,y)) = \begin{cases} \frac{1}{2}x(1-x)F^{02}(y|x)h_1 + O(h_1^2) & \text{if } x \in S_I; \\ \frac{1}{2}(2+k)F^{02}(y|x)h_1^2 + O(h_1^3) & \text{if } x \in S_B \end{cases}$$

and

$$Var(\hat{F}(y|x)) = \begin{cases} \frac{Var(\epsilon_i(y)|X_i=x)}{2\sqrt{\pi x(1-x)}g(x)n\sqrt{h}} + o(\frac{1}{n\sqrt{h}}) & \text{if } x \in S_I; \\ \frac{Var(\epsilon_i(y)|X_i=x)\Gamma(2k+1)}{2^{2k+1}\Gamma^2(k+1)g(x)} + o(\frac{1}{nh}) & \text{if } x \in S_B. \end{cases}$$

Proof. Standard arguments enable to write

$$\begin{aligned} m(x,y) &= E\left(\Omega(\frac{y-Y_i}{h_2})|X_i = x\right) = \int \Omega(\frac{y-u}{h_2})f(u|x)du \\ &= h_2\int \Omega(v)f(y-h_2v|x)dv = \int W(v)F(y-h_2v|x)dv \end{aligned}$$

Then Taylor's expansion together with the fact that $\int W(v)dv = 1$, $\int vW(v)dv = 0$ and $\mu_2(W) = \int v^2W(v)dv$, lead to

$$m(x,y) = F(y|x) + \frac{h_2^2}{2}\mu_2(W)f^{10}(y|x) + o(h_2^2).$$

It follows that for $h_2 \to 0$,

$$m(x,y) \approx F(y|x).$$

By using this approximation and the relationship (9), we might write

$$\Omega(\frac{y-Y_i}{h_2}) = F(y|X_i) + \epsilon_i(y).$$

To estimate $m(x,y)$, we will use the local linear estimator $\hat{F}(y|x)$ proposed by Chen (2002). This estimator uses asymmetric kernel in x direction. The remaining part of the proof follows from Theorem 1 of Chen(2002). □

Lemma 2. *Under the conditions 1-6, we have*

$$Var(\epsilon_i(y)|X_i = x) = F(y|x)\,(1 - F(y|x)) - h_2 f(y|x)\alpha(W) + o(h_2),$$

where $\alpha(W) = 2\int \Omega(v)W(v)vdv$.

Proof. By definition of $\epsilon_i(y)$, we have for any fixed y

$$\begin{aligned} Var(\epsilon_i(y)|X_i = x) &= Var(\Omega(\frac{y-Y_i}{h_2})|X_i = x) \\ &= E(\Omega^2(\frac{y-Y_i}{h_2})|X_i = x) - E^2(\Omega(\frac{y-Y_i}{h_2})|X_i = x). \end{aligned}$$

Straightforward algebra shows that

$$\begin{aligned} E(\Omega(\frac{y-Y_i}{h_2})|X_i = x) &= \int \Omega(\frac{y-u}{h_2}) f(u|x)du \\ &= F(y|x) + \frac{1}{2}\mu_2(W) f^{10}(y|x) h_2^2 + o(h_2^2) \end{aligned}$$

and

$$\begin{aligned} E(\Omega^2(\frac{y-Y_i}{h_2})|X_i = x) &= h_2 \int \Omega^2(v) f(y - h_2 v|x)dv \\ &= 2\int \Omega(v)W(v)F(y - h_2 v|x)dv \\ &= F(y|x) - h_2 f(y|x)\alpha(W) + o(h_2). \end{aligned}$$

where $\alpha(W) = 2\int \Omega(v)W(v)vdv$. This concludes the proof. □

Proof of Theorem 1

Proof. By Lemma 1, we have

$$\begin{aligned} Bias(\hat{F}(y|x)) &= E\left(\hat{F}(y|x) - F(y|x)\right) \\ &= E(\hat{F}(y|x) - m(x,y)) + \frac{1}{2}h_2^2\mu_2(W)f^{10}(y|x) + o(h_2^2) \\ &= E(\hat{F}(y|x) - m(x,y)) + o(h_1^2), \quad \text{because} \quad h_2 = o(h_1). \end{aligned}$$

For any fixed y the term $E(\hat{F}(y|x)-m(x,y))$ is the bias of mean regression and it is given by Lemma 1, so by substituting this term in the last expression, we obtain $Bias(\hat{F}(y|x))$ given in Theorem 1.

Now, from Lemma 1 and Lemma 2 we have:

$$Var\{\hat{F}(y|x)\} = \begin{cases} \frac{\sigma_1^2(x,y)}{2\sqrt{\pi}\sqrt{\psi(x)}g(x)n\sqrt{h_1}} + o(\frac{1}{n\sqrt{h_1}}) & \text{if } x \in S_I; \\ \\ \frac{\sigma_1^2(x,y)\Gamma(2c+1)}{2^{2c+1}\Gamma^2(c+1)g(x)nh_1} + o(\frac{1}{nh_1}) & \text{if } x \in S_B, \end{cases}$$

where $\sigma_1^2(x,y) = F(y|x)(1-F(y|x)) - h_2 f(y|x)\alpha(W)$.

Finally, by remarking that $h_2 = o(h_1)$, we derive the asymptotic variance of $\hat{F}(y|x)$ given in Theorem 1. □

Proof of Theorem 2

Proof. By using Taylor about $q_\tau(x)$ and ignoring the higher terms, we obtain

$$\hat{F}(\hat{q}_\tau(x)|x) \approx \hat{F}(q_\tau(x)|x) + \hat{f}(\hat{q}_\tau(x)|x)(\hat{q}_\tau(x) - q_\tau(x)),$$

then

$$\begin{aligned} \hat{q}_\tau(x) - q_\tau(x) &\approx -\left(\frac{\hat{F}(q_\tau(x)|x) - \hat{F}(\hat{q}_\tau(x)|x)}{\hat{f}(\hat{q}_\tau(x)|x)}\right) && (10) \\ &\approx -\left(\frac{\hat{F}(q_\tau(x)|x) - \tau}{\hat{f}(\hat{q}_\tau(x)|x)}\right) && (11) \\ &= -\left(\frac{\hat{F}(q_\tau(x)|x) - F(q_\tau(x)|x)}{\hat{f}(\hat{q}_\tau(x)|x)}\right) && (12) \\ &\approx -\left(\frac{\hat{F}(q_\tau(x)|x) - F(q_\tau(x)|x)}{f(q_\tau(x)|x)}\right) && (13) \end{aligned}$$

the last approximation using equation given in Lemma 6 of Samanta (1989) :

$$\hat{f}(q_\tau^*(x)|x) = f(q_\tau(x)|x) + o_p(1),$$

where $q_\tau^*(x)$ is some random point between $\hat{q}_\tau(x)$ and $q_\tau(x)$. Now, the proof of the theorem is an application of the Theorem 1. □

Acknowledgement

The authors wish to express their appreciation to the associate editor, the two referees and Dr. Farid Beninel for their helpful suggestions and comments.

References

Abberger,K(1998). Cross Validation in Nonparametrique Quantile Regression. *Allgemeines Statistisches Archiv*, 82 ,149–161.

Brown,B.M., and Chen, S.X. (1999). Beta-Bernstein smoothing for regression curves with compact supports, *Scandinavian Journal of Statistics*, 26, 47–59.

Cai, Z. (2002). Regression quantiles for time series data, *Econometric Theory*, 18, 169–192.

Cai, Z. and Wang, X. (2008). Nonoparametric methods for estimating conditional VaR and expected shortfall, *Journal of Econometrics*, 147, 120–1300.

Chen, S.X. (1999). "A Beta kernel estimator for density functions with compact supports," *Computational Statistics & Data Analysis - Journal*, 31, 131–145.

Chen, S. X. (2000a). Beta kernel smoother for regression curves, *Statistica Sinica* 10, 73–91.

Chen, S. X. (2000b). "Probability density function estimation using gamma kernels," *Annals of the Institute of Statistical Mathematics*, 52, 471–480.

Chen, S. X. (2002). Local Linear Smoother Using Asymmetric Kernels, *Annals of the Institute of Statistical Mathematics*, 54, 312–323.

Fan, J. Q. (1993). Local linear regression smoothers and their minimax efficiencies, *Annals of Statistics*, 21, 196–216.

Fan, J., Hu, T-C., and Troung, Y.K. (1994). Robust Non-parametric Function Estimation, *Scandinavian Journal of Statistics*, 21, 433–446.

Fan, J. and Gijbels, I. (1996). Local polynomial Modeling and Its Applications, *Chapman and Hall*, London (1996).

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83, 189–206.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, 49, 79–99.

Fan, J., and Yao, Q. (2003) Nonlinear time series: nonparametric and parametric methods. *Springer*.

Goh, S.C. (2012). Design-Adaptive Nonparametric Estimation of Conditional Quantile Derivatives, *Journal of Nonparametric Statistics*, 24, 597-612.

Jones, M.C. (1990). The performance of kernel density functions in kernel distribution function estimation, *Statistics and Probability Letters*, 9, 129-132.

Koenker, R., and Basset, G. (1978). Regression quantile, *Econometrica*, 46, 33–50.

Samanta, M. (1989). Non-parametric estimation of conditional quantiles, *Statistics and Probability Letters*, 7, 407–412.

Seifert, B., and Gasser, Th. (1996a). Finite sample variance of local polynomials: analysis and solutions, *Journal of the American Statistical Association*, 91, 267–275.

Seifert, B., and Gasser, Th. (1996b). Variance properties of local polynomials and ensuing modifications, *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Hardle and M. G. Schimek, Heidelberg: Physica-Verlag 5–79.

Yu, K., and Jones, M.C. (1998). Local Linear Quantile Rregression, *Journal of the American Statistical Association*, 93, 228–237.

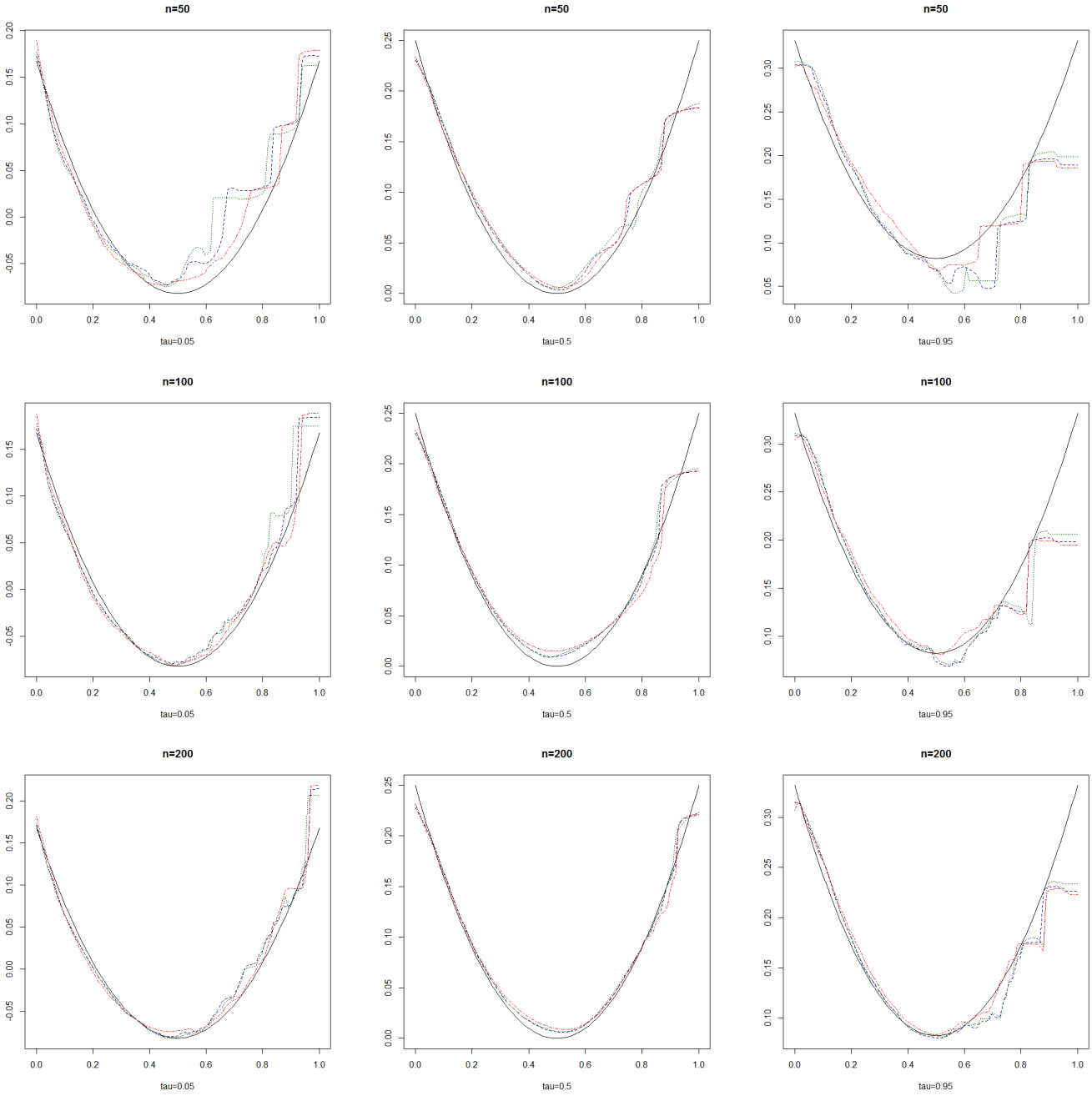


Figure 1: Conditional quantile estimate for example 1. Solid black line: true conditional quantile function, dashed blue line: estimator using a gaussian kernel at x, dashed green line: estimator using an Epanechnikov kernel at x, and dashed-dotted red line: estimator using a Beta kernel at x.

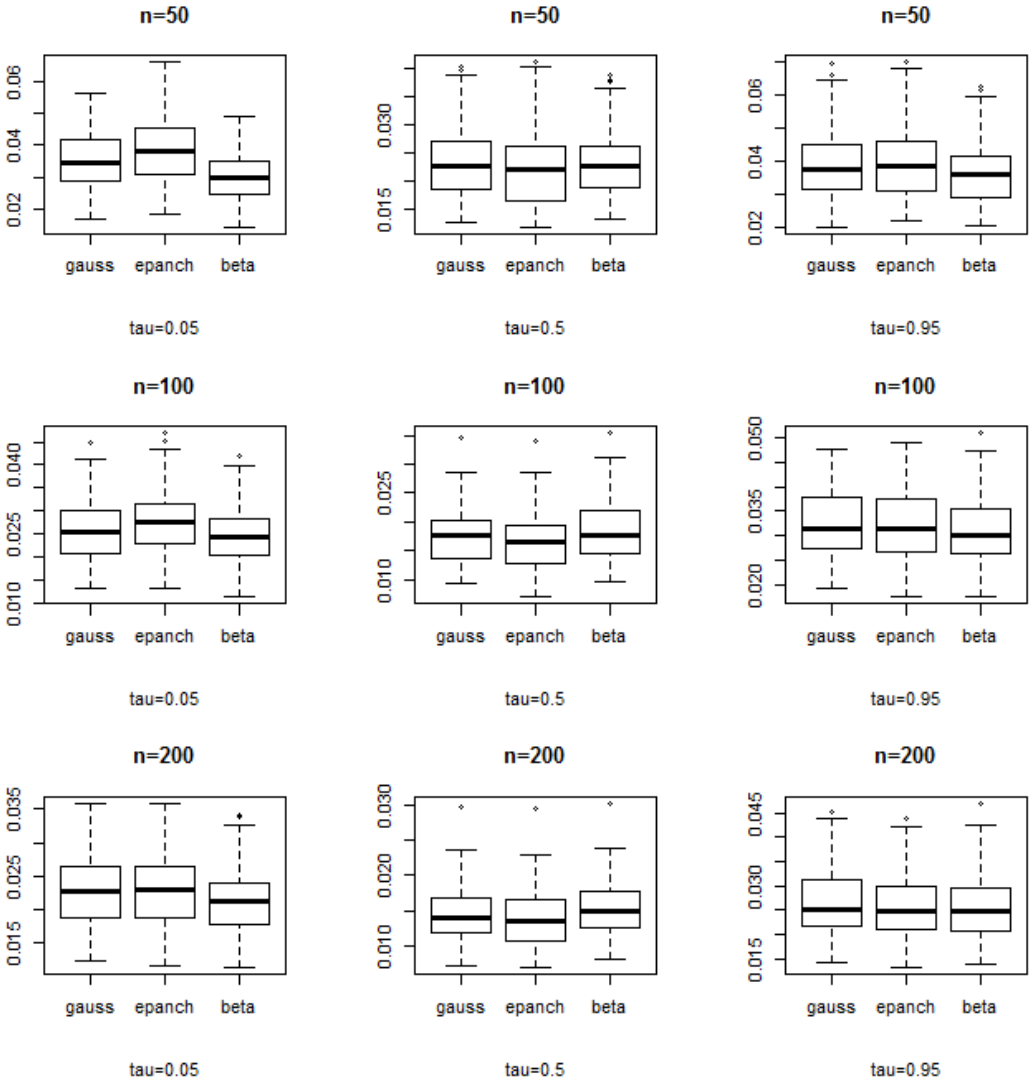


Figure 2: Boxplots of the 100 MADE values for conditional quantile estimate using gaussian, Epanechnikov and Beta kernels at x.

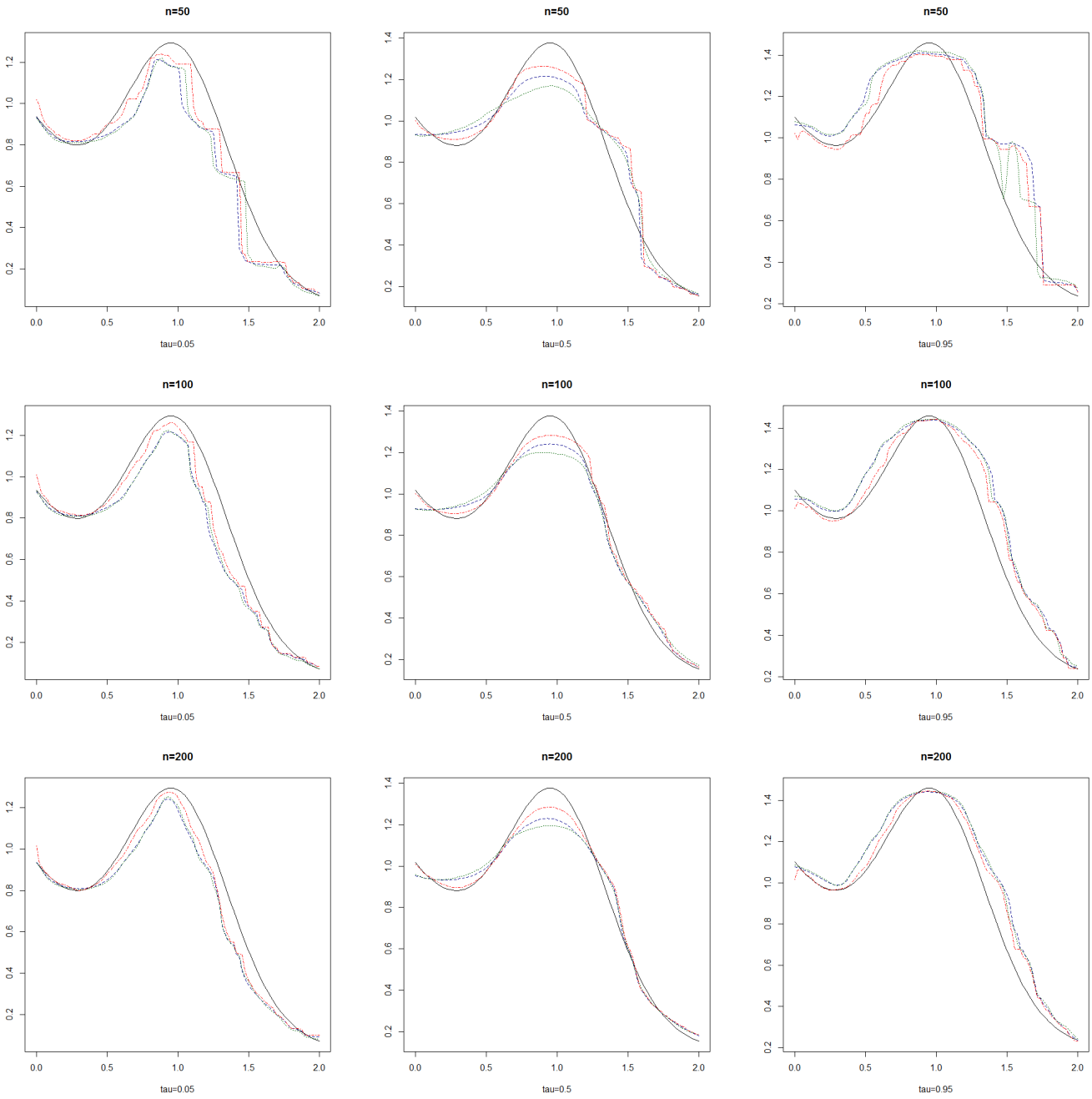


Figure 3: Conditional quantile estimate for example 2. Solid black line: true conditional quantile function, dashed blue line: estimator using a gaussian kernel at x, dashed green line: estimator using an Epanechnikov kernel at x, and dashed-dotted red line: estimator using a Gamma kernel at x

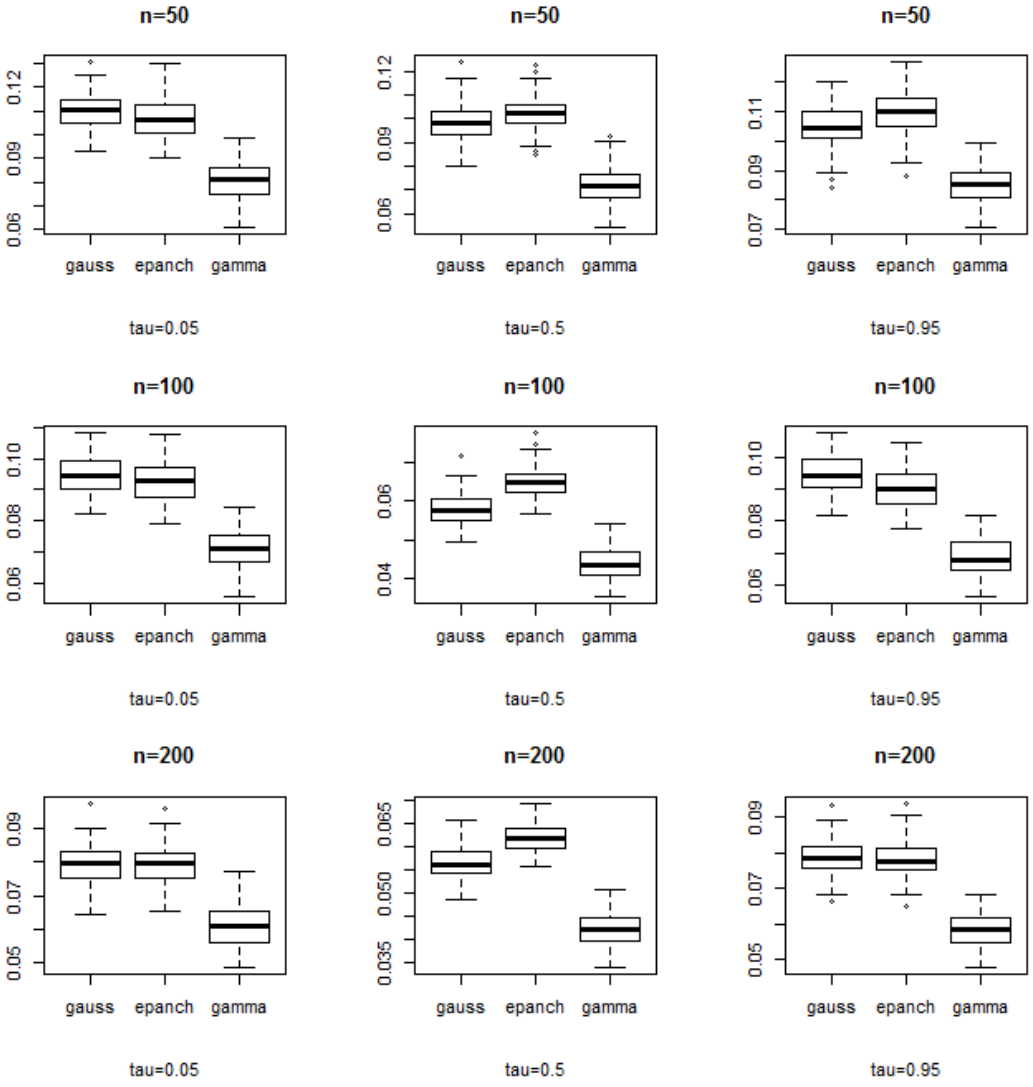


Figure 4: Boxplots of the 100 MADE values for conditional quantile estimate using gaussian, Epanechnikov and Gamma kernels at x.

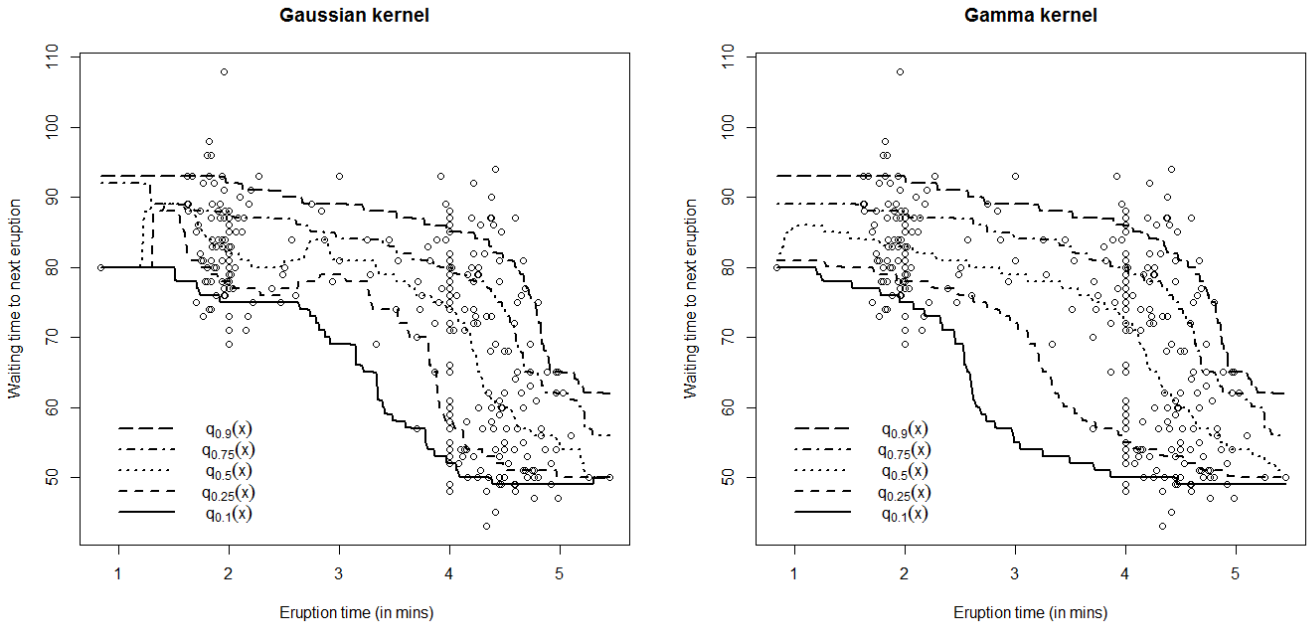


Figure 5: Collection of estimated conditional quantiles with the use of gaussian/gamma kernel for the covariate variable.

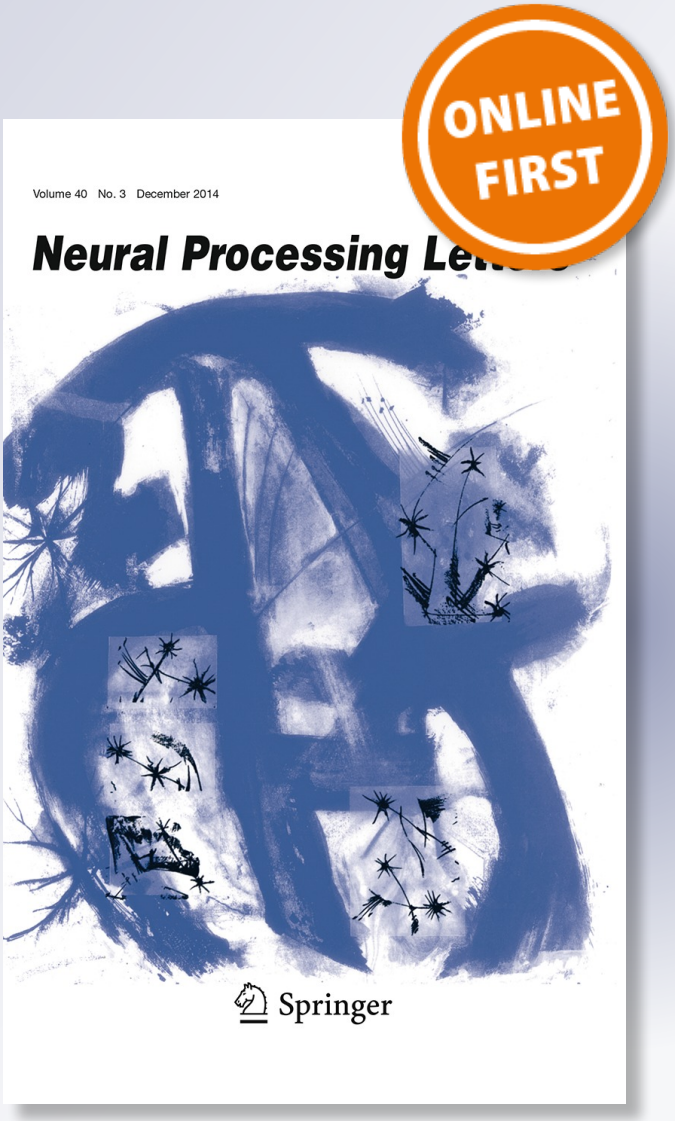
Reference Curves Estimation Using Conditional Quantile and Radial Basis Function Network with Mass Constraint

M.-Anas Knefati, Pierre E. Chauvet, Sylvie N'Guyen & Bassam Daya

Neural Processing Letters

ISSN 1370-4621

Neural Process Lett
DOI 10.1007/s11063-014-9399-9



Springer

Neural Process Lett
DOI 10.1007/s11063-014-9399-9

Reference Curves Estimation Using Conditional Quantile and Radial Basis Function Network with Mass Constraint

M.-Anas Knefati · Pierre E. Chauvet · Sylvie N'Guyen · Bassam Daya

Abstract This paper focuses on the improvement of reference curves building $Y = q(X)$ using a fast algorithm, robust against outliers. Our method consists in plugging a radial basis function neural network in the local linear quantile regression estimation proposed by Yu and Jones (QYJ). This neural network (QRBFc) is designed with a constructive algorithm, introducing a constraint on its integral over the input space. After explaining the different models and algorithms, we compare the QYJ and QRBFc estimators with the quantile regression neural network (QRNN) implemented by A. J. Cannon through simulations with a known underlying model using the R software. We observe that the QRBFc estimator reduces the mean absolute deviation error obtained with other estimators by about 16 %, the introduction of the constraint allowing to lower the number of neurons and therefore the computation time. Finally, using a database of 416 electroencephalograms recorded on preterm infants, we compare the QYJ, QRBFc and QRNN models for the building of brain maturation curves which are based on the dependence of the mean duration of interburst intervals (called IBIs—periods of quiescence between periods of normal electrical activity) with the age. The pathological infants represent 12 % of the total population. Denoting by S_A the set of individuals whose coordinates (age, mean IBI length) are above the 90 %-quantile curve, the QRBFc network

M.-A. Knefati ( )
Département de Mathématiques, P2MI Université de Poitiers, 86962 Futuroscope-Chasseneuil, France
e-mail: maknefati@hotmail.com

P. E. Chauvet
LUNAM Université, Université Catholique de l'Ouest and LARIS EA 4094, 3 place André-Leroy, BP 10808, 49008 Angers, France
e-mail: pierre.chauvet@uco.fr

S. N'Guyen
LUNAM Université, Child Neurology Unit University Hospital and LARIS EA 4094, 4 rue Larrey, 49000 Angers, France
e-mail: sylvie.nguyenthe@gmail.com

B. Daya
IUT Saida, B.P 813, Saida, Lebanon

Published online: 22 November 2014

improves by 16.5 % the number of pathological infants in S_A compared to QYJ, when QRNN proved to be too unstable.

Keywords Quantile regression · Radial basis function neural network · Reference curve · Brain maturation

1 Motivation

The reference curves are part of the basic tools of the physician practices, in that they can decide on the vulnerability of an individual against a certain disease. Their construction, which is based on measurements made on a given set of individuals, shows a partition between normal and at risk individuals. It was during a project in pediatric neurology that we have developed a general method for construction of reference curves based on conditional quantiles and Radial basis function networks (RBFN). Because cerebral injury in newborns tends to be clinically silent, tools and techniques for neurological evaluation are essential. The electroencephalogram (EEG) is such a bedside, non-invasive and low cost technique. The EEG consists in recording the spontaneous electrical activity of the brain through several electrodes placed on the scalp. In preterm infants, the normal background EEG activity has the unique characteristic of being spontaneously discontinuous with periods of electrical activity alternating with periods of quiescence - called interburst intervals (IBIs). Figure 1 illustrates this phenomenon, which is normal in infancy. The diagnostic and prognostic value of neonatal EEG abnormalities in the preterm infant are well established and the IBI duration have been shown to be related to abnormal brain maturation in preterm infants [14]. In a recent study comparing maturation of cerebral activity and cortical folding, IBI duration was the only parameter significantly linked to morphologic brain maturation [1].

We have developed inside a dedicated web portal a Java application allowing the physician to extract automatically all the IBIs from an EEG. In this area, artificial neural networks are used generally for features extraction and classification, particularly for brain computer

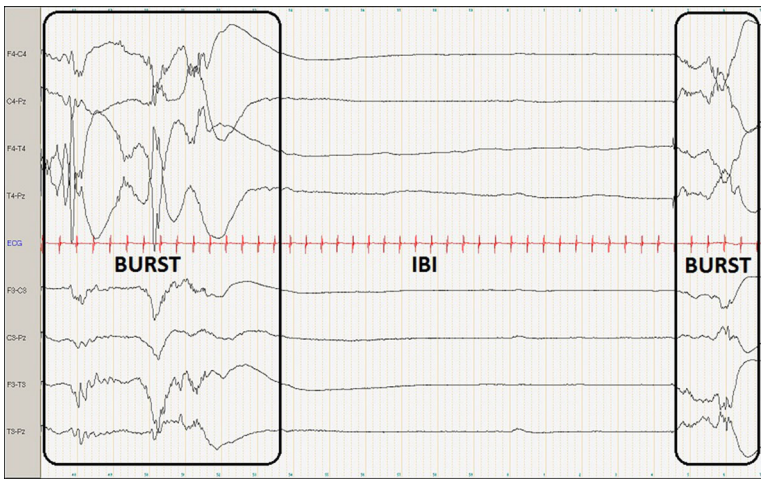


Fig. 1 Normal discontinuous EEG tracing in a preterm infant at 26 weeks of gestationnal age consisting in burst of electrical activities separated by interburst intervals (IBI)

interfaces (see for example [12,13]). In this work, the features (IBIs) are already detected and the objective is to build reference curves for brain maturation using these data, easy to recalculate from new sets of EEG and robust against artefacts and entry mistakes (e.g. age).

Our paper is structured as follows. First we start by reviewing the state of the art in Sect. 2. Then, after a few reminders about the basics of conditional quantile, we introduce in Sect. 3 the RBFN modeling of the conditional quantile curve; the learning algorithm is based on a constructive approach with a control of the integral of the model over its input domain (its mass), and the coupling with the conditional quantile estimator of Yu and Jones (QYJ), to obtain the so-called QRBFc model. Experiments are carried out in Sect. 4, where we compare on simulated data with known properties the QYJ, the QRBFc, the quantile regression neural network models (QRNN), and the possible improvements of the mass constraint. Finally we present and discuss in Sect. 5 the brain maturation reference curves obtained with the different models, based on the correlation between the mean length of interburst intervals in the newborn EEG and its age. We conclude this work in Sect. 6, providing our future work directions.

2 State of the Art

Conditional quantile regression has gained particular attention during the recent three decades due to their useful applications in various disciplines, such as finance, economics, medicine, and biology. See for exemple Fan and Gijbels [9] and Cai and Wang [3]. Of particular interest is the conditional median which is more explainable and more robust than the mean regression function for asymmetric conditional distributions. For $\tau \in (0, 1)$, the quantile regression function gives the τth quantile $q_\tau(x)$ in the conditional distribution of a response variable Y given $X = x$. It measures the effect of covariates not only in the center of the population, but also in the lower and upper tails. For x varying in a given real interval, $q_\tau(x)$ is a reference curve that predicts vulnerability of an individual with the probability τ or $1-\tau$ as its associated measured pair $(X = x, Y = y)$ is below or above the curve. A classical approach to evaluate the conditional quantile function from a sample (X_i, Y_i) on a discrete finite set is the local linear quantile regression method from Yu and Jones (denoted QYJ) in [16]. As its name suggests, $q_\tau(x)$ is approximated by a piecewise linear curve, resulting is a broken line that emphasizes local fidelity to the detriment of regularity, difficult to interpret as a reference chart.

Cannon [2] developed in R the QRNN package (for Quantile Regression Neural Network) based on the work from Taylor [11] in the field of time series forecasting. J. W. Taylor used a one hidden-layer feedforward neural network (FNN) to fit a non-local quantile model. FNN where successfully used in several domains like robotics and image processing (see for example [6,7]) as non linear parametric models. FNN realizes a global approximation of the data, and by varying the number of neurons it is possible to slide between high/low precision and smoothness of the quantile modelling. However, standard non-linear FNN with sigmoid transfer functions does not allow local settings on some parts of the data, and the learning process can be very slow. RBFN has the advantage to allow a faster constructive approach since each neuron puts emphasis on local data: neurons can be added iteratively, as in Chen et al. [4].

3 The RBFN Modeling of the Conditional Quantile

In this section, we first start with an overview of the conditional quantile. Next, the RBFN model is presented by mentioning its parameters which are the centers and the weights, and

we explain how to calculate them introducing constraint on the weights. Finally, we end this section by explaining our method to improve the estimation of the conditional quantile with a RBFN network, leading to the QRBFc method.

3.1 An Overview of the Conditional Quantile

Let (X, Y) be a bivariate random variable, $F(y|x)$ the conditional distribution function of Y given $X = x$, and $\tau \in (0, 1)$. The τth conditional quantile, noted by $q_\tau(x)$, is given by

$$q_\tau(x) = \inf\{y \in \mathbb{R} : F(y|x) \geq \tau\} \equiv F^{-1}(\tau|x) \tag{1}$$

or equivalently, by

$$q_\tau(x) = \arg\min_{\theta} E\{\rho_\tau(Y - \theta)|X = x\}, \tag{2}$$

where ρ_τ is the "check function" $\rho_\tau(u) = 0.5(|u| + (2\tau - 1)u)$.

The building of predictive intervals is an important application of quantile regression. Suppose the observations can be modelled as $Y_i = m(X_i) + \gamma\epsilon_i$ where $m(.)$ is an unknown function, γ is unknown reel parameter and the residual ϵ_i are uncorrelated random variables with zero mean and one variance. Then the quantile function can be written as

$$q_\tau(x)) = m(x) + \gamma F_\epsilon^{-1}(\tau), \tag{3}$$

where $F_\epsilon(.)$ denotes the distribution function of ϵ. A predictive interval is an interval that predicts, with certain coverage probability, the future value of the response variable Y for a given covariate $X = x$. The pairs of extreme conditional quantiles $q_{inf}(x)$ and $q_{sup}(x)$ map out a conditional prediction interval within which one expects the majority of individual points to lie. These "reference curves" are popular in medicine (see, e.g, Cole in [5]) and have provided a stimulus for much of the recent statistical work in this area.

Parametric techniques for estimating conditional quantile can be efficient if the underlying functions are correctly specified. But a misspecification may cause serious bias, and model constraint may distort the underlying distribution. Therefore, we will concentrate in this article on the nonparametric quantile regression, with the advantage that little or no restrictive prior information on functionals is needed.

Nonparametric estimation of conditional Quantile has been tackled by several authors, with direct and indirect methods. Direct methods use the "check" function ρ_τ, taking roots in definition (2). See, for instance, Fan et al. [8] and Yu and Jones [16] for a local linear estimate of conditional quantiles. Indirect methods, inspired from (1), are performed in two steps: the estimation of the conditional distribution is performed first, then the inverse of the obtained estimator is used to estimate the desired conditional quantile. Examples of indirect methods are provided by Cai and Wang [3]. Our work is based on the direct estimator proposed by Yu and Jones [16], because of its robustness in the y direction. The idea is to approximate the unknown τth conditional quantile $q_\tau(x)$ by the linear function

$$q_\tau(z) = q_\tau(x) + q'_\tau(x)(z - x) \equiv a + b(z - x)$$

for z in the neighborhood of x. This motivated us to define an estimator by setting $\hat{q}_\tau(x) = \hat{a}(x)$, with

$$\left(\hat{a}(x), \hat{b}(x)\right) = \arg\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \rho_\tau\left(Y_i - a - b(X_i - x)\right) \times K\left(\frac{x - X_i}{h}\right), \tag{4}$$

where K is a kernel density function and h is the smoothing (also called bandwith) parameter, which is a nonnegative number controlling the size of the local neighborhood. Yu and Jones

have proposed a "rule-of-thumb" for selecting h:

$$h_\tau = h_{\text{mean}} \left(\frac{\tau(1-\tau)}{\phi(\Phi^{-1}(\tau))^2} \right)^{\frac{1}{5}}, \tag{5}$$

where h_{mean} is the optimal choice of h for regression mean estimation, ϕ and Φ are the standard normal density and distribution functions. They recommended to use the technique proposed by Ruppert et al. [10] for selecting h_{mean}.

The solution of the problem (4) can be obtained by using the iteratively reweighted least squares algorithm, see for exemple Yu [15]. We will denote by QYJ this algorithm (as well as the resulting quantile curves) used in conjunction with the rule-of-thumb from Yu and Jones in what follows.

3.2 The RBFN Neural Network Model

A RBFN can be viewed as a FNN with a specific structure that allows an easier constructive approach. It has one hidden layer in which each neuron computes its output using a radial basis function (RBF) receiving the inputs, and an output layer which builds a linear weighted sum of hidden neuron outputs and supplies the network's response.

An RBF function is a function $\phi : [0, \infty) \to \mathbb{R}$ that depends only on the distance from some point c, called a center, so that it has the form $\phi(\|x - c\|)$. In other words, a radial basis function is radially symmetric with respect to a given norm. We generally choose the Gaussian function $\phi(\|x - c\|) = e^{-\frac{\|x-c\|^2}{2}}$.

Because our goal is to produce reference curves, i.e. real functions, we use RBFN with only one output neuron. Its model is given by:

$$\hat{f}(x) = \sum_{j=1}^{N} \omega_j \phi(\|x - c_j\|), \tag{6}$$

where $x \in \mathbb{R}^d$, N is the number of hidden neurons, ω_j $(1 \leq j \leq N)$ are the weights of network output layer, c_j $(1 \leq j \leq N)$ is the center of the jth hidden neuron, $\phi(x)$ is a gaussian RBF function and $\|.\|$ denotes the distance function that is taken, in general, to be the Euclidean norm. In this work we use the Mahalanobis distance, because it takes into account the correlations inside the data set and is scale-invariant; in our case, it contributes to a better adjustment of the neurons width with the data. This distance is defined as

$$\|x - c\| = \sqrt{(x-c)^t \Sigma^{-1} (x-c)}, \tag{7}$$

where Σ is the covariance matrix of $(x_1, \ldots, x_n)$, and $x_i \in \mathbb{R}^d$ $(i = 1, \ldots, n)$ are the training data input.

We use in this work an iterative design of the network: neurons are added one at a time until the mean sum-squared error (MSE) falls beneath an error goal or a maximum number of neurons has been reached. The choice of the number N of neurons on the hidden layer of FNN and RBFN networks is crucial. A low number means a very poor performance, or fidelity, of the network. Instead, a large number of neurons will allow the network to fit exactly all the data (resulting in a very low MSE), including noise and biased observations: its regularity will be low. Because N is an integer defining the structure of the network, it cannot be adjusted like the synaptic weights. It exists for FNN some meta-heuristics (like genetic algorithms) to adjust N, but this is a slow process. Unlike FNN, RBFN authorizes

a faster constructive approach because of the locality of the radial basis transfer function in the hidden layer.

Let $x_i \in \mathbb{R}^d$ be the training data inputs and $y_i \in \mathbb{R}$ be the training data outputs for $i = 1,\ldots, n$. We now explain our RBFN learning algorithm.

3.2.1 Computation of the Centers

The algorithm for computation of the centers can be stated as follows:

- Provide the error threshold (err) and the maximum number of neurons $Nmax$ ($Nmax \leq n$).
- Initialize N to 0: the initial network does not have any radial functions.
- Store the output vector $(y_1, ..., y_n)$ in an other vector $(y_1^*, ..., y_n^*)$.
- While the total error $(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2)$ is greater than err and $N < Nmax$ do:
  - calculate the network output $\hat{f}(x_i)$ using (6) for each input and the error $r_i = |\hat{f}(x_i) - y_i^*|$ for i =1, . . ., n;
  - find the input x_l that causes the greatest error, with $l = \arg\max_{i\in\{1,...,n\}} r_i$;
  - add a radial basis function whose center is this entry ($c_N = x_l,\ N = N + 1$);
  - recalculate the weight vector $W = (\omega_1, ..., \omega_n)$;
  - recalculate $\hat{f}(x_i)$, i =1,...,n;
  - initialize to 0 the output value that causes the greatest error, i.e. $y_l^* = 0$ and $\hat{f}(x_l) = 0$, so it will not be chosen again.

We determined here the number of hidden neurons N and the centers c_j. We explain in the next section the computation of the weights.

3.2.2 Computation of the Weights

Due to the fact that the mapping from hidden layer to output layer is linear, the weights computation becomes a linear problem. Minimization of the MSE yields to the well-known least square solution:

$$W = (\Phi^t\Phi)^{-1}\Phi^t Y, \tag{8}$$

where $Y = (y_1, ..., y_n)^t$ and

$$\Phi = \begin{pmatrix} \phi(x_1 - c_1) & \cdots & \phi(x_1 - c_N) \\ \vdots & \vdots & \cdots \quad \vdots \\ \phi(x_n - c_1) & \cdots & \phi(x_n - c_N) \end{pmatrix}$$

A first possibility for a robust calculation of the weights is to use Tikhonov regularization (also called ridge regression). Another suggestion to avoid possible numerical difficulties due to a singular or near-singular matrix, is to use the pseudo-inverse of a matrix which generalizes the inverse of a squared matrix. Any matrix can be factored using its singular value decomposition (SVD) from which its Moore-Penrose or generalized inverse can be obtained.

After testing these two methods we finally prefer the pseudo-inverse method, that provides the same results with far less computation time.

3.2.3 Computation of the Weights using a Mass Constraint

The objective in this section is to modify the previous algorithm to control the mass c of the RBFN model (its integral over the input domain). This is easier than controlling the energy of the RBFN (integral of the squared function), since we will be able to bring a linear constraint on the weights. In the case where the model is positive for all x, the mass of the model is equivalent to its norm L_1, and c appears as a global smoothing parameter. We calculate the weights of the model (6) such that it satisfies $\int_{\mathbb{R}^d} \hat{f}(x)dx = c$. We have

$$\int_{\mathbb{R}^d} \hat{f}(x)dx = \sum_{j=1}^{N} \alpha_j \omega_j,$$

where $\alpha_j = \int_{\mathbb{R}^d} \phi(\frac{\|x-c_{x_j}\|}{\sigma_j})dx$ for $j = 1, ..., N$. In our case, $\phi(\|x-c_{x_j}\|) = \exp(-\frac{\|x-c_{x_j}\|^2}{2})$ is a Gaussian function, then $\alpha_j = \alpha = (\sqrt{\pi})^d$ for all $j = 1, ..., N$. Therefore, for calculating the weights under the assumption that $\int_{\mathbb{R}^d} \hat{f}(x)dx = c$ we obtain the linear constraint: $\alpha \sum_{j=1}^{N} \omega_j = c$, with $\alpha = (\sqrt{\pi})^d$. To solve this least squares constrained problem, let introduce the Lagrange function given by

$$\begin{aligned} L(W, \lambda) &= \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 + \lambda \left(\alpha \sum_{j=1}^{N} \omega_j - c \right) \\ &= \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{N} \omega_j \phi_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{N} \omega_j - c \right), \end{aligned}$$

where λ is the Lagrange multiplier, $\phi_{ij} = \phi(\frac{\|x_i-c_{x_j}\|}{2})$ and W is the vector of weights.

The necessary conditions for optimality are

$$\frac{\partial L(W, \lambda)}{\partial W} = 0 \Rightarrow -\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{N} \phi_{ij} \omega_j \right) \phi_{ik} + d_0 \lambda = 0$$

and

$$\frac{\partial L(W, \lambda)}{\partial \lambda} = 0 \Rightarrow \alpha \sum_{j=1}^{N} \omega_j - c = 0.$$

Thus, we have to solve the $N + 1$ following equations with $N + 1$ variables:

$$\begin{cases} \sum_{j=1}^{N} \left(\sum_{i=1}^{n} \phi_{ij} \phi_{ik} \right) \omega_j + \alpha \lambda = \sum_{i=1}^{n} \phi_{ik} y_i, \quad k = 1, ..., N \\ \alpha \sum_{j=1}^{N} \omega_j = c \end{cases}$$

Like in paragraph III-B2, we use the SVD method to solve these equations.

3.3 Improvement of Conditional Quantile Estimation with a RBFN Network

We explain below our algorithm to improve the QYJ estimator using the RBFN model. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample from (X, Y), and assume the homoscedastic model:

$$Y_i = m(X_i) + \gamma \epsilon_i, \quad i = 1, ..., n, \tag{9}$$

where $m(.)$ is an unknown function, γ unknown reel parameter and the residuals ϵ_i are uncorrelated random variables of known distribution with zero mean and one variance. Therefore the conditional quantile function is defined by:

$$q_\tau(x) = m(x) + \gamma F_\epsilon^{-1}(\tau). \tag{10}$$

The proposed algorithm is explained as follow :

- Estimate first the conditional quantile function $q_\tau(x)$ using the QYJ estimator with the Yu and Jone method to select its smoothing parameter; we get $\hat{q}_\tau^{QYJ}(x)$.
- From (10), we have

$$\hat{m}^{QYJ}(x_k) = \hat{q}_\tau^{QYJ}(x_k) - \hat{\gamma} F_\epsilon^{-1}(\tau),$$

  where $\hat{\gamma}^2$ is the variance of the data outputs, and x_k, $k = 1, ..., n_0$ are either regularly spaced values along x-axis or the values $(X_1, ..., X_n)$ taken in the random sample itself.
- Use the RBFN model with entries x_k and $y_k = \hat{m}^{QYJ}(x_k)$ to approximate the unknown function $m(x)$ to obtain $\hat{m}^{QRBF}(x)$.
- Plug-in $\hat{m}^{QRBF}(x)$ and $\hat{\gamma}$ in (10) to get the improved RBFN conditional quantile estimator:

$$\hat{q}_\tau^{QRBF}(x) = \hat{m}^{QRBF}(x) + \hat{\gamma} F_\epsilon^{-1}(\tau) \tag{11}$$

  for all x in the input space.

We denote by QRBF (when we do not use the mass constraint) and QRBFc (when using the mass constraint) this algorithm and the resulting quantile curves in what follows.

4 Simulations with a Known Underlying Model

The following applications have been made using the R software, with the package "quantreg" from Koenker [17] for our implementation of the QYJ algorithm and the QRNN R-package implemented by Cannon [18].

Consider the model

$$Y_i = \exp(-X_i) + \exp\left(-4(X_i - 1)^2\right) + \epsilon_i$$

for $i = 1, ..., n$, where $\{\epsilon_i\}$ are independent and identically $\mathcal{N}(0, 1)$ random variables, and $\{X_i\}$ are exponential random variables with mean 1 independent from $\{\epsilon_i\}$. Then the true conditional quantile function is given by

$$q_\tau(x) = \exp(-x) + \exp\left(-4(x - 1)^2\right) + F_\epsilon^{-1}(\tau),$$

where F_ϵ is the cumulative $\mathcal{N}(0, 1)$ distribution function.

First, we compare for different sample sizes n and different quantile values τ the performance of the QRBF estimator without constraint with the QYJ estimator and the QRNN estimator with 5 neurons on the hidden layer (after several tries, 5 neurons appear to be a good compromise between the error goal and the calculation speed). For this purpose, we ran 100 replications per experiment for several combination of the parameters values n and τ. We also demonstrate the performance of the estimators in terms of the mean absolute deviation error (MADE):

$$MADE = \frac{1}{n_0} \sum_{j=1}^{n_0} |\hat{q}_\tau(x_j) - q_\tau(x_j)|,$$

where x_j, $j = 1, ..., n_0$ are regularly spaced values between 0 and 3.5.

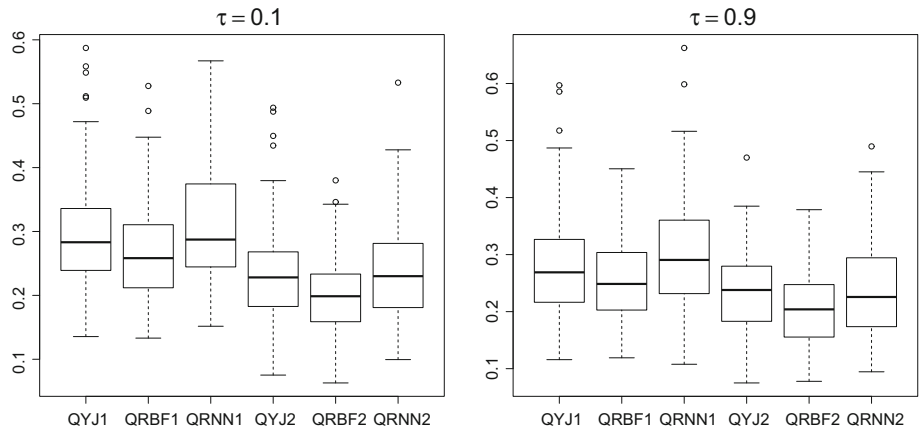


Fig. 2 Mean absolute deviation error boxplots obtained for sample sizes $n = 300$ and $n = 500$, with the two quantile values $\tau = 0.1$ and $\tau = 0.9$. QYJ_i denotes the Yu and Jones QYJ estimator, $QRBF_i$ its QRBF improvement, and $QRNN_i$ denotes the QRNN estimator, where i equals 1 for size 300 and 2 for size 500

Table 1 Median for 100 values of N, N_c and $MADE$ with mass constraint c

Sample size	$\tau = 0.9$	$\bar{N}$	$\bar{N}_1$	$\bar{N}_5$	$\bar{N}_{10}$	$\bar{N}_{20}$	$\bar{N}_{50}$	$\bar{N}_{100}$	$\bar{N}_{500}$	$\bar{N}_{1000}$	$\bar{N}_{5000}$
$n = 300$	Neurons	82	80	78	58	56	59	76	100	100	100
	MADE	0.242	0.238	0.24	0.24	0.24	0.24	0.24	0.284	0.329	2.962
$n = 500$	Neurons	102	87	92	62	51	59	88	166	166	166
	MADE	0.197	0.196	0.199	0.198	0.198	0.198	0.199	0.218	0.254	0.757

$\bar{N}$ denotes the median number of neurons on the RBFN hidden layer with no mass constraint and $\bar{N}_c$ this same median number using a mass constraint equals to c

We can see on Fig. 2 the boxplots of the 100 MADE values for the QYJ, QRBF and QRNN estimators, with $\tau = 0.1$ and $\tau = 0.9$. The QRBF estimator is better than its Yu and Jones counterpart and the QRNN estimator : (i) the MADE median of QRBF is lower than QYJ and QRNN; (ii) the MADE spread of QYJ and QRBF is smaller than QRNN. The computation time for the QRNN model is an average of 3 times that of QRBF and QYJ.

We then studied the impact of the mass constraint on the MADE median and on the number N of neurons in the hidden layer. We denote N_c the number of hidden neurons obtained by the constructive learning method under the constraint $\int_{\mathbb{R}} \hat{f}(x)dx = c$. We ran 100 replications for each simulation with different values of c. We found that the median of N_c values is smaller than the median of N values for all values of τ and some values of c.

In the Table 1, we see the median of $MADE$, N and N_c obtained for several values of c, in the case of $\tau = 0.9$ and sample sizes equal to 300 and 500, after 100 simulations. It appears clearly that controlling the RBFN mass (i.e calculating weights under constraint) uses, for some values of c, less neurons and the MADE obtained is nearly the same one obtained from RBFN without constraint. In our example, we can see from the Table 1 that $c = 10$ or $c = 20$ could be a suitable choice, but we think, according to this table, that an optimal choice should exist. This is the center of our actuel research.

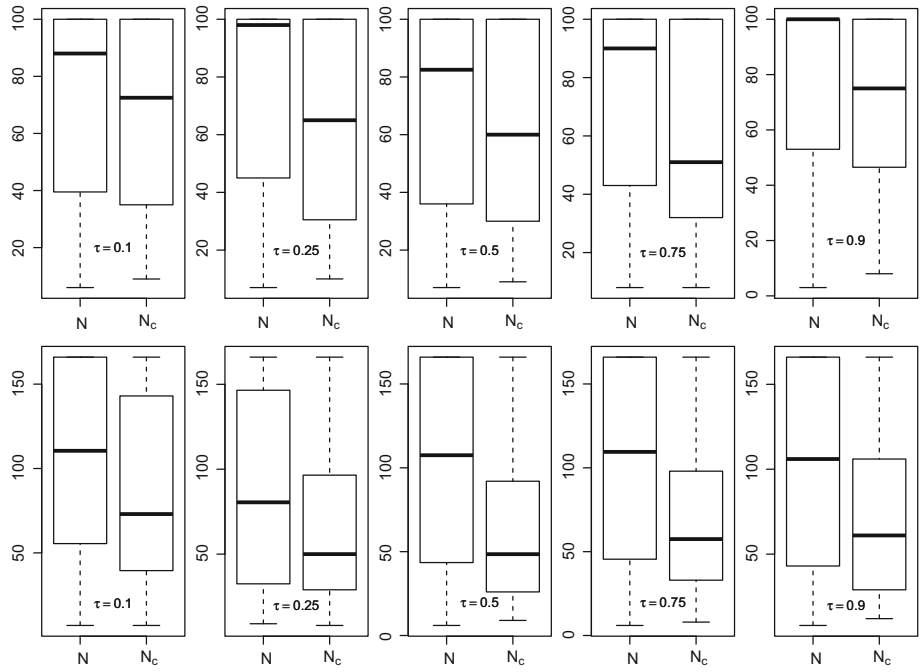


Fig. 3 Boxplot of the number of hidden neurons in the QRBF estimator for both cases: computing the weights without constraint (N) and with constraint (N_c with $c = 10$). First row: sample size equals to 300; second row: sample size equals to 500

Table 2 cfor the different estimators; *QYJ* denotes the classical Yu and Jones linear quantile estimator, *QRBF* its improvement using the RBFN without mass constraint, *QRBFc* the RBFN with mass constraint, and *QRNN* denotes the quantile regression feedforward neural network from A. J. Cannon ($\tau = 0.9$ and sample size: $n = 500$)

	QRNN	QYJ	QRBF	QRBF_{10}
Error	0.237	0.237	0.198	0.198
Time	2.69	0.91	1.385	1.01

The Fig. 3 illustrates, also, this fact for five quantile values, two sample sizes and three values of c. We think that the mass constraint act like a smoothing or regularization parameter and it would reduce the variance of the RBFN estimator by keeping, nearly, the same MADE error obtained by RBFN without constraint.

In Table 2, we put the median of MADE and computation time for every estimator, with the sample size equals to 500 and the quantile order equals to 0.9. According to this table we can see that the QRBF estimator (with or without mass constraint) reduces in this example the relative error by about 16 % for QRNN and for QYJ. Moreover, introducing the mass constraint in the QRBF estimator improves its computation time by reducing significantly its number of neurons. For this reason, we recommend to calculate the RBFN weights using the mass constraint as explained in paragraph (3.2.3), as we reduce the error against the QYJ estimator while we improve the speed against other neural network estimators.

5 Application to the Building of Brain Maturation Reference Curves

The next results were obtained from the automatic analysis of 416 EEG recorded between years 2003 and 2004 on newborn infants, with relative informations such that the infant outcome, age (from birthday and gestational), other neurological evaluations, etc. Moreover, each EEG in this database was visually analyzed and categorized as "artefact", "normal", "doubtful" or "pathological". A specific algorithm was built during this research to detect IBIs using an approach mimicking visual analysis. The algorithm is designed to study on each channel the variation of the signal estimated variance between contiguous short time-windows. An IBI on one channel is detected if this variation is lower than a given threshold and EEG (global) IBIs are finally computed as the intersection of channel's IBIs. More precisely, our algorithm is described by the following sequence of operations:

- each channel is filtered at 50Hz (that corresponds to frequency of electricity supply within European Union) with a second order Butterworth designed IIR filter;
- each channel is smoothed with a moving window using the simple average;
- each channel is processed to produce an estimate of the standard deviation on overlapped windows, a standard deviation series.
- this is a two steps operation; (i) for each resulting standard deviation series, if the difference between two successive values is lower than a given V_T threshold (in μV) the corresponding time intervals are aggregated and an IBI is defined by an aggregated time interval if it lasts at least m_1 seconds; (iii) finally, IBIs separated by less than m_2 seconds are regrouped.
- the intersection of IBIs between all the EEG channels is computed, and only the IBIs of a length greater than m_3 seconds are retained.

By comparisons between the IBIs marked by the specialist and the IBIs detected by this algorithm, we set the parameter values to $V_T = 15\ \mu V$, $m_1 = 1$ s, $m_2 = 0.5$ s and $m_3 = 1$ s.

For each EEG the mean IBI length is computed and plotted versus the gestational age of the infant. From the couple (age , mean IBI length) we have traced the reference curves using the QRNN, the QYJ and the QRBFc estimators. Figures 4 and 5 shows the results obtained with two levels of the quantile (0.1 and 0.9). It clearly appears that the improved

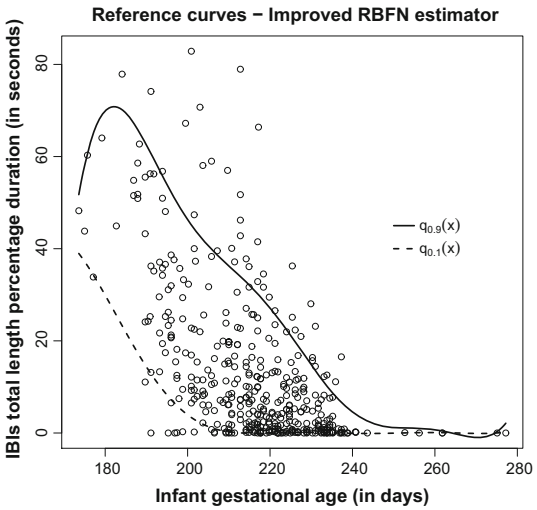


Fig. 4 Brain maturation reference curve for age in days (quantile levels $\tau = 0.1$ and $\tau = 0.9$) obtained with the improved QRBF_{10} estimator using a regular grid with 100 points and the automatic IBI analysis of 416 EEG recorded in 2003–2004

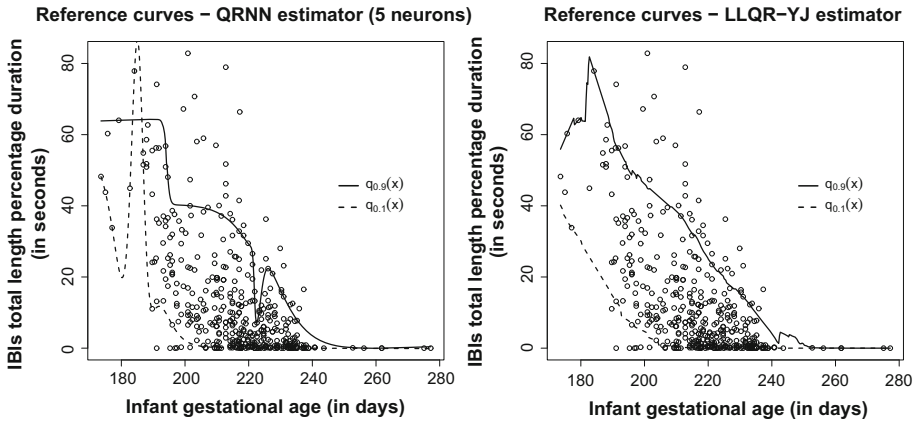


Fig. 5 Brain maturation reference curves obtained with the QRNN (*left*) and the QYJ (*right*) estimators

Table 3 Mean IBI's length and age in each EEG set; the $Total$ row is the initial set of 416 EEG

EEG set	Mean IBI length (s)	Mean age (days)
Total	13.3369	216.0832
QYJ		
S_A	39.4740	202.9307
S_{IN}	9.1188	219.2450
S_U	2.5125	218.0835
QRBF		
S_A	40.4922	216.1256
S_{IN}	11.9003	215.1669
S_U	1.1202	222.3292

QRBFc estimator (Fig. 4) produces smoother curves, easier to use, than the QYJ estimator. The QRNN estimator produces very different cuves from one try to another as the learning algorithm is stochastic. The QRNN curves are generally less smooth, with sometimes a totally flat curve for $\tau = 0.1$, and other times an inversion between the lower and the higher curves.

The two curves define three areas allowing one to build three disjoint EEG sets denoted S_A, S_{IN} and S_U, by taking EEG whose coordinates (age, mean IBI length) are respectively above the higher curve, between the two curves inclusively and under the lower curve. The union of these three sets is equal the initial total set, and the population at risk is defined by the set S_A. To compare these sets we have first calculated for each one the mean of the EEG mean IBI lengths and the mean of the ages (from birthday): the results are provided in Table 3. One can see that the sets S_A defined by QYJ and QRBFc have a longer mean IBI length by about three times that of the total set. This is perfectly consistent with the following clinical outcome: the higher the mean IBI length, the greater the risk of abnormal maturation is important. The mean age of the population in S_A is about 203 days for QYJ and 216 days for QRBFc. Clinical studies have shown that the older the age, the less the IBIs must be long: in an individual with normal cerebral development, the IBIs should disappear. Therefore the QRBFc best defines the population at risk: its mean age is slightly higher than the mean age of the total population, when QYJ provides a younger population for which it is less unusual to have longer IBIs. We have computed too the percentages of EEG for each set in each

Table 4 Percentages of EEG in categories "artefact", "normal", "doubtful" or "pathological" for each set; the $Total$ row is the initial set of 416 EEG

EEG set	Artefact	Normal	Doubtful	Pathological
Total	1.9231	65.8654	20.1923	12.0192
QYJ				
S_A	0	33.3333	29.3333	37.3333
S_{IN}	6.3291	84.8101	5.0633	3.7975
S_U	1.1450	69.4656	22.1374	7.2519
QRBF				
S_A	0	25.6410	20.5128	53.8462
S_{IN}	10.4167	75.0000	6.2500	8.3333
S_U	0.9119	69.3009	22.1884	7.5988

category normal, doubtful and pathological (results in Table 4). The pathological individuals represent 12 % of the total population, 37 % in the S_A set defined by QYJ and 54 % in the S_A set defined by QRBFc. This confirms the improved ability of the QRBFc model to build the population at risk.

6 Conclusion and Perspectives

Our approach to build reference curves is based on the nonparametric linear quantile regression from Yu and Jones. Its advantages are its robustness to outliers and measurement errors. Moreover, the fact that it is non-parametric allows us to construct the reference curve without a priori assumption. We improved ease of use and performance by reshaping the QYJ estimator with a RBFN network. Indeed, the network is defined for all values of x in the input space: whatever the input x presented by the user, the network produces a prediction without going through all the steps of QYJ algorithm. Somehow, we constructed a parametric model based on a non-parametric approach.

We have shown with simulations that the median (and to a lesser extent the spread) of the mean absolute deviation error obtained for 100 replications with the QRBF estimator is less than that obtained with the QYJ and QRNN estimators for different values of the quantile level τ. We have also find in several simulations that the QRNN estimator is not stable specially when n is small. We introduced in paragraph II-B3 a constraint on the "estimator's mass", i.e. its integral value over the input space, that can act like a smoothing or regularization parameter if the estimator is a positive function. This mass constraint becomes a linear constraint on the weights that is easily integrated in the algorithm of the RBFN construction, with very low additional computational cost. This constraint significantly reduces the number of hidden layer's neurons for a given range of the values of the constraint c: the computation time for plug-in a RBFN neural network in the classical quantile regression approach becomes marginal when the efficiency in term of error gains ground. Further work will be carried out to improve the qualities of the QRBFc estimator by automating the settings of c.

Concerning the IBI curves reflecting the brain maturation, further studies are in progress in order to establish the relationships between these values and the neurological outcome of the infants. We ought to propose reference curves that could help the medical physician in the assessment of the EEG and the neurological status of the preterm infants.

Acknowledgments The authors wish to express their appreciation to the three referees for their constructive and helpful comments and suggestions. The writing of this manuscript was supported by the Agence Nationale de la Recherche grant "BB EEG Platform".

References

1. Biagioni E, Frisone MF, Laroche S, Kapetanakis BA, Ricci D, Adeyi-Obe M et al (2007) Maturation of cerebral electrical activity and development of cortical folding in young very preterm infants Clin Neurophysiol. J PubMed 118(1):53–59 M
2. Cannon AJ (2011) Quantile regression neural networks: implementation in R and application to precipitation downscaling. J Comput Geosci 37(4):1277–1284
3. Cai Z, Wang X (2008) Nonparametric methods for estimating conditional VaR and expected shortfall. J Econom 174:120–130
4. Chen S, Cowan CFN, Grant PM (1991) Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans Neural Netw 2(2):302–309
5. Cole TJ (1988) Fitting smoothed centile curves to reference data. J R Stat Soc Ser A 151(3):385–418
6. Daya B, Ismail A (2006) A neural control system of a two joints robot for visual conferencing. Neural Process Lett 23:289–303
7. Daya B (1999) A multilayer perceptrons model for the stability of a bipedal robot. Neural Process Lett 9:221–227
8. Fan J, Hu T-C, Troung YK (1994) Robust Non-parametric function estimation. Scand J Stat 21:433–446
9. Fan J, Yao Q, Tong H (1996) Estimation of conditional densities and sensitivity mesures in nonlinear dynamic systems. j biomet 83:189–206
10. Ruppert D, Seather ST, Wand MP (1995) An effective bandwidth selector for local least squares regression. J Am Stat Assoc 90:1257–1270
11. Taylor JW (2000) A quantile regression neural network approach to estimating the conditional density of multiperiod returns. J Forecast 19(4):299–311
12. Mateo J, Torres AM, Garca MA (2014) Dynamic fuzzy neural network based learning algorithms for ocular artefact reduction in EEG recordings. Neural Process Lett 39:45–67
13. Wang H (2012) Harmonic mean of Kullback–Leibler divergences for optimizing multi-class EEG spatio-temporal filters. Neural Process Lett 36:161–171
14. Watanabe K, Hayakawa F, Okumura A (1999) eonatal EEG: a powerful tool in the assessment of brain damage in preterm infants. J Brain Dev 21(6):361–372
15. Yu K (1997) Smooth regression quantile estimation. unpublished Ph.D. thesis, The open university
16. Yu K, Jones MC (1998) Local Linear quantile rregression. J Am Stat Assoc 93:228–237
17. Koenker R (2013) Quantreg R package, University of Illinois. http://cran.r-project.org/web/packages/quantreg/index.html
18. Cannon AJ (2011) QRNN R package, University of British Columbia. http://cran.r-project.org/web/packages/qrnn/

Transfer of semiparametric single index model in binary classification

Muhammad-Anas Knefati [1] and Farid Beninel [2]

Abstract The semiparametric classification based on single index model is used in several domains of real life data engineering due to its flexibility. However, it has the same drawback as parametric classification : It is not suitable for the case where the training sample is derived from a certain subpopulation and the prediction sample from another one. The aim of this paper is to use the idea of transfer learning to reduce this drawback. Numerical experiments are performed and are intended to show the improvements from the prediction of point of view.

Keywords Binary supervised classification ; Transfer learning ; parametric classification ; semiparametric classification ; Single index model ; Credit scoring ; Morphometry.

1 Introduction

Classification is a an important statistical field in many experimental sciences and real life applications. It aims to build predictive models to separate and classify data points in two or more groups. Here we are interested in adapting or updating binary classification rules for some particular structure of data *i.e.,* the learning sample and the prediction sample arise from two different subpopulations.

Classification methods are based on an estimate of $\mathbb{E}(Y|X = \mathbf{x})$ or more generally $g(\mathbb{E}(Y|X = \mathbf{x}))$ where g is is a link function. These methods are classified into parametric, nonparametric and semiparametric methods.

Parametric methods assume that the function $\mathbb{E}(Y|X = \mathbf{x})$ is known up to a set of constant parameters that can be estimated from data. Several parametric methods have been proposed in this context such as discriminant analysis, logistic regression, see for instance Hastie et al. (2008) for a comprehensive review. Parametric methods have the advantage of being easily interpreted by practitioners, but rarely justified by theoretical or other *a priori* considerations related to the data design.

Nonparametric methods assume that the function of interest is unknown but smooth. No other assumptions about its shape or functional form are postulated. Therefore they will be more flexible when data at hand does not fit strict classical statistical assumptions. In the other hand, these methods have a serious drawbacks. One of them is that the estimation precision decreases rapidly as the dimension of the the covariate vector $\mathbf{X}$ increases (curse of dimensionality). Another serious drawback is that they don't provide predictions of $E(Y|\mathbf{x})$ at points $\mathbf{x}$ that are outside the considered support of the random variable $\mathbf{X}$.

Semiparametric methods are a trade off between the parametric and nonparametric methods. Their assumptions on the form of the function of interest are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing the possibility of

[1]LMA, UMR CNRS 7348, SP2MI Université de Poitiers, 86962 Futuroscope-Chasseneuil, France (e-mail : maknefati@hotmail.com).

[2]LMA, UMR CNRS 7348, SP2MI Université de Poitiers, 86962 Futuroscope-Chasseneuil, France (e-mail : fbeninel@gmail.com).

specification error. Semiparametric methods give greater estimation precision than do nonparametric methods when X is multidimensional. See Li and Racine (2007) for a review on semiparametric methods.

An approach that is very important in this domain is semiparametric single index models (SIM) that summarizes the effects of the feature measurements variable $\mathbf{X} = (X^1, \ldots, X^p)^T$ within a single variable called the index or score for some specialists. In these models the conditional mean function has the form

$$\mathbb{E}(Y|X = \mathbf{x}) = G(\boldsymbol{\beta}^T \mathbf{x}) \ , \tag{1}$$

where $\boldsymbol{\beta}$ is $p-$dimensional vectors of real parameters and $G(\mathbb{R} \to \mathbb{R})$ a real function. These models mean that all the relevant information carried by X is contained in a linear combination of the components of X. Having the estimates of $\boldsymbol{\beta}$ and $G(.)$, we can readily obtain the estimate of conditional mean from equation (1).

In this work, we deal with the binary classification *i.e.,* Y is a binary group label variable. Let $\mathcal{U}$ is the population of interest. The aim is to predict the group label value of a new individual, for which only the feature measurements $\mathbf{x} = (x^1, \ldots, x^p)^T$ are known. We use the model

$$Y = f(x) + \epsilon, \qquad \epsilon \perp\!\!\!\perp X \tag{2}$$

where $f(\mathbf{x}) = \mathbb{E}(Y|X = \mathbf{x})$ is estimated, using the training sample

$$\mathcal{S}_T \equiv (Y, \mathbf{X})(\mathcal{S}_L) = \{(Y_1, \mathbf{X}_1), ..., (Y_n, \mathbf{X}_n)\}.$$

This problem is known as *supervised classification*. Several examples of such a problem are available such as in credit scoring, where we predict borrowers's behavior to pay pack loan by using information related to these customers. Another example in medicine, where we predict the risk of lung cancer recurrence for a patient previously treated, on the basis of used treatment for the first occurence of the cancer and on some clinical and demographic measurements.

A main problem in supervised learning is that we assume that any individual to predict is supposed to be derived from the same statistical population as the training one. Unfortunately, such assumptions are not realistic. For example, in credit scoring, to predict non customers behavior we use a training sample of costumers only. Also in medicine, the risk of lung cancer recurrence is learned from European patients and will be applied to Asian patients.

In order to avoid space limitations due to the available training sample, we use the transfer learning methodology which aims to transfer the knowledge from a source subpopulation to a target subpopulation.

This idea has been first proposed by Biernacki et al. (2002) in the gaussian context, where they consider the case of two subpopulations slightly different. They establish that both subpopulations are linked through stochastic linear relationship. Estimation of the allocation rule (to be applied on the non-labeled sample) is obtained by estimating parameters of this linear relationship, using constraints models on this relation. They proved that this method is efficient and exhibits better performances than classical classification. Beninel and Biernacki (2009) extended this approach to the multinomial logistic discrimination and proposed several additional links model in the case where the two studied subpopulations are gaussian ones. Beninel et al. (2012) went in deep in the previous results with more tests and simulations in the context of credit scoring and they added to another link's model the former ones.

The semiparametric SIM in classification has a potential superiority over the classical classification methods. We develop here the idea of transfer learning to be applied in SIM as in the work of Beninel et al. (2012).

This work is organized as follows : Section (2) is devoted to the construction of the semi parametric single index model(SIM). The methodology of learning transfer and links models between source and target subpopulations are discussed in Section (3). The performance of the proposed method is assessed by means of a numerical experiments on two real examples in credit scoring and biology in Section (4).

2 Semiparametric single index model(SIM)

A semiparametric SIM has the form

$$Y = G(\boldsymbol{\beta}^T X) + \epsilon, \qquad \boldsymbol{\beta} \in \mathbb{R}^p \tag{3}$$

where Y is the dependent variable, ϵ is the error such that $\mathbb{E}(\epsilon|\mathbf{X}) = 0$. The term $\boldsymbol{\beta}^T\mathbf{X}$ is called "single index" or a scoring.

For identification purpose on $\boldsymbol{\beta}$ and G, We suppose that X must include at least one continuously distributed component whose $\boldsymbol{\beta}$ coefficient is non-zero. Also, we suppose that X contain no constant (intercept) component. Finally we set the $\boldsymbol{\beta}$ coefficient of one component of X equal to one. This problem of identification has been tackled by several authors. To name just few, Manski (1988) studied the identification of single index models for the case of binary response models. Ichimura (1993) investigated the general case in which the response variable can be continue and he described a nonlinear least squares estimator for estimating $\boldsymbol{\beta}$. Klein and Spady (1993) investigated the case of binary response models where he described a semiparametric maximum likelihood estimator estimating $\boldsymbol{\beta}$. Delecroix et al. (2003) generalized the idea of Klein and Spady to arbitrary distributions for Y. Delecroix et al. (2006) analyzed a large class of semiparametric M-estimators for single-index models, including semiparametric quasi-likelihood and semiparametric maximum likelihood estimators.

In binary response models (Y=0 or 1), one can model the relationship between Y and X as $Y = \mathbb{1}_{Z^*>0}$ where $Z^* = \boldsymbol{\beta}^T X + u$ is called the latent variable and u is an error term independent of X with an unknown distribution function F. Then, we have

$$\begin{aligned} \mathbb{E}(Y|X) &= P(Z^* > 0|X) \qquad (4)\\ &= P(u \geq -\boldsymbol{\beta}^T X)\\ &= 1 - F(-\boldsymbol{\beta}^T X)\\ &= G(\boldsymbol{\beta}^T X) \qquad (5) \end{aligned}$$

In general $G(.) \neq F(.)$, but if u has a symmetric distribution, then, in this case, $G(.) = F(.)$ as in a Probit model where $u \sim \mathcal{N}(0, 1)$.

Now, we will review the method of Ichimura (1993) and Klein and Spady (1993) to estimate $\boldsymbol{\beta}$ and G. These two methods use M-estimation method as follows

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \psi(Y_i, \hat{G}(\boldsymbol{\beta}^T X_i; \boldsymbol{\beta}))\tau_n(X_i) \tag{6}$$

where

– $\hat{G}(t; \boldsymbol{\beta})$ is a nonparametric estimator of the regression function $E(Y|\boldsymbol{\beta}^T X = t)$. For example, the Nadaraya-Watson estimator can be used. i.e.,

$$\hat{G}(t; \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Y_i K(\frac{t-\boldsymbol{\beta}^T X_i}{h})}{\sum_{j=1}^{n} K(\frac{t-\boldsymbol{\beta}^T X_j}{h})} \tag{7}$$

where $K(.)$ is a symmetic kernel that can be, in general, with compact support as epanechnikov kernel $K(t) = \frac{3}{4}(1 - t^2)\mathbb{1}_{|t|\leq 1}$, and h is the smoothing parameter. In order to avoid degenerate problems the version "leave-one-out" is considered to estimate G

$$\hat{G}^{(-i)}(\boldsymbol{\beta}^T X_i; \boldsymbol{\beta}) = \sum_{k\neq i} \frac{Y_k K(\frac{\boldsymbol{\beta}^T X_i - \boldsymbol{\beta}^T X_k}{h})}{\sum_{j\neq i} K(\frac{\boldsymbol{\beta}^T X_i - \boldsymbol{\beta}^T X_j}{h})} \tag{8}$$

– ψ is the contrast function. In Ichimura (1993) $\psi(y,r) = (y-r)^2$, and in Klein and Spady (1993) $\psi(y,r) = -y\log(r) - (1-y)\log(1-r)$.
– $\tau_n(.)$ is the trimming function.

The estimator of $\boldsymbol{\beta}$ is very sensible by the choice of h. (Ichimura (1993) proved that, we can estimate, simultaneously, $\boldsymbol{\beta}$ and h form (6)).

The decision rule for a new individual as $\mathbf{x}^*$ is given by :

$$\hat{y} = \mathbb{1}_{\hat{G}(\hat{\boldsymbol{\beta}}^T x^*)}$$

3 Learning transfer model

3.1 Methodology

We assume that the data consist of tow samples : the first is $\mathcal{S} = (Y,X) = \{(Y_1,X_1),...,(Y_n,X_n)\}$ with n points and drawn from a source population $\mathcal{U}$, and the second is $\mathcal{S}^* = (Y^*,X^*) = \{(Y_1^*,X_1^*),...,(Y_{n^*}^*,X_{n^*}^*)\}$ with n^* points and drawn from a target population $\mathcal{U}$. The idea of learning Transfer Models is to build a decision rule for the target population using both samples $\mathcal{S}$ and $\mathcal{S}^*$.

From the learning sample $\mathcal{S}$, we use the semoparametric SIM to obtain the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{G}$. Then we affect the sample $\mathcal{S}^*$ as follows :

$$E(Y_j^*|X_j^*) = \hat{G}(L(X_j^*)), \qquad (j = 1,...,n^*) \tag{9}$$

where
– $L(X_j^*) = c + (\Lambda\hat{\boldsymbol{\beta}})^T X_j^*$, $\quad c \in \mathbb{R}$, and $\Lambda = \mathrm{diag}(\lambda_1,...,\lambda_p)$.
– Function $\hat{G}$ is defined by (7).

In order to estimate the $(p+1)$ real parameters of (c,Λ), we use the maximum likelihood method :

$$\begin{aligned}\ell(c,\Lambda) &= \sum_{j=1}^{n^*} Y_j^* \log(P(Y_j^*=1|X_j^*)) + (1-Y_j^*)\log(1-P(Y_j^*|X_j^*)) \\ &= \sum_{j=1}^{n^*} Y_j^* \log(E(Y_j^*|X_j^*)) + (1-Y_j^*)\log(E(Y_j^*|X_j^*))\end{aligned}$$

By maximising ℓ with respect to c and Λ after substituting $E(Y_j^*|X_j^*)$ from (9), we can obtain transfer Model parameters. In what follows, we discuss issues where the pairs (c,Λ) are unkwown and we propose several scenarios for estimating them.

3.2 Links models

The estimation of parameters c and Λ can be done through several models depending on several possible situations for c and Λ that are :
– M_0 : No parameter to be estimated : $c = 0$ and $\Lambda = I_p$ where I_p is the identical matrix in $\mathbb{R}^p$.
– M_1 : Here only c is to be estimated and $\Lambda = I_p$.
– M_2 : c=0 and $\Lambda = \lambda I_d$, where $\lambda \in \mathbb{R}$ and we have to estimate λ.
– M_3 : Two parameters to be estimated : c and $\lambda \in \mathbb{R}$, where $\Lambda = \lambda I_d$,
– M_4 : $c = 0$ and $\Lambda = \{\lambda_1,...,\lambda_p\}$, where $\lambda_1,...,\lambda_p \in \mathbb{R}$ are to be estimated
– M_5 : The most complex model : c is free and $\Lambda = \mathrm{diag}(\lambda_1,...,\lambda_p)$, where $\lambda_j \in \mathbb{R}$ and $j = 1,...,p$.

We add another model M_6 in which we use $S \cup S^*$ as the learning sample to estimate the SIM parameters.

3.3 Algorithm of transfer learning

The algorithm of transfer learning is explained as follow :

1. Calculate
$$L(X_j^*) = c + (\Lambda\hat{\boldsymbol{\beta}})^T X_j^*, \quad j = 1, ..., n^* \tag{10}$$
2. Estimate the set of parameters $c \in \mathbb{R}$ and $\Lambda \in \mathbb{R}^4$ according to the chosen model in table (**??**) by maximizing the empirical Likelihood function in c and Λ :
$$\ell(c, \Lambda) = \sum_{j=1}^{n^*} Y_j^* \log\left(P(Y_j^* = 1|L(X_j^*))\right) + (1 - Y_j^*) \log\left(1 - P(Y_j^* = 1|L(X_j^*))\right) \tag{11}$$
where
$$P(Y_j^* = 1|L(X_j^*)) = \hat{G}(L(X_j^*)) \tag{12}$$
and $\hat{G}$ is given by (7).
3. Replace the estimated parameters in (10) to obtain $L(X^*)$ and then replace $L(X^*)$ in (9) to obtain estimator for the sample S_L^*
4. We predict the sample test $S_T^* = S^* \setminus S_L^*$ using the estimator obtained.
5. To mesure the performance of our method, we calculate the error rate
$$e = \frac{1}{n_T^*} \sum_{j=1}^{n_T^*} \mathbb{1}_{Y_j^* \neq \hat{Y}_j^*}$$
where n_T^* is the length of S_T^* and $\hat{Y}_j^*$ is the prediction of Y_j^*.

4 Numerical experiments

In our numerical experiments, we suppose that the size of $\mathcal{S}$ is large enough ($n \geq 200$) in order to obtain a small variance for $\hat{G}$.

4.1 Biology Data

The data consist of three samples of seebirds that come from three subspecies of Calanectris diomedea species. These samples are :

- Borealis (n=206, 45% female) live in the Atlantic islands
- Diomeda (n=35, 58% female) live in the Mediterranean islands
- Edwardsii(n=92, 52% female) live in the Cape Verde Islands

Five morphological variables were measured to forecast the bird's sex. These variables are :

$\underline{BECH}$ culmen(bill) depth.

$\underline{BECL}$ culmen length.

$\underline{AIL}$ wing length.

$\underline{QUEUE}$ tail length.

All simulations are replicated $R = 50$.

We use the Borealis subspecies as leaning sample($\mathcal{S}$) to calculate $\hat{\boldsymbol{\beta}}$ and $\hat{G}$ which determinate the SIM estimator. Now, we will first take $\mathcal{S}^*$ as the Diomedea subspecies. Length of samples $\mathcal{S}_L^*$ are $10, 11, ..., 20$. The figure (1) illustrate the different results of simulations.

Then, we take $\mathcal{S}^*$ as the Edwardsii subspecies($\mathcal{S}^*$). Length of samples $\mathcal{S}_L^*$ are 10,15,...,70. The figure (2) illustrate the different results of simulations.

4.2 Credit scoring data

To illustrate our proposed methodology, we consider a real data example in credit scoring on private loans from a southern German bank. The data set and the description of the variables are available in the page web `http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kreditvar_e.html` or see also for more description Fahrmeir and Tutz (2010).

This data set consists of 1000 consumer credits. For each consumer the binary response variable "creditability" is available (Kredit=1 for creditworthy and Kredit=0 otherwise). In addition, 20 covariates that are assumed to influence creditability were recorded. Here we are interested in the following six covariates :

$\underline{laufkont}$ *Balance of current account* with the following four categories :

1 : no running account ;

2 : no balance or debit ;

3 : medium running account (less than 200 Deutsche Mark (DM)) ;

4 : large running account (greater or equal to 200 DM or checking account for at least one year)

$\underline{laufzeit}$ **:** *Duration of credit in months (metric)* ;

$\underline{sparkont}$ *Value of savings or stocks* ;

$\underline{moral}$ **:** *Payment of previous credits.* ;

$\underline{beszeit}$ **:** *Duration of employment* with five categories :

1 : unemployed ;

2 : less than one year ;

3 : more than one year and less than four years ;

4 : more than four years and less than seven years ;

5 : more than seven years.

$\underline{weitkred}$ **:** *Further running credits.*

There are 700 observations with $Kredit = 1$ and 300 observations for $Kredit = 0$. For this experiment, we study the borrowers non customers behavior to pay back loans, we use the variable $Laufkont$ to separate the available data set in tow subpopulations : the customers subpopulation $\mathcal{S}$ when $Laufkont > 1$ with 726 observations and the non customers subpopulations $\mathcal{S}^*$ when $Laufkont = 1$ with 274 observations.

We split the non costumers population S^* into two samples : a learning sample S_L^* and a test sample S_T^*. we draw at random 35 learning samples S_L^* of sizes : $n^* = 50, 100, 150, 200$ from the non costumers S^*, each learning sample allows to represent the diverse models and to bring out affectation rules, we take, also, $S_T^* = S^* \setminus S_L^*$ as a test sample that is used to verify the reliability of the established models. The algorithm of transfer learning is repeated $R = 50$ times. The figure (3) illustrate the different results of simulations.

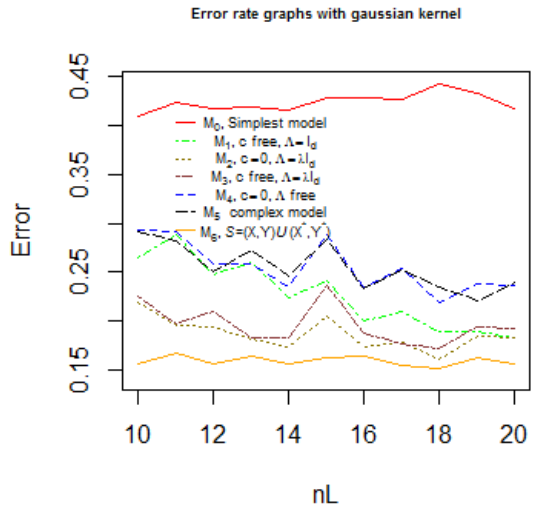


FIGURE 1 – Experiments results : Borealis subspecies are learning sample and Diomedea subspecies are testing sample.

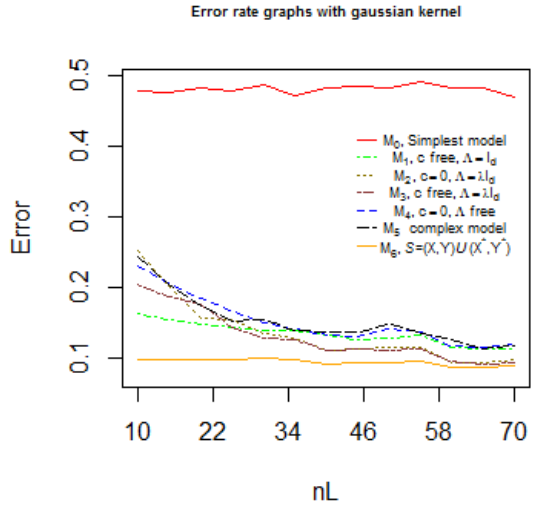


FIGURE 2 – Experiments results : Borealis subspecies are learning sample and Edwardsii subspecies are testing sample.

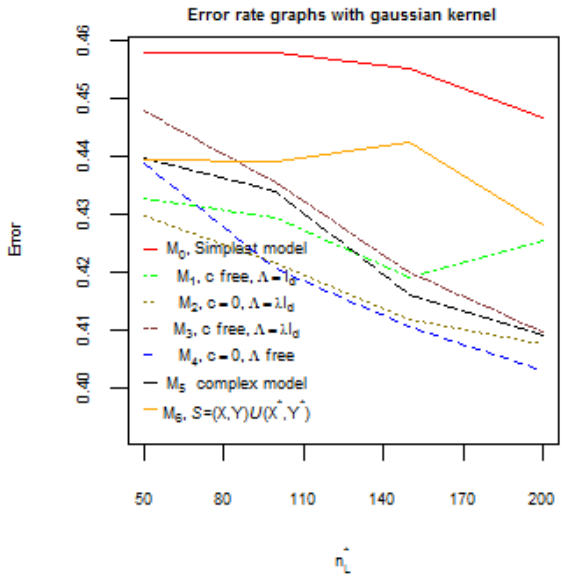


FIGURE 3 – Experiments results : Customers are learning sample and non customers are testing sample.

Références

Biernacki, C., Beninel, F., and Bretagnolle, V. (2002). A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2), 387–397.

Beninel, F. and C. Biernacki (2009). Updating a logistic discriminant rule : Comparing some logistic submodels in credit-scoring. *International Conference on Agents and Artificial Intelligence, Porto, Portugal*, pp. 267–274.

Beninel, F., Bouaguel, W., and Belmufti, G. (2012). Transfer Learning Using Logistic Regression in Credit Scoring. *arXiv preprint arXiv* :1212.6167.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression, *Journal of Multivariate Analysis* 86(2), 213-226.

Delecroix,M, Hristache, M, and Patilea, V.(2006) On semiparametric M-estimation in single-index regression *Journal of Statistical Planning and inference* 136, 730–769.

Dominitz,J and Sherman, R.P.(2005) Some convergence theory for iterative estimation procedures with an application to semiparametric estimation *Econometric Theory*, 21, 838-863.

Fahrmeir, L., and Tutz, G. (2010). Multivariate statistical modelling based on generalized linear models. *(2nd ed.). Springer.*

Hastie, T., Tibshirani, R., Friedman, J. H.(2011). The Elements of Statistical Learning : Data Mining, Inference, and Prediction. *Springer.*

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models, *Journal of Econometrics*, 58, 71–120.