MOHAMED KHIDER UNIVERSITY - BISKRA

Faculty of Exact Sciences and the Natural Sciences and Life

DEPARTMENT OF MATHEMATICS

A Thesis Presented for the Degree of **:**

DOCTOR in Mathematics

In the Filled of : **Statistics**

By

Souraya KHEIREDDINE

Title :

On Boundary Correction in Kernel Estimation

Examination Committee Members :

NECIR Abdelhakim	Prof.	Biskra University	Chairman
YAHIA Djabrane	M.C.A.	Biskra University	Supervisor
YOUSFATE Abderrahmane	Prof.	Sidi Bel-Abbas University	Examinator
SAYAH Abdallah	M.C.A.	Biskra University	Examinator
BRAHIMI Brahim	M.C.A.	Biskra University	Examinator

2016

UNIVERSITÉ MOHAMED KHIDER, BISKRA

Faculté des Sciences Exactes et des Sciences de la Nature et de La Vie

DÉPARTEMENT DE MATHÉMATIQUES

Thèse présentée en vue de l'obtention du Diplôme **:**

Doctorat en Mathématiques

Option: **Statistique**

Par : Souraya KHEIREDDINE

Titre :

Sur la correction des effets de bord dans l'estimation à noyau

Membres du Comité d'Examen :

NECIR Abdelhakim	Professeur	Université de Biskra	Président
YAHIA Djabrane	M.C.A.	Université de Biskra	Encadreur
YOUSFATE Abderrahmane	Professeur	Univ. Sidi Bel-Abbas	Examinateur
SAYAH Abdallah	M.C.A.	Université de Biskra	Examinateur
BRAHIMI Brahim	M.C.A.	Université de Biskra	Examinateurr

2016

To my

Dear Parents

&

my Sisters and Brothers.

Remerciements

Avant tout j'adresse mes remerciements à mon Dieu qui m'a donné la patience et le courage qui m'ont permis de réaliser ma thèse.

Je souhaite exprimer ici ma reconnaissance envers mon directeur de thèse Dr. Yahia Djabrane, pour ses conseils bénéfiques et ses apports précieux tout au long de la réalisation de ma thèse. Merci de votre disponibilité, de votre grande patience et de vos conseils fort judicieux...

Je remercie infiniment le Professeur Necir Abdelhakim de l'Université de Biskra de me faire l'honneur de présider le Jury de ma thèse. Je tiens également à exprimer ma gratitude envers le Professeur Yousfate Abderrahmane d'avoir accepté de participer au Jury de cette thèse. Je le remercie énormément. Dr. Sayah Abdallah et également Dr. Brahimi Brahim Maitres de Conférences à l'Université de Biskra ont bien acceptés de participer au Jury de cette thèse, qu'ils trouvent ici toute ma gratitude.

Je tiens à remercier ma chère mère, et toute ma famille que Dieu les gardent, pour leurs sacrifices, leur patience et leur encouragements. Sans eux, je ne me serais jamais rendu aussi loin.

Mes remerciements sont adressés également à tous les membres de Département de Mathématiques et ceux du Laboratoire de Mathématiques Appliquées. Sans oublier mes amis, dont je garderai de très bons souvenirs.

ملخـــــص

في هذه الرسالة نقوم بدراسة بعض طرق تصحيح الآثار الحدية لمقدرات النواة لدوال الكثافة والانحدار وخصائصهم الإحصائية. تظهر مقدرات النواة مشاكل في التقارب على مستوى حدود حاملها. بعبارة أخرى، فإن هذه الآثار الحدية تؤثر تأثيرا خطيرا على أداء هذه المقدرات .لتصحيح هذه الآثار، تم اقتراح وطرح مجموعة متنوعة من الأساليب ، الأكثر استخداما وعلى نطاق واسع هي الانعكاس ، التحويل والخطية المحلية... في هذه المذكرة قمنا بمزج طريقتي الانعكاس والتحويل من أجل تقديم طريقة جديدة وشاملة لتصحيح الآثار الحدية لدالة الانحدار. مشكلة الآثار الحدية لمقدرات النواة لدوال الكثافة والربيعيات في حالة التوزيعات ذات الذيل الثقيل تمت دراستها أيضا.

Abstract

In this thesis we study some boundary correction methods for kernel estimators of both density and regression functions and their statistical properties. Kernel estimators are not consistent near the finite end points of their supports. In other words, these effects seriously affect the performance of these estimators. To remove the boundary effects, a variety of methods have been developed in the literature, the most widely used is the reflection, the transformation and the local linear methods... In this thesis, we combine the transformation and the reflection methods in order to introduce a new general method of boundary correction when estimating the regression function. Boundary problem for kernel quantile function estimator in heavy-tailed case are also studied in this thesis.

Résumé

Dans cette thèse nous étudions certaines méthodes de correction des effets de bord des estimateurs à noyaux des fonctions de densité et de la régression et leurs propriétés statistiques. Les estimateurs à noyau présentent des problèmes de convergence aux bords de leurs supports. En d'autre termes, ces effets de bord affectent sérieusement les performances de ces estimateurs. Pour corrigé ces effets de bord, une variété de méthodes ont été développées dans la littérature, la plus largement utilisée est la réflexion, la transformation et la linéaire locale... Dans cette thèse, nous combinons les méthodes de transformation et de réflexion, pour introduire une nouvelle méthode générale de correction de l'effet de bord lors de l'estimation de la régression. Le problème de bord de l'estimateur à noyau de la fonction des quantiles en cas de distribution à queue lourde est également étudié dans cette thèse.

Contents

List of Figures

List of Tables

Introduction

In statistical studies, it is often the case that variables represent some sort of physical measure such as time or length. These variables thus have a natural lower boundary, e.g. time of birth or zero point on a scale. Hence, it is also justified to assume that the underlying true density f has a bounded support. There are many applications in particular in economics where densities of positive random variables are the object of interest or an essential model to be estimated from data. For examples, volatility models, duration and survival times data, financial transaction data,... In a lot of these situations, however, appropriate functional forms are unknown, such that a nonparametric estimate is needed. It is often the point estimates close to the boundary which are the focus of practical interest and thus, require good precision.

Nonparametric kernel smoothing belongs to a general category of techniques for nonparametric curve estimations including : density, distribution, regression, quantiles, ... These estimators are now popular and in wide use with great success in statistical applications. Early results on kernel density estimation are due to Rosenblatt (1956) *and Parzen* (1962)*, and the form kernel regression estimator*

has been proposed by Nadaraya (1964) *and Watson* (1964). *Since then, much research has been done in the area e.g., the monographs of Silverman* (1986), *and Wand and Jones* (1995) *and kernel regression estimator can be found in, for instance, Gasser and Müller* (1979), *Eubank* (1988) *and Fan and Gijbels* (1996).

Kernel estimators are not consistent near the finite end points of their supports. In other words, these effects seriously affect the performance of these estimators and these require good precision. In this thesis we study some boundary correction methods for kernel estimators of both density and regression functions and their statistical properties. The so-called 'boundary problem' of kernel density estimators has been thoroughly analyzed and discussed for densities which are continuous on their support $[0, \infty)$. *It arises when the support has at least one finite boundary and it appears e.g. in form of a relatively high bias when calculating the estimate at a point near the boundary. In the density estimation context, a variety of boundary correction methods now exists, and most are referred to in Jones* (1993). *He sets up a unified approach to many of the more straightforward methods using "generalized jackknifing" (Schucany et al.* 1971*). A well-known method of Rice* (1984) *is a special case. A popular linear correction method is another: it has close connections with the boundary properties of local linear fitting (Fan and Gijbels,* 1996*)... Consequently, an idea on how to include boundary corrections in these estimators is presented.*

In the regression function estimation context, Gasser and Müller (1979) *identified the unsatisfactory behavior of the Nadaraya Watson regression estimator for*

points in the boundary region. They proposed optimal boundary kernels but did not give any formulas. However, Gasser and Müller (1979) *and Müller* (1988) *suggested multiplying the truncated kernel at the boundary zone or region by a linear function. The local linear methods developed recently have become increasingly popular in this context (cf. Fan and Gijbels* (1996)*). More recently, in Dai and Sperlich* (2010) *a simple and effective boundary correction for kernel density and regression estimator is proposed, by applying local bandwidth variation at the boundaries. To remove the boundary effects a variety of methods have been developed in the literature, the most widely used is the reflection method, the boundary kernel method, the transformation method, the pseudo-data method and the local linear method. They all have their advantages and disadvantages. One of the drawbacks is that some of them (especially boundary kernels), can produce negative estimators.*

For heavy-tailed distributions, bias or inefficiency problems may occur in the classical kernel estimation when considering. The estimation of population quantiles is of great interest when a parametric form for the underlying distribution is not available. It plays an important role in both statistical and probabilistic applications, namely: the goodness-of-fit, the computation of extreme quantiles and Value-at-Risk in insurance business and financial risk management. Also, a large class of actuarial risk measures can be defined as functionals of quantiles (see, Denuit et al., 2005*). Quantile estimation has been intensively used in many fields, see Azzalini (*1981*), Harrell and Davis (*1982*), Sheather and Marron (*1990*), Ralescu and Sun (*1993*), Chen and Tang (*2005*). Most of the existing estimators*

suffer from either a bias or an inefficiency for high probability levels. Inspired by Wand et al. (1991) , *Buch-Larsen et al. (*2005*) showed that for heavy-tailed distributions, the tail performance of the classical kernel density estimator could be significantly improved by using a tail flattening transformation. They used modified Champernowne distribution to estimate loss distributions in insurance which is categorically heavy-tailed distributions. Sayah et.al.*(2010) *produce a kernel quantile estimator for heavy-tailed distributions using a modification of the Champernowne distribution.*

The rest of the thesis is organized as follows. In chapter 1, we focused on the boundary effect in kernel density estimation, some methods of boundary correction have been discussed. This first chapter consists of preliminary mathematical material which serves the framework for the rest of the thesis. Chapter 2 is concerned with the connections between the kernel regression estimation and boundary effect. Chapter 3 introduces the important part of our research is devoted to the extension of the boundary correction methods based on both transformation and reflection to the regression setting. In chapter 4, We have focused also on the boundary problems for kernel quantile estimator in heavy-tailed data case and presents some asymptotic results.

Chapter 1

Boundary correction in kernel density estimation

In the past, many ways to diminish the boundary problem in the kernel density estimation have been considered. Consequently, an idea on how to include boundary corrections in these estimators is presented. The first statement implies that the density has a support which is bounded on the left hand side. Without loss of generality the support is set to be $[0, \infty)$. Nonparametric kernel density estimation is now popular and in wide use with great success in statistical applications. The reflection method is specifically designed for the case $f^{(1)}(0) = 0$ where $f^{(1)}$ denotes the first derivative of f. The boundary kernel method is more general than the reflection method in the sense that it can adapt to any shape of density.

These included a boundary kernel and its close counterpart the local linear fitting method, the transformation and reflection based method given by Zhang et al.

(1999), Jones and Foster's (1993) nonnegative adaptation estimator, Cowling and Hall's (1996) pseudo-data method, and a recent estimator due to Hall and Park (2002) based on a transformation of the data "inside" the kernel.

1.1 Kernel density estimation

Let $X_{1,}...,X_n$ be independent and identically distributed (iid.) copies of the random variable (rv) X with continuous distribution function

$$F(x) = P[X \leq x]$$

and continuous density function : $f(x) = \frac{d}{dx}F(x)$.

In this chapter, we will consider the problem of estimating the density using non-parametric kernel estimation, wich is a rather simple but very powerful and thus broadly used method to estimate a density non-parametrically. It was first defined in Rosenblatt (1956) and Parzen (1962), the latter providing a more detailed analysis of this new and innovative method.

A very natural estimator of the distribution function is the empirical estimator

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}},$$

where

$$1_{\{X_i \leq x\}} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x \end{cases}$$

Definition 1.1.1 *(Standard kernel density estimator). Let n be the sample size and K be a kernel function of support $[-1,1]$, symmetric around the origin. The standard kernel density estimator based on $X_1, ..., X_n$ is*

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1.1}$$

where $h := h_n$ ($h \to 0$ and $nh \to \infty$) is the bandwidth and $K_h(.) := K(./h)$, where K is an integrable smoothing kernel which usually is nonnegative, i.e., a symmetric probability density function .

Conditions 1.1 *f has two derivatives and f'' is bounded and uniformly continuous in a neighborhood of zero or x when x is a boundary or interior point, respectively.*

K satisfies $\int K^2(t)\,dt < \infty$ and $\int |t^2 K(t)|\,dt < \infty$.

Propriety 1.1.1 *For any real-valued function χ on $\mathbb{R}$, $c \in \mathbb{R}$ and $l = 0, 1, 2$, define $\mu_{l,c}(\chi) = \int_{-\infty}^{c} t^l \chi(t)\,dt$ and $\mu_l(\chi) = \int_{-\infty}^{\infty} t^l \chi(t)\,dt$. Suppose that Condition 1.1 holds, then for $x = ch$, $c \geq 0$, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, $\hat{f}_h(x)$ has expected value*

$$E\left(\hat{f}_h(x)\right) = \mu_{0,c}(K) f(x) - h\mu_{1,c}(K) f^{(1)}(x) + \frac{h^2}{2}\mu_{2,c}(K) f^{(2)}(x) + o\left(h^2\right) \tag{1.2}$$

and variance

$$Var\left(\hat{f}_h(x)\right) = \frac{1}{nh} f(x)\,\mu_{0,c}\left(K^2\right) + o\left(\frac{1}{nh}\right). \tag{1.3}$$

Propriety 1.1.2 *Suppose that K is supported on $[-1,1]$. Then for any $c \in [0,1)$, $\mu_{0,c}(K) < 1$ and $\hat{f}_h(x)$, $x = ch$, as an estimator of $f(x)$, has a nonzero constant bias unless $f(0+)$ is equal to zero. And for $c \geq 1$, (1.2) and (1.3) become*

$$E\left(\hat{f}_h(x)\right) = f(x) + \frac{h^2}{2}\mu_2(K)\, f^{(2)}(x) + o\left(h^2\right) \tag{1.4}$$

and

$$Var\left(\hat{f}_h(x)\right) = \frac{1}{nh} f(x)\,\mu_0\left(K^2\right) + o\left(\frac{1}{nh}\right). \tag{1.5}$$

The mean square error (MSE) is a widely used measure of discrepancy. For $\hat{f}_h$ as an estimator of f it is defined as

$$MSE\left(\hat{f}_h(x)\right) = E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] = Bias^2\left(\hat{f}_h(x)\right) + Var\left(\hat{f}_h(x)\right).$$

The asymptotic mean integrated square error ($AMISE$) is

$$AMISE\left(\hat{f}_h(x)\right) = \frac{h^4}{4}\mu_2^2(K)\int f^{(2)}(x)^2\,dx + \frac{1}{nh}\mu_0\left(K^2\right). \tag{1.6}$$

The bandwidth which minimizes the AMISE can be calculated by differentiating (1.6), setting the equation to 0 and solving it for h. The result is referred to as

the optimal global bandwidth:

$$h_{opt} = \left(\frac{\mu_0 \left(K^2 \right)}{\mu_2^2 \left(K \right) \int f^{(2)} \left(x \right)^2 dx} \right)^{1/5} n^{-1/5}.$$

Remark 1.1.1 *The formula for the bias and the variance show that some sort of bias-variance trade-off is present. Taking assumption* $h \to 0$ *and* $nh \to \infty$ *for* $n \to \infty$ *into account, the following behavior can be observed:*
1) h becomes too small: bias gets smaller, variance gets larger.
2) h becomes too large: bias gets larger, variance gets smaller.
An example of the impact of the bandwidth on the estimator can be seen in the figure (1.1).

From (1.6) another useful result can be derived: the optimal kernel function. Since the moments have a defining impact on the AMISE and the function itself has restrictions from its own definition, an optimal kernel function can be derived. Some popular kernels functions used in the literature are the following (see Silverman, 1986):

Quartic or Biweight kernel	$K_{biw}(t) = \frac{15}{16}\left(1 - t^2\right)^2 1_{\lvert t \rvert \leq 1}$
Triangular kernel	$K_{trian}(t) = \frac{35}{32}\left(1 - t^2\right)^3 1_{\lvert t \rvert \leq 1}$
Gaussian kernel	$K_{gauss}(t) = \frac{1}{\sqrt{2\pi}} e^{-1/2t^2}, \qquad for \;\; t \in \mathbb{R}$
Epanechnikov kernel	$K_{Epa}(t) = \frac{3}{4}\left(1 - t^2\right) 1_{\lvert t \rvert \leq 1}$

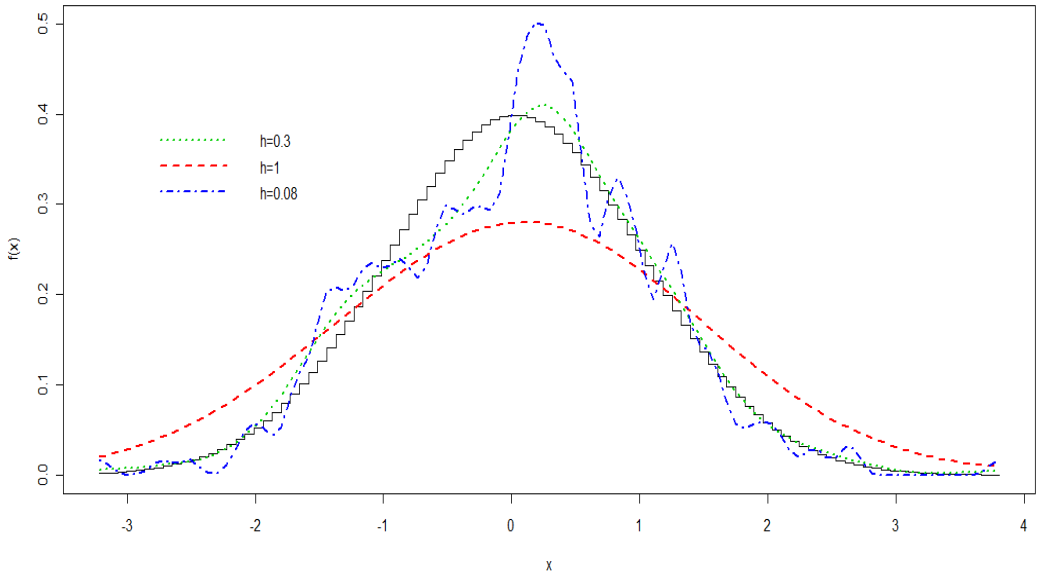

Figure 1.1: Kernel density estimator using three different bandwidths

1.2 Boundary effects

In statistical studies, it is often the case that variables represent some sort of physical measure such as time or length. These variables thus have a natural lower boundary, e.g. time of birth or zero point on a scale. Hence, it is also justified to assume that the underlying true density f has a bounded support. Boundary effects are a well known problem in nonparametric curve estimation, no matter if we think of density estimation or regression. Moreover, both density estimator and regression usually show a sharp increase in variance and bias when estimating them at points near the boundary region, i.e., for $x \in [0, h)$, this phenomenon is referred to the "*boundary effects*".

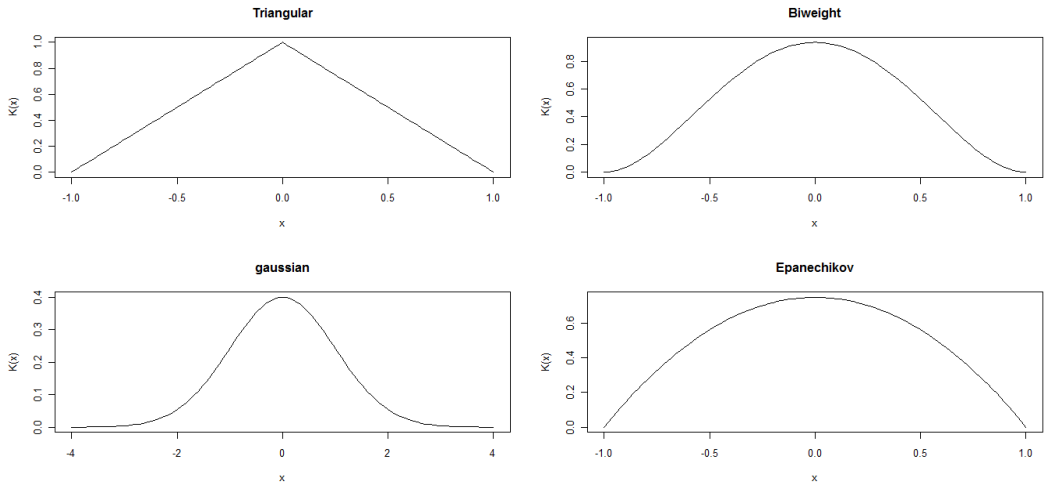

Figure 1.2: Rate of kernels: Triangular, Biweight, Gaussian and Epanechnikov

To remove those boundary effects in kernel density estimation, a variety of methods have been developed in literature. Some well-known methods are summarized below:

- Reflection method (Cline and Hart, 1991, Schuster, 1985, Silverman, 1986).
- Boundary kernel method (Gasser and Müller, 1979, Gasser et al., 1985, Jones 1993, Müller, 1991, Zhang and Karunamuni, 2000).
- Transformation method (Marron and Ruppert, 1994, Wand et al., 1991).
- Pseudo-data method (Cowling and Hall, 1996).
- Local linear method (Cheng, 1997, Zhang and Karunamuni, 1998).
- Rice's Boundary Modification (Cheng, 2006).

Consider a density function which is continuous on $[0,\infty)$ and is 0 for $x<0$. Given a bandwidth h, the interval $[0,h)$ is defined to be the *boundary interval* and $[h,\infty)$ the *interior interval*. The kernel density estimator is in conformity in the interior interval. As will be shown, problems will arise if x is smaller than the chosen bandwidth h. In order to analyze this situation, consider now only $\hat{f}_h(c.h)$, for $c\in[0,1)$. This can be understood as some sort of rescaling. The expected value of $\hat{f}_h(x)$ is computed just as before, but when substituting the variables, one must pay attention to the limits of the integral:

$$\mathbb{E}\left(\hat{f}_h(x)\right)=\int_{-1}^{c}K(t)\,f(x-ht)\,dt,\qquad x=ch\ \text{ for }\ c\in(0,1).$$

Assuming that f'' exists and is continuous in a neighborhood of x, the density in the integral can be approximated by its second order Taylor expansion evaluated at x:

$$f(x-ht)=f(x)+(x-ht-x)\,f^{(1)}(x)+\frac{1}{2}(x-ht-x)^2\,f^{(2)}(x)+o\left(h^2\right).$$

given for $h\to 0$ and being uniform in $t\in[-1,1]$,

$$\begin{aligned}\mathbb{E}\left(\hat{f}_h(x)\right)&=f(x)\int_{-1}^{c}K(t)\,dt-hf^{(1)}(x)\int_{-1}^{c}tK(t)\,dt\\&\quad+\frac{h^2}{2}f^{(2)}(x)\int_{-1}^{c}t^2K(t)\,dt+o\left(h^2\right).\end{aligned}\tag{1.7}$$

Unless $x\geq h$, i.e. $c\geq 1$, the estimator is not asymptotically unbiased and inconsistent. At the left most boundary the expected value asymptotically reaches

only half the original value:

$$\mathbb{E}\left(\hat{f}_h(0)\right) = \frac{1}{2} f(0) + O(h).$$

Example 1.2.1 *The boundary problem can be detected in figure (1.3). The theoretical curve is that of the exponential density.*

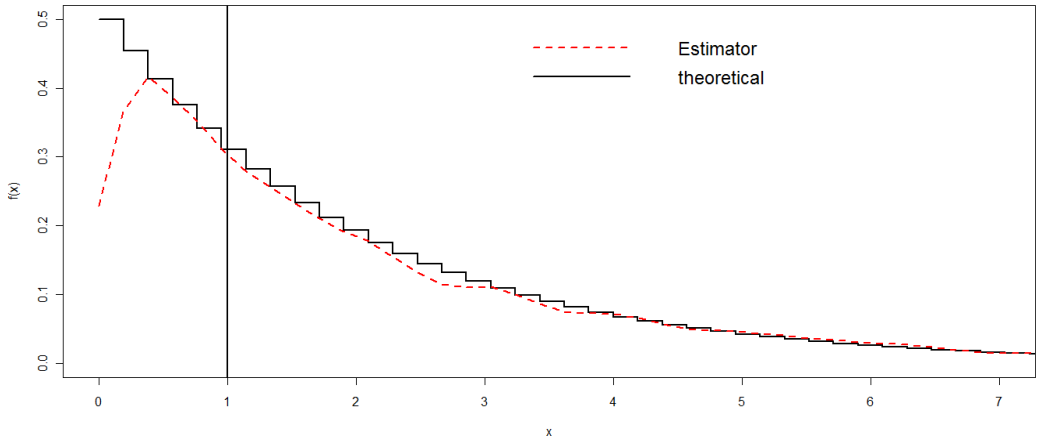

Figure 1.3: Boundary problem in kernel density estimation

1.3 Boundary corrections in kernel density estimation

As will be shown, it is difficult to find an approach which fulfills both requirements without bringing along other restrictions. Thus, one must usually set some

kind of priority: is it more important to find an estimate which is rather precise but is not a real density or to find an actual density setting the exactness of the estimate as a second priority?. Since the standard kernel density estimator performs satisfyingly for $x \geq h$, the goal is to find a method which adapts near the boundary in a beneficial way, but coincides with the standard estimator in the interior interval, i.e. if $x \geq h$. It is natural to desire a smooth crossover from the boundary to the internal estimator. This is justified by the simplicity it brings along: one would have to select a kernel function, bandwidth and possibly tuning factors for the boundary but would not require two or more different algorithms for the estimation on the whole support. In this section, some methods were selected which seemed to be reasonable. There were methods which were rather complicated and others which on the other hand felt quite natural. The order in which these approaches are presented is not chronological but is rather an attempt to create a coherent order.

If not explicitly stated otherwise, $K(.)$ is taken to be a smooth kernel function of support $[-1, 1]$, symmetric with respect to the origin, the sample consists of n i.i.d. copies of the random variable X with continuous density f on $[0, \infty)$, the bandwidth is a strictly positive number $h > 0$, depending on n, fulfilling the conditions ($h \to 0$, $n \to \infty$ and $nh \to \infty$) and $\hat{f}_h(.)$ is the standard kernel density estimator as in Definition (1.1.1).

1.3.1 Cut-and-Normalized method

As can be seen in (1.7), the kernel density estimator is asymptotically biased in $[0, h)$,

$$E\left(\hat{f}_h(x)\right) = f(x)\int_{-1}^{c} K(t)\,dt - hf^{(1)}(x)\int_{-1}^{c} tK(t)\,dt + \frac{h^2}{2}f^{(2)}(x)\int_{-1}^{c} t^2K(t)\,dt + o\left(h^2\right)$$

Due to Gasser and Müller (1979), a very naive correction could then be to divide the original estimator (1.1) by this factor $\int_{-1}^{c} K(t)dt$. The order of the bias is then h, which still is not very satisfying since in $[h, \infty)$ it becomes of order h^2. The goal is to achieve such an order in the boundary interval. This is a local correction since the integral depends on the relative position of x with respect to the bandwidth h

$$\hat{f}_{CN}(x) = \frac{1}{nh}\frac{1}{\int_{-1}^{c} K(t)\,dt}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \qquad x \geq 0. \tag{1.8}$$

1.3.2 Reflection of the data method

This method is introduced by Schuster (1985), then study by Cline and Hart (1991). A simple but ingenious idea is to reflect the data points $X_{1,}..., X_n$ at the

origin and then to work with the rv's:

$$Y_i = \begin{cases} -X_j, & j = 1, ..., n \\ X_{2n-j}, & j = n+1, ..., 2n \end{cases}$$

This not only yields a twice as large sample size but most importantly yields a 'sample' drawn from a density with unbounded support. Therefore, a standard kernel estimator can be applied to the data which is now of sample size $2n$:

$$f^*_{refl}(x) = \frac{1}{2nh} \sum_{j=1}^{2n} K\left(\frac{x - Y_j}{h}\right), \qquad x \in \mathbb{R} \tag{1.9}$$

This is the standard kernel density estimator. Moreover it is also easy to see that this estimate is symmetric around the origin. Thus, the natural way to get an estimate with support $[0, \infty)$:

$$\hat{f}^*_{refl}(x) := \begin{cases} 2f^*_{refl}(x), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

This is usually referred to as the reflection estimator and it can also be formulated as

$$\hat{f}_{refl}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right) \right\}, \qquad x \geq 0 \tag{1.10}$$

Due to the symmetry of the kernel function, it is very easy to prove that this results in *the reflection estimator*, i.e. $\hat{f}_{refl}(x) = \hat{f}_h(x) + \hat{f}_h(-x)$. This equality allows to calculate the bias and the variance of the estimator in the following way:

$$Bias\left(\hat{f}_{refl}(x)\right) = \frac{h^2}{2} f^{(2)}(x) \int t^2 K(t)\, dt + o\left(h^2\right).$$

$$Var\left(\hat{f}_{refl}(x)\right) = \frac{1}{nh} f(x) \int K^2(t)\, dt + O\left(n^{-1}\right).$$

Example 1.3.1 *Taking boundary problem for rv X with exponential distribution with parameter $\lambda = 0.5$ and sample size $n = 300$. Graphical output figure(1.4) illustrates the boundary correction by the reflection method.*

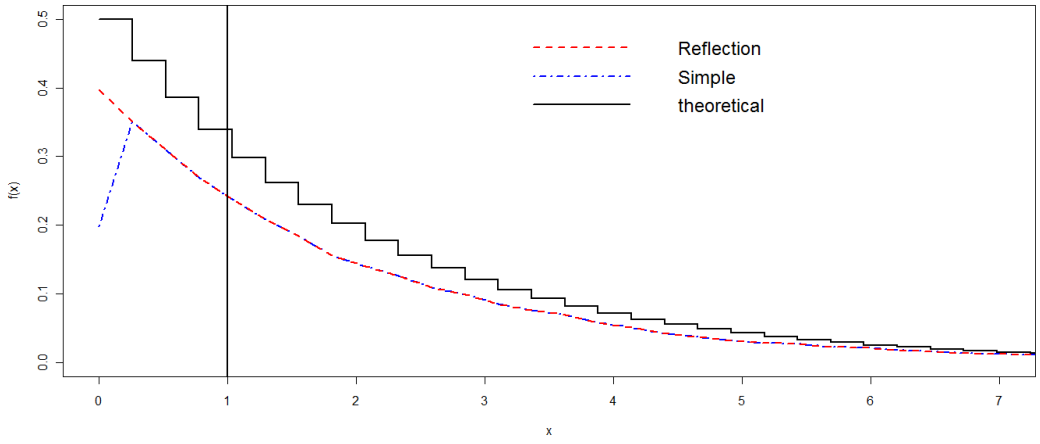

Figure 1.4: Classical (simple) and reflection estimator.

Remark 1.3.1 *As in the reflection estimator, the estimate is set to 0 for $x < 0$. Of course, (1.8) reduces to the standard kernel density estimator (1.1) as soon*

as $x \geq h$. An interesting property is that at 0 this estimator coincides with the reflection estimator (1.10):

$$\hat{f}_{CN}(0) = \frac{1}{nh}\frac{1}{\int_{-1}^{0} K(t)\,dt}\sum_{i=1}^{n} K\left(\frac{0-X_i}{h}\right) = \frac{1}{nh}\frac{1}{1/2}\sum_{i=1}^{n} K\left(\frac{-X_i}{h}\right)$$
$$= \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{-X_i}{h}\right) + K\left(\frac{X_i}{h}\right)\right\} = \hat{f}_{refl}(0)$$

since K is a symmetric function.

1.3.3 Generalized Jackknifing method

Jones (1993) sets up a unified approach to many of the more straightforward methods using *"generalized jackknifing"* developed by Schucany *et al.,* (1971). Let $\hat{f}_h$ be the standard kernel density estimator. We have

$$\bar{f}(x) = \frac{\hat{f}_h(x)}{a_0(c)} = \hat{f}_{CN}(x)$$

with $a_l(c) = \int_{-1}^{(c\wedge 1)} t^l K(t)\,dt$, and $k_l(c) = \int_{-1}^{(c\wedge 1)} t^l L(t)\,dt$.

Let also $\tilde{f}$ be like $\bar{f}$ only with kernel function L

$$\tilde{f}(x) = \frac{\hat{f}_h(x)}{k_0(c)}$$

Think of $\bar{f}$ and $\tilde{f}$ as being defined only on $[0,\infty)$. Then, in a minor reformulation of the presentation of Jones (1993), generalized jackknifing seeks a linear

combination

$$\hat{f}(x) = \alpha_x \bar{f}(x) + \beta_x \tilde{f}(x) \tag{1.11}$$

with good asymptotic bias properties. Away from the boundary, kernel density estimation typically affords a bias of order h^2 as $h = h(n) \to 0$. It turns out that the choices

$$\alpha_x = k_1(c)\, a_0(c) / \{k_1(c)\, a_0(c) - a_1(c)\, k_0(c)\}$$

$$\beta_x = -a_1(c)\, k_0(c) / \{k_1(c)\, a_0(c) - a_1(c)\, k_0(c)\}$$

allow $O(h^2)$ bias at and near the boundary also. (Note that $k_1(c)\, a_0(c)$ must not equal $a_1(c)\, k_0(c)$). Observe that boundary corrected kernel density estimates typically do not integrate to unity, but could be renormalised to do so.

There are many possible choices for L. It is usually preferred to make L a function of K because then one has a boundary correction derived solely from the "interior kernel" K. Examples include taking $L(t)$ to be $K_c(t) = c^{-1}K(c^{-1}t)$ or $K(t)$ or $K(2p - t)$ or $tK(t)$.

Remark 1.3.2 *A disadvantage of all generalized jackknife boundary corrections, however, is their propensity for taking negative values near the boundary. See the dashed curves in figure (1.5) where* $n = 50$ *data points are simulated from the* $Gamma(3, 1)$ *distribution (but only the boundary region* $0 < x < h$ *is shown). Here,* K *is the biweight kernel and* $h = 1.3$. *The proposed modified boundary*

corrected estimator

$$\hat{f}_C(x) = \bar{f}(x) \exp\left\{\frac{\hat{f}(x)}{\bar{f}(x)} - 1\right\}. \tag{1.12}$$

It is clearly nonnegative because, since K *and* $a_0(c)$ *are nonnegative,* $\bar{f}$ *is nonnegative, and the rest of the formula is exponentiated. That it is a modification of* $\bar{f}$ *"in the direction of"* $\hat{f}$ *is clear, and thus to each* $\hat{f}$ *there corresponds a nonnegative* $\hat{f}_C(x)$*. Indeed, there is no requirement here of generalized jackknifing to obtain* $\hat{f}$*, so the proposal is a completely general non negativisation.*

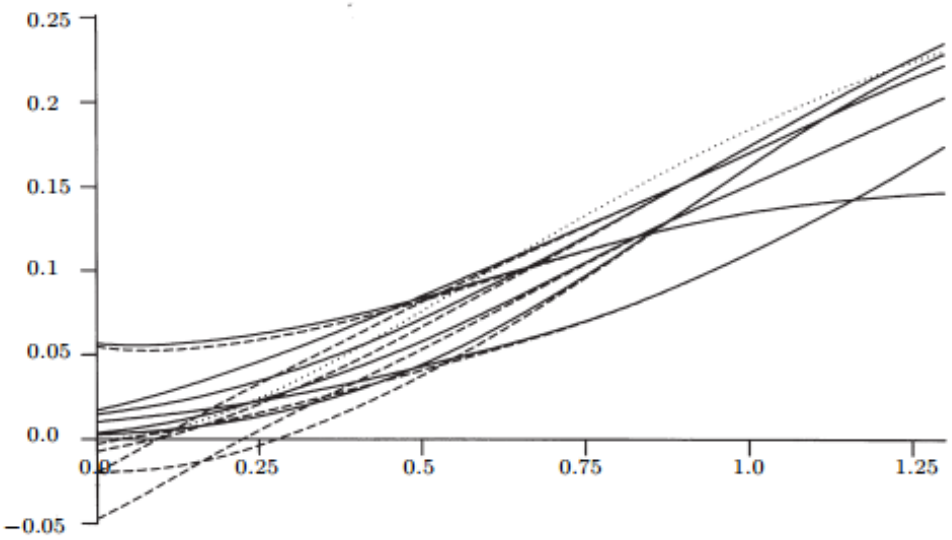

Figure 1.5: Generalized jackknife boundary corrections methods.

The asymptotic means and variances of both $\hat{f}(x)$ and $\hat{f}_C(x)$ are given in the

following theorem. Let

$$B(c) = \frac{k_1(c)\,a_2(c) - a_1(c)\,k_2(c)}{k_1(c)\,a_0(c) - a_1(c)\,k_0(c)},$$

$$V(c) = \frac{k_1^2(c)\,b(c) - 2k_1(c)\,a_1(c)\,e(c) + a_1^2(c)\,z(c)}{\{k_1(c)\,a_0(c) - a_1(c)\,k_0(c)\}^2},$$

where $b(c) = \int_{-1}^{(c\wedge 1)} K^2(t)\,dt$, $e(c) = \int_{-1}^{(c\wedge 1)} tK^2(t)\,dt$, and $z(c) = \int_{-1}^{(c\wedge 1)} t^2K^2(t)\,dt$.

Theorem 1.3.1 *Suppose that f has at least two continuous derivatives. Then, as $n \to \infty$, $h \to 0$ and $nh \to \infty$,*

$$\begin{aligned} Bias\left(\hat{f}(x)\right) &\simeq \tfrac{1}{2}h^2 B(c)\,f^{(2)}(x), \\ Bias\left(f_C(x)\right) &\simeq \tfrac{1}{2}h^2\left(B(c)\,f^{(2)}(x) + \tfrac{a_1^2(c)}{a_0^2(c)}\tfrac{f'^2(x)}{f(x)}\right) \end{aligned}$$

and

$$Var\left(\hat{f}_E(x)\right) \simeq \frac{1}{nh}V(c)\,f(x)$$

where $\hat{f}_E$ denotes either $\hat{f}$ given by (1.11) and $f_C(x)$ given by (1.12).

1.3.4 Translation in the argument of the kernel method

The cut-and-normalized estimator (1.8) converges slowly to the true density function. In Hall and Park (2002) an adaptation of this estimator is presented, density functions with an upper bound were considered. Nevertheless, the estimator will

now be presented for densities with support $[0, \infty)$. For $c = x.h^{-1}$:

$$\hat{f}_{Tag}(x) := \frac{1}{nh} \frac{1}{\int_{-1}^{c} K(t)\, dt} \sum_{i=1}^{n} K\left(\frac{x - X_i + \alpha(x)}{h}\right), \qquad x \geq 0 \tag{1.13}$$

where $\alpha(x)$ is a function to determine.

Remark 1.3.3 *It is clear that if $\alpha(x) = 0$, this estimator would reduce to the cut-and-normalized estimator (1.8). The aim is to find a suitable $\alpha(x)$ such that for $x \geq h$, the estimator (1.13) reduces again to the standard kernel density estimator (1.1). Hence, an estimator must be used and in Hall and Park* (2002) *the following is proposed*

$$\hat{\alpha}(x) = h^2 \frac{\hat{f}^{(1)}(x)}{\hat{f}_{CN}(x)} \frac{1}{K(c)} \int_{-1}^{c} tK(t)dt, \qquad x = ch$$

with $\hat{f}_{CN}(x)$ is given by (1.8) and $\hat{f}^{(1)}(x)$ is an estimate of the first derivative of the density evaluated at x.

Definition 1.3.1 *By a translation in the argument of the kernel, Hall and Park* (2002) *give a boundary correction kernel density estimator, for $c = x.h^{-1}$:*

$$\hat{f}_{TAK.}(x) := \frac{1}{nh} \frac{1}{\int_{-1}^{c} K(t)\, dt} \sum_{i=1}^{n} K\left(\frac{x - X_i + \hat{\alpha}(x)}{h}\right), \qquad x \geq 0.$$

Example 1.3.2 *In figure (1.6), the boundary problem is studied for the sample size of* 300 *and rv* $X \sim \exp(0.2)$. *Classical kernel estimator (Simple), by translation in the argument of the kernel (TAK) and by cut and normalized (CN)*

approach are considered. In the graphical output, we give a comparison of three methods also mentioned. The graph shows that the three methods to coincide for $x > h = 0.3$. *In the boundary region, both CN and TAK estimators improve the simple one.*

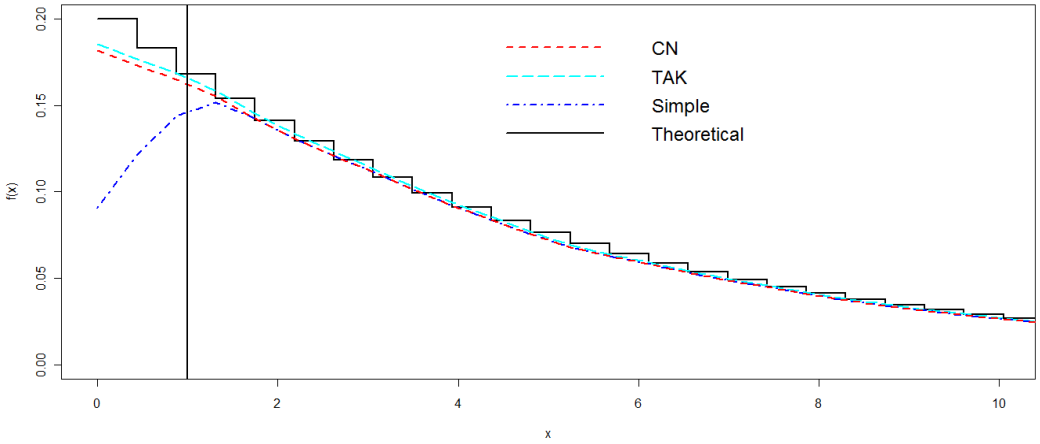

Figure 1.6: Classical (simple), by translation in the argument of the kernel and by the approach of the cut and normalized estimators.

1.3.5 Reflection and transformation methods

The reflection estimator computes the estimate density based on the original and the reflected data points. Unfortunately, this does not always yield a satisfying result since this estimator enforces the shoulder condition and still contains a bias

of order h if the density does not fulfill this condition. The generalized reflection and transformation density estimators is given by

$$\hat{f}_g(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{x+g(X_i)}{h}\right)+K\left(\frac{x-g(X_i)}{h}\right)\right\}, \qquad x \geq 0 \quad (1.14)$$

where g is a transformation that need to be determined.

Remark 1.3.4 *By simply looking at this formula, one could also question the need of using the same function in both arguments: why not use different functions* g_1 *and* g_2*? If chosen in a smart way the bias could possibly be reduced to a higher order with respect to* h*. This idea was pursued in the technical report of Karunamuni and Alberts* (2003) *and later on, in an abbreviated manner, in Karunamuni and Alberts* (2005)*. Special cases were analyzed in Zhang et al.* (1999) *and in Karunamuni and Alberts* (2006)*. The general form of such estimator is the following:*

$$\hat{f}_{grt}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{x+g_1(X_i)}{h}\right)+K\left(\frac{x-g_2(X_i)}{h}\right)\right\}, \qquad x \geq 0 \quad (1.15)$$

g_1, g_2 *are two transformations that need to be determined. The kernel function* K *is nonnegative, symmetric function with support* $[-1,1]$*, and satisfying*

$$\int K(t)\,dt = 1, \int tK(t)\,dt, \text{ and } 0 < \int t^2 K(t)\,dt < \infty.$$

A1 Karunamuni and Alberts (2005) assumed that the transformations g_1, g_2 in (1.15) are non-negative, continuous and monotonically increasing functions

defined on $[0,\infty)$. Further assumed that g_k^{-1} exists, $g_k(0)=0$, $g_k^{(1)}=1$, and that $g_k^{(2)}$ and $g_k^{(3)}$ exist and are continuous on $[0,\infty)$, where $g_k^{(j)}$ denotes the $j^{th}-$derivative of g_k, with $g_k^{(0)}=g_k$ and g_k^{-1} denoting the inverse function of g_k (for $k=1,2$).

A2 Particularly, supposed that $g_1=g_2:=g$ and

$$\begin{aligned} g^{(2)}(0) &= 2\frac{f^{(1)}(0)}{f(0)}\int_c^1 (t-c)K(t)\,dt\left(c+2\int_c^1 (t-c)K(t)\,dt\right)^{-1} \\ &:= d.K_c', \end{aligned} \tag{1.16}$$

where

$$d := \frac{f^{(1)}(0)}{f(0)} \tag{1.17}$$

and

$$K_c' := 2\int_c^1 (t-c)K(t)\,dt\left(c+2\int_c^1 (t-c)K(t)\,dt\right)^{-1}.$$

A3 Supposed further that, $f^{(j)}$ the $j^{th}-$derivatives of f exists and is continuous on $[0,\infty)$, $j=0,1,2$, with $f^{(0)}=f$.

Theorem 1.3.2 *Under the above conditions on f, g_1, g_2, h and K (e.g., A1-A3). For the estimate $\hat{f}_{grt}(x)$ defined in (1.15), we have for $x = ch$, $0 \leq c \leq 1$:*

$$\begin{aligned} Bias\left(\hat{f}_{grt}(x)\right) = \frac{1}{2}h^2 \Bigg\{ & f^{(2)}(0) \int_{-1}^{1} K^2(t)\, dt \\ & - \left[g_c^{(3)}(0) f(0) + g_c^{(2)}(0) \left(f^{(1)}(0) - g_c^{(2)}(0) f(0) \right) \right] \\ & \left(c^2 + \int_{-1}^{1} K^2(t)\, dt \right) \Bigg\} + o\left(h^2\right) \end{aligned} \tag{1.18}$$

and

$$Var\left(\hat{f}_{grt}(x)\right) = \frac{f(0)}{nh} \left(2\int_{c}^{1} K(t) K(2c-t)\, dt + \int_{-1}^{1} K^2(t)\, dt \right) + o\left(\frac{1}{nh}\right). \tag{1.19}$$

1.3.6 Rice's boundary modification density estimator

Rice (1984) proposed a boundary modification of kernel regression estimators. In the boundary area, the method takes a linear combination of two kernel regression estimators based on different bandwidths such that the bias is of the same order of magnitude as in the interior. The idea is similar to the bias reduction technique discussed in Schucany and Sommers (1977). Cheng (2006) adapted the method to the context of density estimation.

Definition 1.3.2 *Given $\alpha > 0$, the Rice's boundary modified kernel estimator of*

$f(x)$, $x = ch$, $c \geq 0$ *is*

$$\overline{f}_{\alpha,h}(x) = a\hat{f}_h(x) - b\hat{f}_{\alpha h}(x) = n^{-1}\sum\nolimits_{i=1}^{n}(aK_h - bK_{\alpha h})(x - X_i) \tag{1.20}$$

where

$$a = \frac{\alpha\mu_{1,c/\alpha}(K)}{\alpha\mu_{0,c/\alpha}(K)\,\alpha\mu_{1,c/\alpha}(K) - \mu_{0,c/\alpha}(K)\,\alpha\mu_{1,c}(K)}, b = \frac{\mu_{1,c}(K)}{\alpha\mu_{1,c/\alpha}(K)}a \tag{1.21}$$

Here, a and b depend on c and are obtained by requiring to have a bias $\overline{f}_{\alpha,h}(x)$ of order h^2, see Rice (1984) *for more details.*

Let

$$\overline{K}_\alpha(.) = aK(.) - \frac{b}{\alpha}K\left(\frac{\cdot}{\alpha}\right)$$

Asymptotic bias and variance of $\overline{f}_{\alpha,h}(x)$ are given in the following theorem.

Theorem 1.3.3 *Under Condition (1.1), for $x = ch, c \geq 0$, as $n \to \infty$, $h \to \infty$ and $nh \to \infty$,*

$$Bias\left(\overline{f}_{\alpha,h}(x)\right) = \frac{h^2}{2}f^{(2)}(0+)\,\mu_{2,c}\left(\overline{K}_\alpha\right) + o\left(h^2\right)$$

and

$$Var\left(\overline{f}_{\alpha,h}(x)\right) = \frac{f(0+)}{nh}\mu_{0,c}\left(\overline{K}_\alpha^2\right) + o\left(\frac{1}{nh}\right).$$

Remark 1.3.5 *Under the above Theorem,* $\overline{f}_{\alpha,h}$ *retains the same rate of convergence in mean squared error everywhere. This method introduces an extra parameter* α*, the ratio of the two bandwidths. Rice* (1984) *recognized that it is difficult to find the best solution for each* c *and suggested taking* $\alpha = 2 - c$*, where* K *is supported on* $[-1, 1]$.

Remark 1.3.6 *In the case of Normal kernels, keeping the bandwidth ratio fixed, for ease and speed of implementation, and a specific bandwidth ratio are suggested.*
i) When the kernel is Gaussian, our asymptotic studies recommend taking $\alpha \equiv 1$.
ii) Hence $\alpha \equiv 1$ *is recommended as a general choice.*
iii) Cheng (2006) *discussed advantages of Rice's boundary modification. For that method, best choice of the bandwidth ratio* α *depends on the density, the sample size, the kernel and the location in a complicated way. He provided both asymptotic and exact formulae of the mean squared errors to analyze the problem. Cheng* (2006) *also performed some analyses in the case of Normal kernel and made some useful suggestions.*

Chapter 2

Boundary correction in kernel regression estimation

This chapter is concerned with the connections between the kernel regression estimation and boundary effect. In the regression function estimation context, Gasser and Müller (1979) identified the unsatisfactory behavior of the Nadaraya Watson regression estimator for points in the boundary region. They proposed optimal boundary kernels but did not give any formulas. However, Gasser and Müller (1979) and Müller (1988) suggested multiplying the truncated kernel at the boundary zone or region by a linear function. The local linear methods developed recently have become increasingly popular in this context (cf. Fan and Gijbels 1996). More recently, in Dai and Sperlich (2010) a simple and effective boundary correction for kernel density and regression estimator is proposed, by applying local bandwidth variation at the boundaries. To remove the boundary effects a

variety of methods have been developed in the literature, the most widely used is the reflection method, the boundary kernel method, the transformation method, the pseudo-data method and the local linear method.

2.1 Nadaraya-Watson estimator

Let Y be a real random variable (rv), and let X be a continuous covariable with probability density function f which is supported within $[0, \infty)$. The real rv's Y and X are repectively called variable of interest and predictor. Our goal is to estimate the regression function, which is the conditional expectation $m(x) := E(Y|X = x)$ (assuming $f(x) \neq 0$). Then the model can be written as

$$Y = m(X) + \epsilon, \tag{2.1}$$

where ϵ is a rv such that $E(\epsilon|X) = 0$ and $Var(\epsilon|X) = \sigma^2 < \infty$.

There exist many interesting nonparametric estimators for the unknown regression function m. Examples of these last can be found in, for instance, Gasser and Müller (1979), Eubank (1988) and Fan and Gijbels (1996). Given a sample of independent replicates of (X, Y), the popular Nadaraya-Watson estimator Nadaraya (1964) and Watson (1964) of m is given by

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}, \tag{2.2}$$

where $h := h_n$ ($h \to 0$ and $nh \to \infty$) is the bandwidth and $K_h(.) := K(./h)$, where K is an integrable smoothing kernel which usually is nonnegative, i.e., a symmetric probability density function with compact support. There have been numerous activities to study $m_n(x)$, see Härdle (1990) and Wand and Jones (1995) for a review.

Conditions 2.1

- $E(Y^2) < \infty$ *and* $E(X^2) < \infty$.
- m *is twice continuously differentiable in a neighborhood of* x.

Theorem 2.1.1 *We have, under conditions (2.1), as* $h \to 0$, $nh \to \infty$ *for* $n \to \infty$

$$Bias\left(\hat{m}_h(x)\right) = \frac{h^2\left(2f^{(2)}(x)\, m^{(1)}(x) + f(x)\, m^{(2)}(x)\right)\mu_2(K)}{2f(x)} + o\left(h^2\right) \quad (2.3)$$

and

$$Var\left(\hat{m}_h(x)\right) = \frac{\sigma^2}{nhf(x)}\mu_0\left(K^2\right) + o\left(\frac{1}{nh}\right). \quad (2.4)$$

2.2 Some boundary corrections methods in kernel regression estimation

Nonparametric regression function estimators usually show a sharp increase in variance and bias when estimating $m(.)$ at points near the boundary of the support of the function (e.g., $x < h$). Gasser and Müller (1979, 1984) identified the

crucial nature of these effects. They proposed optimal boundary kernels but did not give any formulas. However, Gasser and Müller (1979) and Müller (1988) suggested multiplying the truncated kernel at the boundary by a linear function. Rice (1984) proposed another approach using a generalized jackknife, also known as Richardson extrapolation which linearly combines the two bandwidths. Schuster (1985) introduced a reflection technique for density estimation. Eubank and Speckman (1991) have given a method for removing boundary effects using a "*bias reduction theorem*". The fundamental idea of their work is to use a biased estimator to improve another estimator in some sense. Müller (1991) proposed an explicit construction for a boundary kernel which is the solution of a variational problem under asymmetric support. He tables many polynomials that are optimal in a specified sense. Moreover, Müller (1993) introduced a general method of constructing a boundary kernel which is the solution of a variational problem involving a certain weight function. More recently, Müller and Wang (1994) gave explicit formulas for a new class of polynomial boundary kernels.

In the context of density estimation, Wand and Schucany (1990) and Berlinet (1993) worked with the Gaussian kernel which exhibits a first-order boundary effect because the Gaussian kernel has noncompact support. In fact, Berlinet (1993) proposed a framework for building kernels of increasing order apart from some specific methods based on moment relationships.

2.2.1 Gasser and Müller estimator

Gasser and Müller (1979) proposed the following estimator

$$\hat{m}_n(x) = \frac{1}{h}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du Y_i$$

with $s_i, i = 1, ..., n$ a sequence defined as follows:

$$s_0 = 0, \quad s_{i-1} \leq X_i \leq s_i, \quad (i = 1, ..., n) \quad , \; s_n = 1$$

A natural choice for s_i $(i = 1, ..., n-1)$ is:

$$s_i = \frac{1}{2}(X_i + X_{i+1})$$

Conditions 2.2 *K fulfills a Lipschitz condition of order γ_K* $(0 < \gamma_K \leq 1)$.

The basic requirement for the design is:

$$\max_i |X_i - X_{i-1}| = o\left(\frac{1}{n}\right)$$

but at some points they require form of asymptotic equidistant with rate $\delta > 1$:

$$\max_i \left| s_i - s_{i-1} - \frac{1}{n} \right| = o\left(\frac{1}{n^\delta}\right)$$

For an equidistant design they can put the terms involving $o\left(\frac{1}{n^{\delta}}\right)$ equal to zero.

Remark 2.2.1 *The Gasser and Müller regression estimator is a modification of an earlier version of Priestly and Chao* (1972)*, and is similar to that of Cheng and lin* (1981)*. The special case of* $s_i = X_i$ *has been investigated by Cheng and Lin* (1981) *.*

Theorem 2.2.1 *assuming* m *to be Lipschitz continuous of order* γ_m

$$E\left(\hat{m}_n\left(x\right)\right) = \frac{1}{h}\int_0^1 K\left(\frac{x-u}{h}\right) m\left(u\right) du + o\left(\frac{1}{n^{\gamma_m}}\right)$$

and

$$Var\left(\hat{m}_n\left(x\right)\right) = \frac{\sigma^2}{nh}\int_{-1}^1 K^2\left(u\right) du + o\left[\frac{1}{n^{1+\gamma_K}.h^{1+\gamma_K}} + \frac{1}{n^{\delta}h}\right]$$

2.2.2 Cut-and-Normalized regression estimator

A method of *cut-and-normalize* was first introduced by Gasser and Müller (1979). For simplicity, only the left boundary effects, i.e., $c = x/h < 1$, we discussed here. The right boundary effects proceed in the same manner. Since Gasser and Müller (1979) investigated the *cut-and-normalize* method, we briefly explain the general approach described above.

Therefore, a boundary kernel modification of m is

$$\hat{m}_{CN}(x) = \frac{1}{nh}\sum_{i=1}^{n} K_{1c}\left(\frac{x - X_i}{h}\right) Y_i, \tag{2.5}$$

where

$$K_{1c}(t) = \frac{K(t)}{\int_{-1}^{c} K(u)\, du}, \qquad -1 \leq t \leq c \tag{2.6}$$

Further, this is '*normalized*' in the sense that it is rescaled to integrate into $(0, 1)$. Then, the corresponding Bias is

$$Bias\left(\hat{m}_{CN}(x)\right) = -hm^{(1)}(x)\int_{-1}^{c} tK_{1c}(t)\, dt + \frac{h^2 m^{(2)}(x)}{2!}\int_{-1}^{c} t^2 K_{1c}(t)\, dt + o\left(h^2\right), \tag{2.7}$$

where $\int_{-1}^{c} tK_{1c}(t)\, dt \neq 0$.

Remark 2.2.2 *The dominant part of* $Bias\left(\hat{m}_{CN}(x)\right)$ *in (2.7) is of order* h*, so* $\hat{m}_{CN}(x)$ *is still subject to more boundary bias. The asymptotic variance of* $\hat{m}_{CN}(x)$ *can be obtained by the same method as for he non boundary, i.e.,*

$$Var\left(\hat{m}_{CN}(x)\right) = \frac{\sigma^2}{nh}\int_{-1}^{c} K_{1c}^2(t)\, dt + o\left(\frac{1}{nh}\right). \tag{2.8}$$

Hence, the asymptotic mean square error has the form

$$AMSE\left(\hat{m}_{CN}(x)\right) = \frac{\sigma^2}{nh}\int_{-1}^{c} K_{1c}^2(t)\, dt + \left[hm^{(1)}(x)\int_{-1}^{c} tK_{1c}(t)\, dt\right]^2. \tag{2.9}$$

2.2.3 Rice's boundary modified regression estimator

The Rice's boundary modified kernel regression estimator (cf. Rice, 1984) is

$$\bar{m}_{\alpha,h}(x) = \hat{m}_h(x) + \beta\left[\hat{m}_h(x) - \hat{m}_{\alpha h}(x)\right],$$

where

$$\beta = \frac{R(c)}{\alpha R(c/\alpha) - R(c)},$$

$R(c) = \omega_1(c)/\omega_0(c)$, and $\omega_l(c) = \int_{-1}^{c} t^l K(t)\,dt$.

Theorem 2.2.2 *The leading bias of* $\bar{m}_{\alpha,h}(x)$ *is*

$$bias\left(\bar{m}_{\alpha,h}(x)\right) = hm^{(1)}(x)\left[-R(c) - \beta R(c) + \alpha\beta R(c/\alpha)\right]$$

Remark 2.2.3 *For the choice of* α*, Rice has recommended the following* : $\alpha = 2 - c$.

Remark 2.2.4 *Rice presents a simple and effective solution to the following problem: if a given kernel,* K *is used in the interior of the interval, how can* K *be smoothly modified near the boundary? one may not choose to use the optimal kernel (Epanechnikov,* 1969*) because of its non differentiability at* ± 1 *and the relatively small dean in MSE, Tapia and Thompson* (1978) *(Although Epanechnikov's kernel was derived to be optimal for the problem of density estimation, a similar derivation shows its optimality for regression).*

Next, to obtain the same local asymptotic behavior, the generalized jackknife method is applied to reduce the order of the bias.

2.2.4 Generalized Jackknif regression estimator

In this section we describe the boundary effects and present a simple and effective solution to the boundary problem. This solution is due to Rice (1984) and uses the (generalized) jackknifing technique. Boundary phenomena have also been discussed by Gasser and Müller (1979) and Müller (1984*b*) who proposed "boundary kernels" for use near the boundary. In the setting of spline smoothing Rice and Rosenblatt (1983) computed the boundary bias. Consider the fixed design error model with kernels having support $[-1, 1]$. Take the kernel estimator

$$\hat{m}_{jh}(x) = (nh)^{-1} \sum_{i=1}^{n} K_h (X_i - x) Y_i$$

which has the expectation

$$E(\hat{m}_{jh}(x)) = \int_{(x-1)/h}^{x/h} K(u) m(x - uh) du + O\left(\frac{1}{nh}\right), \quad as \;\; nh \to \infty.$$

Now let $x = ch \leq 1 - h$, then by a Taylor series expansion the expected value of $\hat{m}_{jh}(x)$ can be approximated by

$$\begin{aligned} & m(x) \int_{-1}^{c} K(u)du - hm^{(1)}(x) \int_{-1}^{c} uK(u)\,du \\ & + \frac{1}{2}h^2 m^{(2)}(x) \int_{-1}^{c} u^2 K(u)\,du \\ & = m(x)\,\omega_0(c) - hm^{(1)}(x)\omega_1(c) + \frac{1}{2}h^2 m^{(2)}(x)\,\omega_2(c) \end{aligned} \tag{2.10}$$

Of course, if $c \geq 1$

$$\begin{cases} \omega_0(c) &= 1 \\ \omega_1(c) &= 0 \\ \omega_2(c) &= d_k \end{cases}$$

and we have the well-known bias expansion for the estimator. The idea of John Rice is to define a kernel depending on the relative location of x expressed through the parameter c. Asymptotic unbiasedness is achieved for a kernel: $K_c(\cdot) = K(\cdot)/\omega_0(c)$.

Remark 2.2.5 *If x is away from the left boundary, that is, $c \geq 1$, then the approximate bias is given by the third term. If $c < 1$, the second term is of dominant order $O(h)$ and thus the bias is of lower order at the boundary than in the center of the interval. The generalized jackknife technique (Gray and Schucany, 1972) allows one to eliminate this lower order bias term.*

Let $\hat{m}_{jh,c}(x)$ be the kernel estimator with kernel K_c and let

$$\hat{m}_{jh}^{J}(x) =(1-R)\hat{m}_{jh,c}(x) + R\hat{m}_{\alpha jh,c}(x)$$

be the *jackknife estimator* of $m(x)$, a linear combination of kernel smoothers with bandwidth h and αh. From the bias expansion (2.10), the leading bias term of $\hat{m}_{jh}^{J}(x)$ can be eliminated if

$$R = -\frac{\omega_1(c)/\omega_0(c)}{\alpha\omega_1(c/\alpha)/\omega_0(c/\alpha) - \omega_1(c)/\omega_0(c)}$$

This technique was also used by Bierens (1987) to reduce the bias inside the observation interval. In effect, the jackknife estimator is using the kernel function

$$K_c^J(x) = (1-R)K(t) - (R/\alpha)K(t/\alpha)$$

where R and α and thus K_c^J depend on c. In this sense, K_c^J can be interpreted as a "boundary kernel". For the choice of α, Rice (1984) has recommended to take :$\alpha = 2 - c$.

Example 2.2.1 *As an example, take as the initial kernel the quartic kernel given by*

$$K(t) = (15/16)\left(1-t^2\right)^2 1_{[-1,1]}.$$

The numbers $\omega_0(c)$, $\omega_1(c)$ can be computed explicitly. Figure (2.1) shows the sequence of boundary kernels K_c^J for $c = 0.1, 0.2, 0.4, 0.6, 0.8$. Note that the kernels

have negative side lobes. Figure (2.2) shows the nonparametric estimate of the function $m(x) = x^2$ *from* $n = 15$ *observations (Gaussian noise,* $\sigma = 0.05$*). The bandwidth* h *is* 0.4*, thus* 60 *percent of the observation interval are due to boundary effects.*

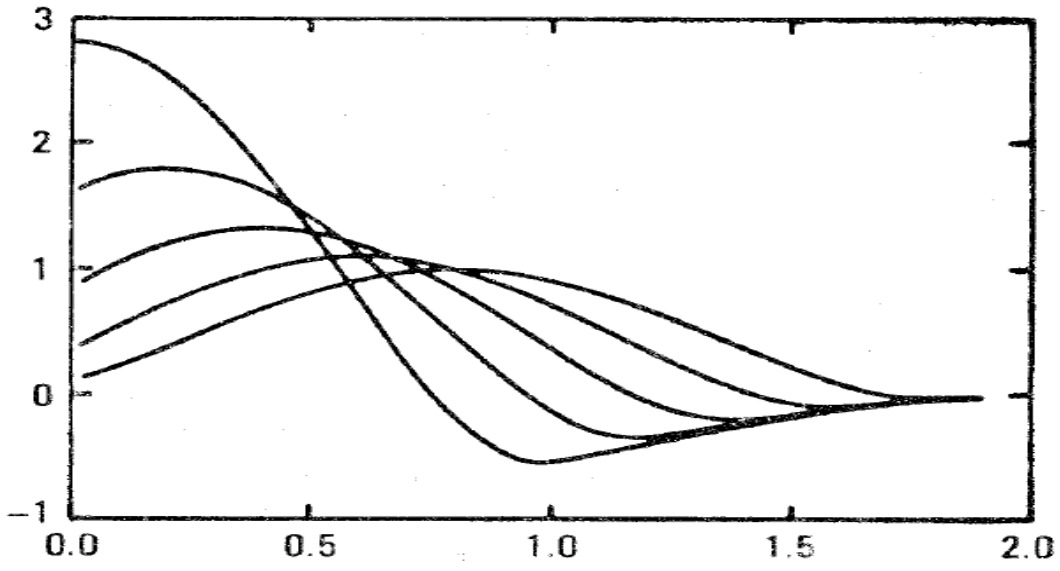

Figure 2.1: Sequence of boundary kernels.

2.2.5 Local linear regression estimator

Most regression estimators studied in the literature are of the form

$$\sum_{i=1}^{n} w_i(x, X_1, ..., X_n) Y_i.$$

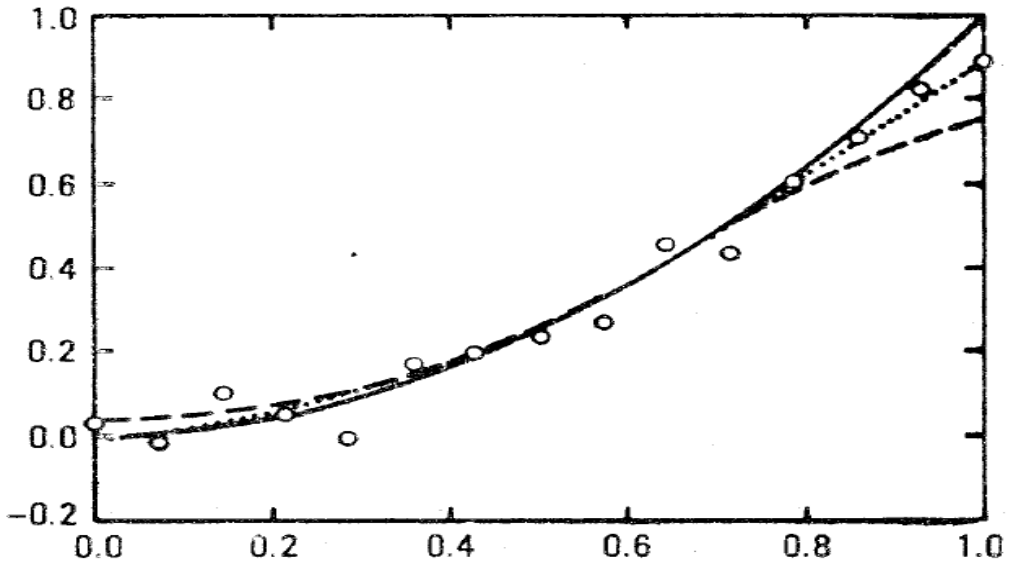

Figure 2.2: Boundary correction in kernels regression estimation: quartic case.

Such a kind of estimator is called a linear smoother (cf. Fan and Gijbels 1996), since it is linear in the response. Consider a linear smoother which is obtained via a local linear approximation to the mean regression function. More precisely, the estimator is defined as $\hat{m}(x)$ where $\hat{a}$ together with $\hat{b}$ minimizes

$$\sum_{i=1}^{n} (Y_i - a - b(x - X_i))^2 K_h(x - X_i) \tag{2.11}$$

It turns out that $\hat{m}(x)$ is the best linear smoother, in the sense that it is the asymptotic minimax linear smoother when the unknown regression function is in the class of functions having bounded second derivative. This property is established in Fan (1992*b*). The preceding idea is an extension of Stone (1977), who used the kernel $K(x) = 1_{[|x|\leq 1]}/2$, resulting in the running line smoother. For a

further motivation and study of linear smothers obtained via a local polynomial approximation to the regression function see Cleveland (1979), Lejeune (1985), Müller (1987), Cleveland and Devlin (1988) and Fan (1992*b*, 1993). Fan and Gijbels (1992) referred to the estimator $\hat{m}(x)$ as a local linear smoother.

The smoothing parameter in (2.11) remains constant, that is, it depends on neither the location of x nor on that of the data X_i. Such an estimator does not fully incorporate the information provided by the density of the data points. Furthermore, a constant bandwidth is not flexible enough for estimating curves with a complicated shape. All these considerations lead to introducing a variable bandwidth $h/\alpha(X_i)$, where $\alpha(.)$ is some nonnegative function reflecting the variable amount of smoothing at each data point. This concept of variable bandwidth was introduced by Breiman, Meisel and Purcell (1977) in the density estimation context. Further related studies can be found in Abramson (1982), Hall and Marron (1988), Hall (1990) and Jones (1990). It is expected that the proposed estimator has all the advantages of both the local linear smoothing method and the variable bandwidth idea. Fan and Gijbels (1992) gave a formal introduction of the estimator. Instead of (2.11), they minimized

$$\sum_{i=1}^{n} \left(Y_i - a - b\left(x - X_i\right)\right)^2 \alpha\left(X_i\right) K_h\left(\left(x - X_i\right)\alpha\left(X_i\right)\right), \tag{2.12}$$

with respect to a and b. Denote the solution to this problem by $\hat{a}$, $\hat{b}$. Then the

regression estimator is defined as $\hat{a}$, which is given by

$$\hat{m}(x) = \hat{a} = \sum_{i=1}^{n} w_i Y_i / \sum_{i=1}^{n} w_i \tag{2.13}$$

where

$$w_i \equiv \alpha(X_i) K_h((x - X_i)\alpha(X_i))[S_{n,2} - (x - X_i) S_{n,1}] \tag{2.14}$$

with

$$S_{n,k} = \sum_{i=1}^{n} \alpha(X_i) K_h((x - X_i)\alpha(X_i))(x - X_i)^k, \qquad k = 0, 1, 2 \tag{2.15}$$

If Fan and Gijbels (1996) take $\alpha(.) = 1$, the preceding result slightly generalizes the known result for the estimator with a constant bandwidth (see Fan, 1992*b*)

$$\hat{m}_l(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)[S_{n,2} - (x - X_i) S_{n,1}] Y_i}{\sum_{i=1}^{n} K_h(x - X_i)[S_{n,2} - (x - X_i) S_{n,1}]}$$

where

$$S_{n,k} = \sum_{i=1}^{n} K_h(x - X_i)(x - X_i)^k, \qquad k = 0, 1, 2 \tag{2.16}$$

Chapter 3

General method of boundary correction in kernel regression estimation

Abstract[1]. Kernel estimators of both density and regression functions are not consistent near the finite end points of their supports. In other words, boundary effects seriously affect the performance of these estimators. In this paper, we combine the transformation and the reflection methods in order to introduce a new general method of boundary correction when estimating the mean function. The asymptotic mean squared error of the proposed estimator is obtained. Simulations show that our method performes quite well with respect to some other existing methods.

[1]This chapter is an Article appeared in Afrika Statistika. Vol. 10, 2015, pages 688–699.(Authors : S. Kheireddine, A. Sayah and D. Yahia).

3.1 Introduction

Let Y be a real random variable (rv), and let X be a continuous covariable with probability density function f which is supported within $[0, \infty)$. Then the model can be written as $Y = m(X) + \epsilon$ where ϵ is a rv such that $E(\epsilon|X) = 0$ and $Var(\epsilon|X) = \sigma^2 < \infty$.

Given a sample of independent replicates of (X, Y), the popular Nadaraya-Watson estimator Nadaraya (1964) and Watson (1964) of m is given by

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)} \tag{3.1}$$

where $h := h_n$ ($h \to 0$ and $nh \to \infty$) is the bandwidth and $K_h(.) := K(./h)$, where K is an integrable smoothing kernel which usually is nonnegative.

Boundary effects are a well known problem in the nonparametric curve estimation setup, no matter if we think density estimation or regression. Moreover, both density and regression estimator usually show a sharp which increase in bias and variance when estimating them at points near the boundary region, i.e., for $x \in [0, h)$, this phenomenon is referred as "boundary effects". In the context of the regression function estimation, Gasser and Müller (1979) identified the unsatisfactory behavior of (2.2) for points in the boundary region. They proposed optimal boundary kernels but did not give any formulas. However, Gasser and Müller (1979) and Müller (1988) suggested multiplying the truncated kernel at the boundary zone or region by a linear function. Rice (1984) proposed

another approach using a generalized jackknife. Schuster (1985) introduced a reflection technique for density estimation. Eubank and Speckman (1991) presented a method for removing boundary effects using a bias reduction theorem. Müller (1991) proposed an explicit construction of a boundary kernel which is the solution of a variational problem under asymmetric support. Moreover, Müller and Wang (1994) gave explicit formulas for a new class of polynomial boundary kernels and showed that these new kernels have some advantages over the smooth optimum boundary kernels in Müller (1991), i.e., these new kernels have higher mean squared error (MSE) efficiency. The local linear methods developed recently have become increasingly popular in this context (cf. Fan and Gijbels, 1996). More recently, in Dai and Sperlich (2010) a simple and effective boundary correction for kernel densitiy and regression estimator is proposed, by applying local bandwidth variation at the boundaries.

To remove the boundary effects a variety of methods have been developed in the literature, the most widely used is the reflection method, the boundary kernel method, the transformation method, the pseudo-data method and the local linear method. They all have their advantages and disadvantages. One of the drawbacks is that some of them (especially boundary kernels), can produce negative estimators. The recent work of Karunamuni and Alberts (2005) provides excellent selective review article on boundary kernel methods and their statistical properties in nonparametric density estimation. In the latter reference, a new boundary correction methodology in density estimation is proposed and studied. It is the purpose of this paper to extend this approach to the regression case.

The rest of the chapter is organized as follows. Section 3.2 introduces our new nonparametric regression estimator and presents some asymptotic results. In Section 3.3, extensive simulations are carried out to compare the proposed estimator with outher ons. Proofs are relegated to Section 3.4.

3.2 Main results

In this paper, we combine the transformation and reflection boundary correction methods to estimate the mean function $\hat{m}_h(x)$. At each point in the boundary region (i.e., for $x = ch$, $0 \leq c \leq 1$), we propose to investigate a class of estimators of the form

$$\begin{aligned} \widetilde{m}_n(x) &= \frac{\sum_{i=1}^{n} Y_i \left\{K_h\left(x + g_1\left(X_i\right)\right) + K_h\left(x - g_1\left(X_i\right)\right)\right\}}{\sum_{i=1}^{n} \left\{K_h\left(x + g_2\left(X_i\right)\right) + K_h\left(x - g_2\left(X_i\right)\right)\right\}} \\ &:= \frac{\widetilde{\varphi}_n(x)}{\widetilde{f}_n(x)} \end{aligned} \tag{3.2}$$

where h is the bandwidth, $K_h(.) := K(./h)$ and K is a kernel function and g_1, g_2 are two transformations that need to be determined. Also, let the kernel function K in (3.2) be a non-negative, symmetric function with support $[-1, 1]$, and satisfying

$$\int K(t)\, dt = 1, \int tK(t)\, dt = 0, \text{ and } 0 < \int t^2 K(t)\, dt < \infty,$$

that is, K is a kernel of order 2.

For $x \geq h$, $\widetilde{m}_n(x)$ reduces to the traditional kernel estimator $\hat{m}_h(x)$ given in (2.2). Thus $\widetilde{m}_n(x)$ is a natural boundary continuation of the usual kernel estimator (2.2). Moreover, estimator (3.2) is non-negative as long as the kernel K is non-negative. Most importantly, the proposed estimator improves the bias while the variance remains almost unchanged.

We assume that the transformations g_1, g_2 in (3.2) are non-negative, continuous and monotonically increasing functions defined on $[0, \infty)$. Further assume that g_k^{-1} exists, $g_k(0) = 0$, $g_k' = 1$, and that g_k'' and g_k''' exist and are continuous on $[0, \infty)$, where g_k^{-1} denoting the inverse function of g_k (for $k = 1, 2$). Particularly, suppose that

$$g_1''(0) = \frac{\varphi'(0)}{\varphi(0)} C_{K,c} \quad \text{and} \quad g_2''(0) = \frac{f'(0)}{f(0)} C_{K,c} \tag{3.3}$$

where

$$C_{K,c} := 2\int_c^1 (t-c) K(t)\, dt \left(2\int_c^1 (t-c) K(t)\, dt + c \right)^{-1}.$$

Suppose further that, $f^{(j)}$, $\varphi^{(j)}$ and $m^{(j)}$ the $j^{th}-$derivatives of f, φ and m exist and are continued on $[0, \infty)$, $j = 0, 1, 2$, with $f^{(0)} = f$, $\varphi^{(0)} = \varphi$ and $m^{(0)} = m$.

The bias and variance of our estimator are given in the following theorem, which is the main result of this paper.

Theorem 3.2.1 *Under the above conditions on f, φ, m, g_1, g_2, and K. For the*

estimate $\widetilde{m}_n(x)$ *defined in (3.2), we have for* $x = ch$, $0 \leq c \leq 1$:

$$Bias(\widetilde{m}_n(x)) = \frac{h^2 \left(A_1 - m\left(x\right) A_2\right)}{f\left(x\right)} + o(h^2), \tag{3.4}$$

and

$$Var\left(\widetilde{m}_n(x)\right) = \frac{f(0)\sigma^2(0)}{nhf^2(x)} \left(\int_{-1}^{1} K^2\left(t\right) dt + 2 \int_{c}^{1} K\left(t\right) K\left(2c - t\right) dt \right) + o\left(\frac{1}{nh}\right). \tag{3.5}$$

where

$$A_1 := \varphi''\left(0\right) \int_{-1}^{1} t^2 K\left(t\right) dt - \left[g_1'''\left(0\right) \varphi\left(0\right) + 3 g_1''\left(0\right) \left(\varphi^{(1)}\left(0\right) - g_1''\left(0\right) \varphi\left(0\right) \right) \right]$$
$$\left(\int_{-1}^{1} t^2 K\left(t\right) dt + c^2 \right), \tag{3.6}$$

$$A_2 := f''\left(0\right) \int_{-1}^{1} t^2 K\left(t\right) dt - \left[g_2'''\left(0\right) f\left(0\right) + 3 g_2''\left(0\right) \left(f'\left(0\right) - g_2''\left(0\right) f\left(0\right) \right) \right]$$
$$\left(\int_{-1}^{1} t^2 K\left(t\right) dt + c^2 \right), \tag{3.7}$$

and $\sigma^2(x) = Var\left(Y/X = x\right)$.

Hence, the MSE of $\widetilde{m}_n(x)$ is

$$MSE\left(\widetilde{m}_n(x)\right) = Bias^2(\widetilde{m}_n(x)) + Var\left(\widetilde{m}_n(x)\right)$$

The asymptotic MSE of $\widetilde{m}_n(x)$ is

$$AMSE\left(\widetilde{m}_n(x)\right) = \frac{h^4\left(A_1 - m\left(x\right)A_2\right)^2}{f^2\left(x\right)} + \frac{f(0)\sigma^2(0)}{nhf^2(x)}\left(\int_{-1}^{1} K^2\left(t\right)dt + 2\,.\right.$$
$$\left.\int_{c}^{1} K\left(t\right)K\left(2c-t\right)dt\right)$$

On the basis of Theorem 3.2.1, the asymptotic optimal bandwidth that minimizes the $AMSE$ is

$$h_{opt} = Cn^{-1/5} \quad \text{with} \quad C = \left(\frac{\sigma^2\left(0\right)f\left(0\right)\left(2\int_c^1 K\left(t\right)K\left(2c-t\right)dt + \int K^2\left(t\right)dt\right)}{4\left(A_1 - m\left(x\right)A_2\right)^2}\right)^{1/5}. \tag{3.8}$$

Remark 3.2.1 *Functions satisfying the conditions (3.3) can be easily constructed. We employ the following transformation in our investigation. For* $0 \leq c \leq 1$, *define*

$$g_k(y) = y + \frac{1}{2}d_k y^2 + \lambda_0 d_k^2 y^3, \qquad k = 1, 2 \tag{3.9}$$

where $d_1 = g_1''\left(0\right)$ *(resp.* $d_2 = g_2''\left(0\right)$*) and* λ_0 *is a positive constant such that* $12\lambda_0 > 1$. *This condition on* λ_0 *is necessary for* $g_k(y)$ *of (3.9) to be an increasing function in* y.

Remark 3.2.2 *The choice h_{opt} of h is only possible in a simulation study, when all required quantities are known, but not in a real data situation. To select the bandwidth for the new method in practice, we can replace the unknown quantities in (3.8) by their estimates. Another method is to use leave-one-out cross-validation (cf. Härdle and Vieu,* 1992*) to select the bandwidth h, i.e., we find h by minimizing*

$$CV\,(h) = \sum\nolimits_{i=1}^{n} \left(y_j - \widetilde{m}_{i,h}(x_i)\right)^2,$$

here $\widetilde{m}_{i,h}(.)$ is the proposed regression estimate by leaving the ith observation out.

3.3 Simulation results

In this section, we present some simulation results which are designed to illustrate the performance of our estimator (3.2) for small sample and large sizes. For comparison purposes, the local linear and the classical Nadaraya–Watson estimators (2.2) were also considered. Recently, local polynomial fitting, and particularly its special case - local linear fitting - have become increasingly popular in light of recent works by Cleveland and Loader (1996), Fan (1992*b*) and several others. It has the advantages of achieving full asymptotic minimax effeciency and automatically correcting for boundary bias. A review of local polynomial smoothing is given in Fan and Gijbels (1996). The local linear regression estimator is given by

$$\hat{m}_l(x) = \sum_{j=1}^{n} w_j Y_j / \sum_{j=1}^{n} w_j, \quad w_j := K_h\left(X_j - x\right)\left(S_{n,2} - S_{n,1}\left(X_j - x\right)\right),$$

where $S_{n,k} := \sum_{j=1}^{n} K_h (X_j - x) (X_j - x)^k$, for $k = 1, 2$.

To assess the effect of the correction methods near the boundaries, the following models are investigated:

$$Model\ 1 : m_1(x) = 2x + 1 \quad \text{and} \quad Model\ 2 : m_2(x) = 2x^2 + 3x + 1$$

and errors ε_j, assumed to be standard normally distributed independent rv's. Likewise, consider two cases of density f with support $[0, \infty)$ of the continuous covariable X :

$$density\ 1 : f_1(x) = \exp(-x) \quad \text{and} \quad density\ 2 : f_2(x) = \frac{2}{\pi (1 + x^2)} \quad x \geq 0.$$

For each density f_1, f_2 and models m_1, m_2 we calculate the absolute biases and $MSE's$ of the proposed general transformation and reflection (GTR), the local linear (LL) and Nadaraya-Watson (NW) estimators, in left boundary region (i.e., $x = ch$; for $c = 0.1$, 0.2, 0.3, and 0.4). The bandwidth selection is based on cross-validation procedure. The main reason for this choice is that it provides a fair basis for comparison among the different estimators regardless of bandwidth effects.

Throughout our simulations, we use the Epanechnikov kernel (cf. Epanechnikov, 1969)

$$K(t) = (3/4)\left(1 - t^2\right) 1_{[-1,1]}(t),$$

where $1_A(.)$ denotes the indicator function of a set A.

The simulated sample sizes are $n = 50$ (small) and $n = 500$ (large). All results are calculated by averaging over 1000 simulation runs. For each model and each density, we calculate the absolute bias and the MSE of the estimators at the points in the mentioned boundary region. The results are shown in Tables **3.1** and **3.2**. We see that in all cases the standard Nadaraya-Watson estimator $\hat{m}_h(x)$ is the worst one. This is clearly due to the boundary effect. Furthermore, when looking at the $MSE's$, our new method outperforms the others. The bias is about the same for our method and the local linear one.

3.4 Proofs

Proof of (3.4). For $x = ch$, $0 \leq c \leq 1$, we have

$$\begin{aligned}\widetilde{m}_n(x) &= \frac{\sum_{i=1}^{n} Y_i \left\{K_h\left(x + g_1\left(X_i\right)\right) + K_h\left(x - g_1\left(X_i\right)\right)\right\}}{\sum_{i=1}^{n} \left\{K_h\left(x + g_2\left(X_i\right)\right) + K_h\left(x - g_2\left(X_i\right)\right)\right\}} \\ &:= \frac{\widetilde{\varphi}_n(x)}{\widetilde{f}_n(x)},\end{aligned}$$

where g_1 and g_2 are given in (3.3). For the numerator $\widetilde{\varphi}_n(x)$, we have

$$\begin{aligned} E\left[\widetilde{\varphi}_n(x)\right] &= \frac{1}{h}\int\int \left\{K_h\left(x+g_1(u)\right)+K_h\left(x-g_1(u)\right)\right\} yf\left(u,y\right)dydu \\ &= \frac{1}{h}\int \left\{K_h\left(x+g_1(u)\right)+K_h\left(x-g_1(u)\right)\right\}\varphi\left(u\right)du \\ &= \frac{1}{h}\int K_h\left(x+g_1(u)\right)\varphi\left(u\right)du+\frac{1}{h}\int K_h\left(x-g_1(u)\right)\varphi\left(u\right)du \\ &=: I_1+I_2, \end{aligned}$$

where $\varphi\left(u\right)=\int yf\left(u,y\right)dy$.

Let $t=\left(x+g_1(u)\right)/h$, then

$$I_1=\int_c^1 K\left(t\right)\frac{\varphi\left(g_1^{-1}\left(h\left(t-c\right)\right)\right)}{g_1^{(1)}\left(g_1^{-1}\left(h\left(t-c\right)\right)\right)}dt.$$

A Taylor expansion of order 2 of the function $\varphi\left(g_1^{-1}\left(.\right)\right)/g_1^{(1)}\left(g_1^{-1}\left(.\right)\right)$ at

$t=c$ gives

$$\begin{aligned} I_1 = \int_c^1 K\left(t\right)\Big[&\varphi\left(0\right)+h\left(t-c\right)\left(\varphi'\left(0\right)-g_1''\left(0\right)\varphi\left(0\right)\right) \\ &+\frac{h^2\left(t-c\right)^2}{2}\left\{\varphi''\left(0\right)-g_1'''\left(0\right)\varphi\left(0\right)-3g_1''\left(0\right)\left(\varphi'\left(0\right)-g''\left(0\right)\varphi\left(0\right)\right)\right\}\Big]dt \\ &+o\left(h^2\right) \end{aligned}$$

$$I_1 = \varphi(0)\int_c^1 K(t)\,dt + h\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\int_c^1 (t-c)\,K(t)\,dt$$
$$+ \frac{h^2}{2}\left\{\varphi''(0) - g_1'''(0)\,\varphi(0) - 3g_1'''(0)\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\right\}$$
$$\int_c^1 (t-c)^2\,K(t)\,dt + o\left(h^2\right). \tag{3.10}$$

Similarly,

$$I_2 = \varphi(0)\int_{-1}^c K(t)\,dt - h\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\int_{-1}^c (t-c)\,K(t)\,dt$$
$$+ \frac{h^2}{2}\left\{\varphi''(0) - g_1'''(0)\,\varphi(0) - 3g_1''(0)\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\right\}$$
$$\times\int_{-1}^c (t-c)^2\,K(t)\,dt + o\left(h^2\right). \tag{3.11}$$

Using the properties of K, we have

$$\int_{-1}^c tK(t)\,dt = -\int_c^1 K(t)\,dt \quad \text{and} \quad \int_{-1}^c K(t)\,dt = 1 - \int_c^1 K(t)\,dt.$$

Also, by the existence and the continuity of $\varphi''(.)$ near 0, we have for $x = ch$,

$$\begin{aligned}
&\varphi(0) = \varphi(x) - ch\varphi'(x) + \tfrac{(ch)^2}{2}\varphi''(x) + o\left(h^2\right),\\
&\varphi'(x) = \varphi'(0) + ch\varphi''(0) + o(h),\\
&\varphi''(x) = \varphi''(0) + o(1).
\end{aligned} \tag{3.12}$$

Now combining (3.10) and (3.11) and using the properties of K along with (3.12),

we have for $x = ch$, $0 \leq c \leq 1$

$$\begin{aligned} E\left[\widetilde{\varphi}_n(x)\right] &= \frac{1}{h} E\left[K_h\left(x + g_1\left(X_i\right)\right) Y_i\right] + \frac{1}{h} E\left[K_h\left(x - g_1\left(X_i\right)\right) Y_i\right] \\ &= \varphi(0) \int_c^1 K(t)\, dt + \varphi(0) \int_{-1}^c K(t)\, dt + h\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right) \\ &\times \int_c^1 (t-c) K(t)\, dt - h\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right) \int_{-1}^c (t-c) K(t)\, dt \\ &+ \frac{h^2}{2}\left\{\varphi'(0) - g_1'''(0)\,\varphi(0) - 3g_1''(0)\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\right\} \\ &\times \int_c^1 (t-c)^2 K(t) + \frac{h^2}{2}\left\{\varphi''(0) - g_1'''(0)\,\varphi(0) - 3g_1''(0)\right. \\ &\left.\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\right\} \int_c^1 (t-c)^2 K(t)\, dt + o\left(h^2\right). \end{aligned} \tag{3.13}$$

Furthermore, the kernel K provides

$$\int_{-1}^1 (t-c)^2 K(t)\, dt = \int_{-1}^1 t^2 K(t)\, dt + c^2,$$

and

$$\int_c^1 (t-c) K(t)\, dt - \int_{-1}^c (t-c) K(t)\, dt = 2\int_c^1 (t-c) K(t)\, dt + c.$$

From (3.13) we have

$$
\begin{aligned}
E\left[\widetilde{\varphi}_n(x)\right] &= \varphi(x) + h\left(\varphi'(0) - g_1'(0)\,\varphi(0)\right)\{2\int_c^1 (t-c)\,K(t)\,dt + c\} \\
&\quad + \frac{h^2}{2}\left\{\varphi''(0) - g_1'''(0)\,\varphi(0) - 3g_1''(0)\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\right\} \\
&\quad \times \{\int_{-1}^1 t^2 K(t)\,dt + c^2\} + o\left(h^2\right) \\
&= \varphi(x) + h\left\{2\varphi'(0)\int_c^1 (t-c)\,K(t)\,dt - g_1''(0)\,\varphi(0)\right. \\
&\quad \left.\times\ \{2\int_c^1 (t-c)\,K(t)\,dt + c\}\right\} + \frac{h^2}{2}\left\{\varphi''(0)\int_{-1}^1 t^2 K(t)\,dt\right. \\
&\quad \left. -\ \left[g_1'''(0)\,\varphi(0) + 3g_1''(0)\left(\varphi'(0) - g_1''(0)\,\varphi(0)\right)\right](\int_{-1}^1 t^2 K(t)\,dt + c^2)\right\} \\
&\quad + o\left(h^2\right). \qquad (3.14)
\end{aligned}
$$

Under the condition (3.3) on the transformation g_1, the second order term of the right-hand side of (3.14) is zero. It can be shown that

$$
E\left[\widetilde{\varphi}_n(x)\right] - \varphi(x) \;=:\; h^2 A_1 + o\left(h^2\right),
$$

where A_1 is given in (3.6).

Similarly, we can get

$$E\left[\widetilde{f}_n(x)\right] = f(x) + h\left\{2f'(0)\int_c^1 (t-c)K(t)\,dt - g_2''(0)f(0) \times \{2\int_c^1 (t-c)K(t)\,dt + c\}\right\} + \frac{h^2}{2}\left\{f''(0)\int_{-1}^1 t^2K(t)\,dt - [g_2''(0)f(0) + 3g_2''(0)(f'(0) - g_2''(0)f(0))]\{\int_{-1}^1 t^2K(t)\,dt + c^2\}\right\} + o\left(h^2\right) \tag{3.15}$$

Substitute $g_2''(0)$, the second term of the right-hand side of (3.15) is zero. Then

$$E\left[\widetilde{f}_n(x)\right] - f(x) =: h^2A_2 + o\left(h^2\right)$$

where A_2 is given in (3.7). Hence

$$\widetilde{m}_n(x) = \frac{h^2A_1 + o(h^2)}{h^2A_2 + o(h^2)} = m(x) + \frac{h^2(A_1 - m(x)A_2)}{f(x)} + o\left(h^2\right).$$

The asymptotic bias result (3.4) follows directly.

Proof of (3.5**).** In order to find the asymptotic variance of the proposed estimator (3.2), we may write

$$\widetilde{m}_n(x) = \sum_{i=1}^n W_{ni}(x)Y_i,$$

with

$$W_{ni}(x) = \frac{K_h\left(x + g_1\left(X_i\right)\right) + K_h\left(x - g_1\left(X_i\right)\right)}{\sum_{i=1}^{n}\left\{K_h\left(x + g_2\left(X_i\right)\right) + K_h\left(x - g_2\left(X_i\right)\right)\right\}}.$$

The weights $W_{ni}(x)$ are nonnegative and satisfy $\sum_{i=1}^{n} W_{ni}(x) = 1$, for all $x \in \mathbb{R}$. Moreover, we have

$$\begin{aligned}\widetilde{m}_n\left(x\right) - m\left(x\right) &= \sum_{i=1}^{n} W_{ni}(x)\left\{Y_i - m\left(X_i\right)\right\} + \sum_{i=1}^{n} W_{ni}(x)\left\{m\left(X_i\right) - m\left(x\right)\right\} \\ &=: \mathcal{J}_1 + \mathcal{J}_2.\end{aligned}$$

Here $\mathcal{J}_1$ is the variance which is study here. Recall that the predictable quadratic variation of $\mathcal{J}_1$ equals

$$\widetilde{f}_n^2(x)\sum_{i=1}^{n} W_{ni}^2(x)\sigma^2(X_i) = (nh)^{-2}\sum_{i=1}^{n}\sigma^2(X_i)\left\{K_h\left(x + g_1\left(X_i\right)\right) + K_h\left(x - g_1\left(X_i\right)\right)\right\}^2,$$

where $\sigma^2(.)$ is the conditional variance i.e., $\sigma^2(.) = Var\left(Y|X = .\right)$.

For $x = ch$, $0 \leq c \leq 1$, we have, using a Taylor expansion of order 2

$$
\begin{aligned}
& E\left[(nh)^{-2}\sum_{i=1}^{n}\sigma^2(X_i)\left\{K_h\left(x+g_1\left(X_i\right)\right)+K_h\left(x-g_1\left(X_i\right)\right)\right\}^2\right] \\
&= \frac{1}{nh^2}E\left[\sigma^2(X_i)\left\{K_h\left(x+g_1\left(X_i\right)\right)+K_h\left(x-g_1\left(X_i\right)\right)\right\}^2\right] \\
&= \frac{1}{nh^2}\int \sigma^2(u)\left\{K_h\left(x+g_1(u)\right)+K_h\left(x-g_1(u)\right)\right\}^2 f(u)du \\
&= \frac{1}{nh^2}\left[\int \sigma^2(u)K_h^2\left(x+g_1(u)\right)f(u)du+\int \sigma^2(u)K_h^2\left(x-g_1(u)\right)f(u)du\right] \\
&+ \frac{2}{nh^2}\int \sigma^2(u)K_h\left(x+g_1(u)\right)K_h\left(x-g_1(u)\right)f(u)du \\
&=: \mathcal{J}_{11}+\mathcal{J}_{12}.
\end{aligned}
$$

Firstly,

$$
\begin{aligned}
\mathcal{J}_{11} &= \frac{1}{nh^2}\left[h\int_c^1 \sigma^2(g_1^{-1}\left((t-c)\,h\right))K^2\left(t\right)\frac{f\left(g_1^{-1}\left((t-c)\,h\right)\right)}{g_1'\left(g_1^{-1}\left((t-c)\,h\right)\right)}dt\right. \\
&\quad \left.+h\int_{-1}^c \sigma^2(g_1^{-1}\left((c-t)\,h\right))K^2\left(t\right)\frac{f\left(g_1^{-1}\left((c-t)\,h\right)\right)}{g_1'\left(g_1^{-1}\left((c-t)\,h\right)\right)}dt\right] \\
&= \frac{f(0)\sigma^2(0)}{nh}\int_{-1}^1 K^2\left(t\right)dt+o(\frac{1}{nh}).
\end{aligned} \tag{3.16}
$$

Next we consider $\mathcal{J}_{12}$. By the continuity property of g_1'' and by a Taylor expansion of order 2 of g_1, we have

$$
\begin{aligned}
g_1\left(\left(c-t\right)h\right) &= g_1\left(0\right)+\left(t-c\right)\left(-h\right)g_1'\left(0\right)+O\left(h^2\right) \\
&= \left(c-t\right)h+O\left(h^2\right),
\end{aligned}
$$

since $g_1(0) = 0$ and $g_1'(0) = 1$. Using (3.16) and by the change of variables, $x + g_1(y) = ht$, we obtain

$$\begin{aligned}
\mathcal{J}_{12} &= \frac{2}{nh^2}\int_0^\infty \sigma^2(u) K_h\left(x + g_1\left(X_i\right)\right) K_h\left(x - g_1\left(X_i\right)\right) f(u)du \\
&= \frac{2}{nh}\int_c^1 \sigma^2\left(g_1^{-1}\left(th - x\right)\right) K\left(t\right) K_h\left(x - g_1(g_1^{-1}\left(th - x\right))\right) \\
&\times f(g_1^{-1}\left(th - x\right))dt \\
&= \frac{2}{nh}\int_c^1 \sigma^2\left(g_1^{-1}\left(th - x\right)\right) K\left(t\right) K_h\left(x - \left(t - c\right)h + O\left(h^2\right)\right) \\
&\times f(g_1^{-1}\left(th - x\right))dt \\
&= \frac{2}{nh}\int_c^1 \sigma^2\left(0\right) K\left(t\right) K\left(2c - t + O\left(h\right)\right)\left(f(0) + O\left(h\right)\right)dt \\
&= \frac{2\sigma^2\left(0\right) f(0)}{nh}\int_c^1 K\left(t\right) K\left(2c - t\right)dt + o\left(\frac{1}{nh}\right). \qquad (3.17)
\end{aligned}$$

The proof of (3.5) now follows from (3.16) and (3.17), which achieves the proof of Theorem 3.2.1.

Table 3.1: Bias and MSE of the indicated regression estimators at boundary

		$\lvert Bias\rvert$	MSE	$\lvert Bias\rvert$	MSE	$\lvert Bias\rvert$	MSE	$\lvert Bias\rvert$	MSE
		$c = .1$		$c = .2$		$c = .3$		$c = .4$	
	GTR	.0141	.0613	.0381	.0631	.0461	.0584	.0512	.0571
$n = 50$	NW	.2678	.1362	.2192	.1120	.1775	.0894	.1365	.0700
	LL	.0375	1.2375	.0064	.3649	.0064	.1468	.0101	.1167
$Model\ 1$	$density1$								
	GTR	.0109	.0083	.0126	.0090	.0162	.0091	.0199	.0084
$n = 500$	NW	.1747	.0393	.1503	.0313	.1217	.0233	.0954	.0163
	LL	.0127	.0386	.0025	.0240	.0006	.0169	.0049	.0123
	GTR	.1361	.0595	.1877	.0786	.1841	.0785	.1304	.0656
$n = 50$	NW	.5705	.3934	.4356	.2540	.3413	.1782	.2923	.1453
	LL	.0694	.1252	.1334	.0903	.1991	.1002	.2158	.1054
$Model\ 2$	$density1$								
	GTR	.0940	.0141	.0948	.0146	.0778	.0131	.0582	.0107
$n = 500$	NW	.3520	.1320	.2854	.0887	.2343	.0627	.1951	.0454
	LL	.0458	.0300	.0955	.0263	.1327	.0305	.1517	.0358

Table 3.2: Bias and MSE of the indicated regression estimators at boundary

		$\lvert Bias\rvert$	MSE	$\lvert Bias\rvert$	MSE	$\lvert Bias\rvert$	MSE	$\lvert Bias\rvert$	MSE
		$c = .1$		$c = .2$		$c = .3$		$c = .4$	
	GTR	.1131	.1189	.1064	.1084	.0876	.0813	.1216	.0769
$n = 50$	NW	.8697	.8014	.6174	.4329	.4455	.2529	.2630	.1279
	LL	.2662	.7402	.0818	.2944	.0245	.2785	.0044	.1955
$Model\ 1$	$density2$								
	GTR	.0496	.0196	.0520	.0167	.0455	.0131	.0391	.0107
$n = 500$	NW	.7162	.5180	.5016	.2577	.3621	.1374	.2511	.0690
	LL	.0063	.0601	.0054	.0373	.0040	.0238	.0007	.0159
	GTR	.1257	.1758	.1205	.1411	.1339	.1119	.1479	.0946
$n = 50$	NW	.6272	.4902	.6220	.4794	.6345	.4856	.7076	.5848
	LL	.1849	1.8400	.0657	.2573	.1288	.1719	.2294	.1398
$Model\ 2$	$density2$								
	GTR	.0505	.0260	.0562	.0201	.0663	.0179	.0575	.0135
$n = 500$	NW	.3744	.1526	.3511	.1344	.3218	.1136	.3077	.1037
	LL	.0004	.0752	.0274	.0489	.0806	.0359	.1517	.0418

Chapter 4

Boundary correction using the Champernowne transformation

Inspired by Wand et al. (1991) , Buch-Larsen *et al.* (2005) showed that for heavy-tailed distributions, the tail performance of the classical kernel density estimator could be significantly improved by using a tail flattening transformation. They used modified Champernowne distribution to estimate loss distributions in insurance which is categorically heavy-tailed distributions. Sayah et.al.(2010) produce a kernel quantile estimator for heavy-tailed distributions using a modification of the Champernowne distribution.

4.1 Champernowne transformation

The original Champernowne distribution has density

$$t_{\alpha,M}(x) := \frac{\alpha x^{\alpha-1} M^{\alpha}}{\left(x^{\alpha} + M^{\alpha}\right)^2}, \quad x \geq 0,$$

The cumulative distribution function (cdf) is

$$T_{\alpha,M}(x) := \frac{x^{\alpha}}{x^{\alpha} + M^{\alpha}}, \quad x \geq 0,$$

with parameters $\alpha > 0$. M is the median of the distribution.

The Champernowne distribution converges to a Pareto distribution in the tail, whil looking more like a lognormal distribution near 0 when $\alpha > 0$. The distribution was mentioned for the first time in 1936 by D.G. Champernowne when he spoke on The Theory of Income Distribution at the Oxford Meeting of the Econometric Society.

Remark 4.1.1 *In the transformation kernel density estimation method, if we transform the data with the Champernowne cdf, the inflexible shape near* 0 *results in boundary problems. We argue that a modification of the Champernowne cdf can solve this inconvenience. The modified Champernowne cdf as proposed by Buch-Larsen et al. (*2005*) is:*

$$T_{\alpha,M,c}(x) := \frac{(x+c)^{\alpha} - c^{\alpha}}{(x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}}, \quad x \geq 0, \tag{4.1}$$

with parameters $\alpha > 0$, $M > 0$ *and* $c \geq 0$. *The associated pdf is*

$$t_{\alpha,M,c}(x) := \frac{\alpha (x+c)^{\alpha-1} \left((M+c)^{\alpha} - c^{\alpha}\right)}{\left((x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}\right)^2}, \quad x \geq 0.$$

Remark 4.1.2 *The effect of the additional parameter* c *is different for* $\alpha > 1$ *and for* $\alpha < 1$ *(see figures 4.1 and 4.2). Moreover, this distribution is of Pareto type, that is*

$$t_{\alpha,M,c}(x) \sim \frac{\alpha \left((M+c)^{\alpha} - c^{\alpha}\right)}{x^{\alpha+1}}, \quad \text{as } x \to \infty.$$

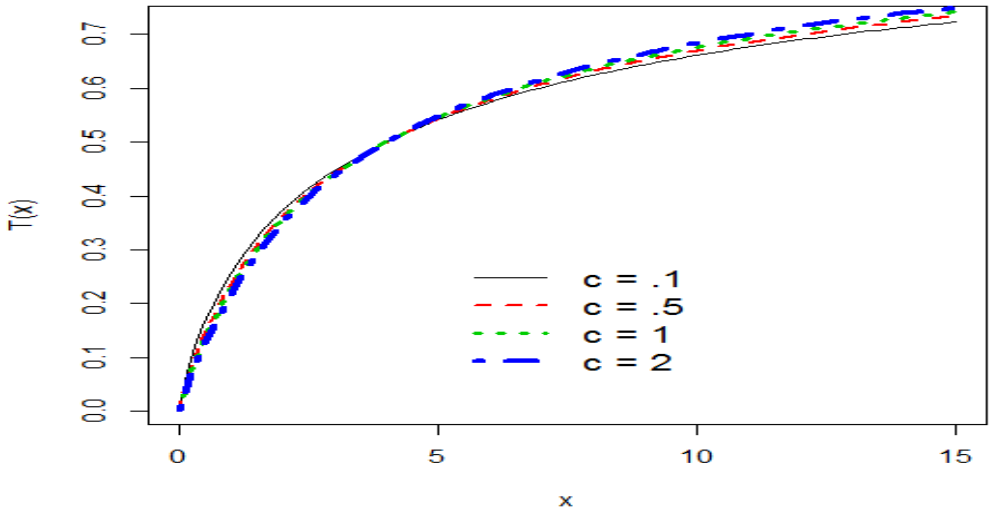

Figure 4.1: Modified Champernowne distribution function, $(M = 5, \alpha = 2)$.

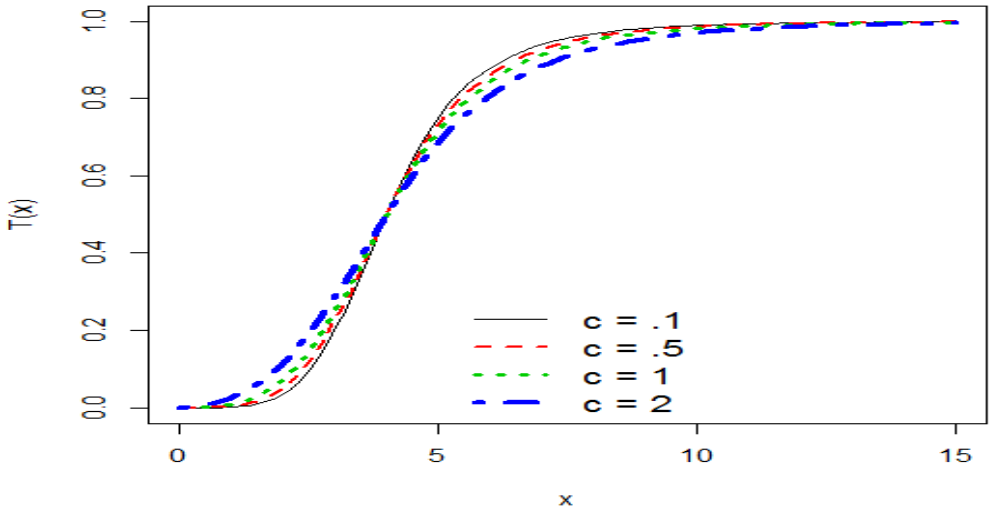

Figure 4.2: Modified Champernowne distribution function, $(M = 5, \alpha = 5)$.

4.2 Boundary correction for heavy-tailed distributions

Kernel density estimator which is of the form

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{4.2}$$

where $h := h_n$ ($h \to 0$ and $nh \to \infty$) is the bandwidth and K is an integrable smoothing kernel.

Definition 4.2.1 *Given a set of data* $X_1, X_2, ..., X_n$, *cdf* $T_{\alpha,M,c}(x)$, *modified*

Champernowne distribution, then

$$\{Z_1, ..., Z_n\} := \{T_{\alpha,M,c}(X_1), ..., T_{\alpha,M,c}(X_n)\}$$

are new variable, Z is in the interval $(0,1)$ and uniform distributed. The kernel density estimation for the transforms data is given by

$$\hat{f}_T(z) = \frac{1}{nhk_z} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{h}\right),$$

K is kernel function. The transformation kernel density estimator of $f(x)$:

$$\hat{f}_{TCh}(x) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{T_{\alpha,M,c}(x) - T_{\alpha,M,c}(X_i)}{h}\right) T'_{\alpha,M,c}(x), \tag{4.3}$$

where $T_{\alpha,M,c}(.)$ is the modified Champernowne transformation function, $T'_{\alpha,M,c}(.)$ it's derivative.

Remark 4.2.1 *Boundary correction, k_z is needed since z are in the interval $(0,1)$ so that we have to divide by the area under the kernel that lies in this interval, which defined by*

$$k_z := \int_{\max(-1,-z/h)}^{\max(1,(1-z)/h)} K(t)\, dt.$$

Theorem 4.2.1 (Buch-Larsen *et al.*, 2005) *The bias and the variance of $\hat{f}_{TCh}(x)$*

are given by

$$Bias\left(\hat{f}_{TCh}(x)\right) = \frac{1}{2}\mu_2(K)\,h^2\left(\left(\frac{f(x)}{T'_{\alpha,M,c}(x)}\right)'\frac{1}{T'_{\alpha,M,c}(x)}\right)' + o\left(h^2\right),$$
$$Var\left(\hat{f}_{TCh}(x)\right) = \frac{1}{nh}\int K^2(t)\,dt T'_{\alpha,M,c}(x)\,f(x) + o(1/nh),$$

where $\mu_2(K) := \int t^2 K(t)\,dt < \infty$.

4.3 Boundary correction in kernel quantile estimation

4.3.1 Kernel quantile estimation

The estimation of population quantiles is of great interest when a parametric form for the underlying distribution is not available. It plays an important role in both statistical and probabilistic applications, namely: the goodness-of-fit, the computation of extreme quantiles and Value-at-Risk in insurance business and financial risk management. Also, a large class of actuarial risk measures can be defined as functionals of quantiles (see, Denuit *et al.*, 2005).

Quantile estimation has been intensively used in many fields, see Azzalini (1981), Harrell and Davis (1982), Sheather and Marron (1990), Ralescu and Sun (1993), Chen and Tang (2005). Most of the existing estimators suffer from either a bias or an inefficiency for high probability levels. To solve this inconvenience,

we suggest to use the so-called transformed kernel estimate, firstly used in the density estimation context, by Devroye and Györfi (1985) for heavy-tailed observations. The idea is to transform the initial observations $\{X_1, ..., X_n\}$ into a sample $\{Z_1, ..., Z_n\} := \{T(X_1), ..., T(X_n)\}$, where T is a given function having values in $(0, 1)$. Buch-Larsen *et al.* (2005) suggested to choose T so that $T(X)$ is close to the uniform distribution. They proposed a kernel density estimation of heavy-tailed distributions based on a transformation of the original data set with a modification of the Champernowne cumulative distribution function (cdf) (see, Champernowne, 1936 and 1952). While Bolancé *et al.* (2008) proposed the Champernowne-inverse beta transformation in kernel density estimation to model insurance claims and showed that their method is preferable to other transformation density estimation approaches for distributions that are Pareto-like.

In order to correct the bias problems, Charpentier and Oulidi (2010) suggested several nonparametric quantile estimators based on the beta-kernel and applied them to transformed data. For nonparametric estimation, the bandwidth controls the balance between two considerations: bias and variance. Furthermore, the mean squared error (MSE) which is the sum of squared bias and variance, provides a composite measure of performance. Therefore, optimality in the sense of MSE is not seriously swayed by the choice of the kernel but is affected by that of the bandwidth (for more details, see Wand and Jones 1995). Sayah et al. (2010) proposed a new estimator of the quantile function, based on the modified Champernowne transformation and obtained an expression for the value of the smoothing parameter that minimizes the AMSE of the obtained estimator. They

show that, the use of this transformation in kernel estimation of quantile functions for heavy-tailed distributions improves the already existing results.

Let $X_1, X_2, ...,$ be independent and identically distributed (iid) random variables (rv's) drawn from an absolutely continuous (cdf) F with probability density function (pdf) f. For each interger n, let $X_{1,n} \leq ... \leq X_{n,n}$ denote the order statistics pertaining to the sample $X_1, ..., X_n$. We define the pth quantile $Q_X(p)$ as the left-continuous inverse of F as

$$Q_X(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad 0 < p < 1.$$

A basic estimator of $Q_X(p)$, is the sample quantile $Q_n(p) = X_{[np]+1,n}$ where $[x]$ denotes the integer part of $x \in \mathbb{R}$. Suppose that K is a pdf symmetric about 0 and $h := h_n$ is a sequence of real numbers (called bandwidth) such that $h \to 0$ as $n \to \infty$. The classical kernel quantile estimator (CKQE) was introduced by Parzen (1979) in the following form:

$$\tilde{Q}_{n,X}(p) := \sum_{i=1}^{n} X_{i,n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} K_h(x-p)\, dx, \tag{4.4}$$

where $K_h(t) := K(t/h)/h$. Yang (1985) established the asymptotic normality and the mean squared consistency of $\tilde{Q}_{n,X}(p)$, while Falk (1984) showed that the asymptotic performance of $\tilde{Q}_{n,X}(p)$ is better than that of the empirical sample quantile. Sheather and Marron (1990) gave the AMSE of $\tilde{Q}_{n,X}(p)$. For further details on kernel-based estimation, see Silverman (1986) and Wand and Jones (1995).

4.3.2 Estimation procedure

In the context of quantile estimation, if T is strictly increasing, the p^{th} quantile of $T(X)$ is equal to $T(Q_X(p))$. The idea is to transform the initial data $\{X_1, ..., X_n\}$ into $\{Z_1, ..., Z_n\}$, where $Z_i := T(X_i)$, $i = 1, ..., n$. This can be assumed to have been produced by a $(0, 1)$-uniform rv Z. Thus, (4.4) yields the transformed kernel quantile estimator

$$\hat{Q}_{n,X}(p) := T^{-1}\left(\hat{Q}_{n,Z}(p)\right),$$

where T^{-1} is the inverse of T and

$$\hat{Q}_{n,Z}(p) := \sum_{i=1}^{n} Z_{i,n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} K_h(z-p)\, dz. \tag{4.5}$$

The estimation procedure is described as follows:

1. Compute the estimates $\left(\hat{\alpha}, \hat{M}, \hat{c}\right)$ of the parameters of the modified Champernowne distribution (4.1). Notice that $T_{\alpha,M,0}(M) = 0.5$, this suggests that M can be estimated by the empirical median (see Lehmann, 1991). Then, estimate the pair (α, c) which maximizes the log-likelihood function (see, Buch-Larsen *et al.*, 2005):

$$\begin{aligned} l(\alpha, c) = n\log\alpha + n\log\left((M+c)^{\alpha} - c^{\alpha}\right) + (\alpha-1)\sum_{i=1}^{n}\log(X_i + c) \\ - 2\sum_{i=1}^{n}\log\left((X_i+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}\right). \end{aligned} \tag{4.6}$$

2. Transform the data $X_1, ..., X_n$ into $Z_1, ..., Z_n$ by

$$Z_i = T_{\hat{\alpha},\hat{M},\hat{c}}(X_i), \qquad i = 1, ..., n.$$

The resulting transformed data belong to the interval $(0, 1)$.

3. Using (4.5), calculate the kernel quantile estimator $\hat{Q}_{n,Z}(p)$ of the transformed data: $Z_1, ..., Z_n$.

4. The resulting of the original data $X_1, ..., X_n$ is given by

$$\hat{Q}_{n,X}(p) = T^{-1}_{\hat{\alpha},\hat{M},\hat{c}}\left(\hat{Q}_{n,Z}(p)\right). \tag{4.7}$$

4.3.3 Asymptotic theory and bandwidth selection

Let $X_1, ..., X_n$ be iid rv's with cdf F and pdf f. For each p in $(0, 1)$, let $\hat{Q}_{n,X}(p)$ be the transformed estimator (4.7) of $Q_X(p)$.

Theorem 4.3.1 (Sayah et al. 2010) *Assume that $Q_Z(\cdot)$ is two-times diffirentiable in a nieghborhood of $p \in (0, 1)$ with continuous second derivative. Assume further that the kernel K has compact support and fulfills:*

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0 \quad \textit{and} \quad \int t^2K(t)dt < \infty.$$

Then the bias and the variance of $\hat{Q}_{n,X}(p)$ *are respectively*

$$Bias\left(\hat{Q}_{n,X}(p)\right) = \frac{h^2}{2}\left[\left(T^{-1}\right)''\left(Q_Z(p)\right)Q_Z'^2(p) + \left(T^{-1}\right)'\left(Q_Z(p)\right)Q_Z''(p)\right] \times \mu_2(K) + o\left(h^2\right),$$

and

$$Var\left(\hat{Q}_{n,X}(p)\right) = \left(\left(T^{-1}\right)'\left(Q_Z(p)\right)Q_Z'(p)\right)^2\left(\frac{p(1-p)}{n} - \frac{h}{n}\varphi(K)\right) + o\left(\frac{h}{n}\right),$$

where $\mu_2(K) := \int t^2K(t)\,dt$, $\varphi(K) := 2\int tK(t)\left(\int_{-\infty}^{t}K(s)ds\right)dt$, Q_Z' *and* Q_Z'' *are the first and the second derivatives of* Q_Z. *The value of* h *that minimizes the* $AMSE$ *of* $\hat{Q}_{n,X}(p)$ *is*

$$h_{opt,X} := Cn^{-1/3}, \qquad C = \left(\frac{\left(\left(T^{-1}\right)'\left(Q_Z(p)\right)Q_Z'(p)\right)^2\varphi(K)}{n\Psi_{T,Q}^2(p)\,\mu_2^2(K)}\right)^{1/3}, \tag{4.8}$$

where

$$\Psi_{T,Q}(p) := \left(T^{-1}\right)''\left(Q_Z(p)\right)Q_Z'^2(p) + \left(T^{-1}\right)'\left(Q_Z(p)\right)Q_Z''(p).$$

Remark 4.3.1 *If* $Q_X'(p) > 0$, *the asymptotically optimal bandwidth for silmple estimator* $\tilde{Q}_{n,X}(p)$ *is*

$$h_{opt,C} = \left(\frac{Q_X'^2(p)\,\varphi(K)}{nQ_X''^2(p)^2\,\mu_2(K)^2}\right)^{1/3}. \tag{4.9}$$

Remark 4.3.2 *The first and the second derivatives of* Q_Z *are*

$$Q'_Z(p) = \frac{1}{g(Q_Z(p))} = \frac{T'(Q_X(p))}{f(Q_X(p))},$$

and

$$\begin{aligned} Q''_Z(p) &= \frac{-g'(Q_Z(p))}{g^3(Q_Z(p))} \\ &= -\frac{f'(Q_X(p))T'(Q_X(p)) - f(Q_X(p))T''(Q_X(p))}{f^3(Q_X(p))}. \end{aligned}$$

4.4 Examples and comparative study

For comparison perpose between $\tilde{Q}_{n,X}(p)$ and the transformed estimator $\hat{Q}_{n,X}(p)$, we consider the distributions described in Table 4.1.

Table 4.1: Examples of heavy-tailed distributions

Distribution	Density for $x > 0$
Burr $(2, 3, 1)$	$\dfrac{6x^3}{x\left(1+x^3\right)^3}$
Paralogistic $(3, 0.5)$	$\dfrac{27x^3}{x\left(1+8x^3\right)^4}$
Mixture of 70% log-normal$(0, 1)$	$0.7\dfrac{1}{\sqrt{2\pi}x}\exp\left\{-\left(\log x\right)^2/2\right\}$
and 30% Pareto$(1, 1)$	$+0.3\dfrac{x}{x\left(1+x\right)^2}$

Remark 4.4.1 *Note that, the mixture of log-normal and Pareto distributions was*

*previously used in Buch-Larsen et al. (*2005*) and Charpentier and Oulidi (*2010*). The performance of the estimators is measured by the AMSE criteria:*

$$AMSE := \frac{1}{N}\sum_{s=1}^{N}\left(\hat{Q}_{n,X}^{(s)}(p) - Q\left(p\right)\right)^2,$$

where $\hat{Q}_{n,X}^{(s)}(p)$ *is the quantile correponding to the* s^{th} *simulated sample* $\left\{X_1^{(s)}, ..., X_n^{(s)}\right\}$ *and* N *is the number of replications. The algorithm used to estimate the quantile function with level* $p \in (0,1)$ *is described as follows:*

1. Generate a sample $X_1, ..., X_n$ of size n.

2. Estimate M by the empirical median $\hat{M}$, solution of $T_{\alpha,M,0}\left(M\right) = 0.5$.

3. Estimate the pair (α, c) maximizing the log-likelihood function (4.6).

4. Transform $X_1, ..., X_n$ into $Z_1, ..., Z_n$:

$$Z_i = T_{\hat{\alpha},\hat{M},\hat{c}}\left(X_i\right), \qquad i = 1, ..., n.$$

5. Compute the estimate $\hat{Q}_{n,Z}(p)$ by choosing the Epanechnikov kernel:

$$K(t) = \tfrac{3}{4}\left(1 - t^2\right)\mathbf{1}_{(|t|<1)}.$$

1. The resulting transformed quantile estimator of the original data is

$$\hat{Q}_{n,X}(p) = T^{-1}_{\hat{\alpha},\hat{M},\hat{c}}\left(\hat{Q}_{n,Z}(p)\right).$$

2. The classical quantile estimator is directly obtained from the original data, where the bandwidth $h := h_{opt,C}$ is such as in (4.9).

Let the sample size be 200 and compute both the transformed (TQ) and the classical (CQ) quantile estimators for probability levels $p \in \{.05, .10, .25, .50, .75, .90, .95\}$. All results are calculated by averaging over 200 simulation runs.. The results are summarized in Tables 4.2–4.5 where we see that the transformed estimator is better than the classical one for high probability levels $p \in \{.75, .90, .95\}$. Table 4.4 is based on the mixture 30% log-normal and 70% Pareto distributions. Both estimators are equal for $p \in \{.05, .10, .25, .50\}$.

Table 4.2: Burr distribution, 200 samples of size 200.

p	0.05	0.1	0.25	0.5	0.75	0.9	0.95
$Q(p)$	0.2962	0.3782	0.5368	0.7454	1.0000	1.2931	1.5143
$TKQE$	0.2966	0.3728	0.5345	0.7480	0.9946	1.2928	1.5150
$CKQE$	0.2988	0.3741	0.5345	0.7503	0.9852	0.5464	0.0367

Table 4.3: Paralogistic distribution, 200 samples of size 200.

p	0.05	0.1	0.25	0.5	0.75	0.9	0.95
$Q(p)$	0.1075	0.1551	0.2622	0.4291	0.6667	0.9803	1.2422
$TKQE$	0.7983	0.1278	0.2526	0.4263	0.6705	0.9676	1.1626
$CKQE$	0.1088	0.1547	0.2641	0.4330	0.7024	0.6079	0.4421

Table 4.4: Mixtures (rho= 0.3) distribution, 200 samples of size 200.

p	0.05	0.1	0.25	0.5	0.75	0.9	0.95
$Q(p)$	0.0948	0.1611	0.3862	1.0000	2.6889	7.3807	14.8541
$TKQE$	0.2380	0.3391	0.6213	1.2560	2.7743	7.2812	15.2085
$CKQE$	0.2350	0.3380	0.6273	1.3246	16.4845	28.9263	21.5483

Table 4.5: Mixtures (rho= 0.7) distribution, 200 samples of size 200.

p	0.05	0.1	0.25	0.5	0.75	0.9	0.95
$Q(p)$	0.1509	0.2277	0.4566	1.0000	2.2741	5.2216	9.3262
$TKQE$	0.2987	0.4200	0.7230	1.3483	2.5389	5.1070	8.4522
$CKQE$	0.3239	0.3981	0.7293	1.3805	2.6514	6.6738	29.6183

Next, we sample, 200 times, from the four distributions sets of sizes 50, 100 and compute the transformed and the classical quantile estimators with their $AMSE's$ for levels $p \in \{.90, .95\}$. The respective results are given in Tables 4.6 and 4.7. It is clear that, for large probability levels, the transformation-based approach gives results of higher quality with respect to the classical procedure. Note that, under classical estimation, some $AMSE's$ are seriously bad when samples come from mixture distributions, especially when 70% of Pareto distribution is considered. The same remark can be observed in Charpentier and Oulidi (2010) (see their table's 13-18 pages 52–53).

Table 4.6: Classical and transformed pth quantile estimators (p= .9)

Distribution			Burr	Paralogistic	ρlog normal+$(1-\rho)$Pareto	
					$\rho = 30\%$	$\rho = 70\%$
$p = .90$		$Q(p)$	1.2931	0.9803	7.3807	5.2216
$n = 50$	$value$	TQ	1.2941	0.9796	7.8530	5.2474
		CQ	0.3864	0.4683	10.668	9.5797
	$AMSE$	TQ	0.0201	0.0277	15.545	3.2335
		CQ	0.8230	0.2655	298.59	179.86
$n = 100$	$value$	TQ	1.2985	0.9819	7.3484	5.1982
		CQ	0.4690	0.5341	12.540	11.3100
	$AMSE$	TQ	0.0084	0.0113	5.3956	1.5319
		CQ	0.6798	0.2012	352.99	324.23

Table 4.7: Classical and transformed pth quantile estimators (p=.95)

Distribution			Burr	Paralogistic	ρlog normal+$(1-\rho)$Pareto	
					$\rho = 30\%$	$\rho = 70\%$
$p = .95$		$Q(p)$	1.5143	1.2422	14.8541	9.3262
$n = 50$	$value$	TQ	1.5506	1.0945	16.6389	9.0187
		CQ	0.0232	0.3396	12.2710	12.0748
	$AMSE$	TQ	0.0443	0.0751	165.422	19.7341
		CQ	2.2232	0.8165	1025.83	466.674
$n = 100$	$value$	TQ	1.5332	1.1352	14.8011	8.6076
		CQ	0.0291	0.3889	16.0566	17.5289
	$AMSE$	TQ	0.0211	0.0702	42.2056	4.8286
		CQ	2.2057	0.7294	1129.14	669.036

Conclusion

Kernel estimators are not consistent near the finite end points of their supports. In other words, these effects seriously affect the performance of these estimators. In this thesis, we have studied the boundary effect in the kernel density and regression estimations. We have mentioned some methods for correcting this effect. Both density and regression functions are considered and their statistical properties are given.

For heavy-tailed distributions, bias or inefficiency problems may occur in the classical kernel quantile estimation when considering high probability levels. To solved this incontinence, the use of the transformation data modified based on the Champernowne distribution is recommended.

The variables studied are fully observed, the case of incomplete data: truncated or censored is interesting for future study. Note also that the density function is often encountered in the estimation of the distribution function, quantile and conditional densities. Therefore, the study of the boundary effect in the estimation of these functions offers good perspectives.

Bibliography

[1] Abramson, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. The annals of Statistics, 1217-1223.

[2] Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika, 68(1), 326-328.

[3] Berlinet, A. (1993). Hierarchies of higher order kernels. Probability theory and related fields, 94(4), 489-504.

[4] Bierens, H. J. (1987). Kernel estimators of regression functions. In Advances in econometrics: Fifth world congress (Vol. 1, pp. 99-144).

[5] Bolancé, C., Guillén, M., & Nielsen, J. P. (2008). Inverse Beta transformation in kernel density estimation. Statistics & Probability Letters, 78(13), 1757-1764.

[6] Breiman, L., Meisel, W., & Purcell, E. (1977). Variable kernel estimates of multivariate densities. Technometrics, 19(2), 135-144.

[7] Buch-Larsen, T., Nielsen, J. P., Guillén, M., & Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. Statistics, 39(6), 503-516.

[8] Champernowne, D. G. (1936). The Oxford meeting, September 25–29, by Brown P. Econometrica, 5, 361-383.

[9] Champernowne, D. G. (1952). The graduation of income distributions. Econometrica: Journal of the Econometric Society, 591-615.

[10] Charpentier, A., & Oulidi, A. (2010). Beta kernel quantile estimators of heavy-tailed loss distributions. Statistics and computing, 20(1), 35-55.

[11] Cheng, M. Y. (1997). Boundary aware estimators of integrated density derivative products. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59(1), 191-203.

[12] Chen, S. X., & Tang, C. Y. (2005). Nonparametric inference of value-at-risk for dependent financial returns. Journal of Financial Econometrics, 3(2), 227-255.

[13] Cheng, M. Y. (2006). Choice of the bandwidth ratio in Rice's boundary modification. Journal of the Chinese Statistical Association, 44, 235-251.

[14] Cheng, K. F., & Lin, P. E. (1981). Nonparametric estimation of a regression function. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 57(2), 223-233.

[15] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association, 74(368), 829-836.

[16] Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association, 83(403), 596-610.

[17] Cleveland, W. S., & Loader, C. (1996). Smoothing by local regression: Principles and methods. In Statistical theory and computational aspects of smoothing (pp. 10-49). Physica-Verlag HD.

[18] Cline, D. B. H., & Hart, J. D. (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives. Statistics: A Journal of Theoretical and Applied Statistics, 22(1), 69-84.

[19] Cowling, A., & Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), 551-563.

[20] Dai, J., & Sperlich, S. (2010). Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation. Computational Statistics & Data Analysis, 54(11), 2487-2497.

[21] Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2006). Actuarial theory for dependent risks: measures, orders and models. John Wiley & Sons.

[22] Devroye, L., & Gyorfi, L. (1985). Nonparametric density estimation: the L1 view (Vol. 119). John Wiley & Sons Incorporated.

[23] Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153-158.

[24] Eubank, R. L. (1988). Spline smoothing and nonparametric regression.

[25] Eubank, R. L., & Speckman, P. L. (1991). A bias reduction theorem with applications in nonparametric regression. Scandinavian Journal of Statistics, 211-222.

[26] Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. The Annals of Statistics, 261-268.

[27] Fan, J. (1992b). Design-adaptive nonparametric regression. Journal of the American statistical Association, 87(420), 998-1004..

[28] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. The Annals of Statistics, 196-216.

[29] Fan, J., & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. The Annals of Statistics, 2008-2036.

[30] Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications: monographs on statistics and applied probability 66 (Vol. 66). CRC Press.

[31] Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions (pp. 23-68). Springer Berlin Heidelberg.

[32] Gasser, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. Scandinavian Journal of Statistics, 171-185.

[33] Gasser, T., Muller, H. G., & Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. Journal of the Royal Statistical Society. Series B (Methodological), 238-252.

[34] Gray, H. L., & Schucany, W. R. (1972). The generalized jackknife statistic. New York: Marcel Dekker.

[35] Hall, P., & Marron, J. S. (1988). Variable window width kernel estimates of probability densities. Probability Theory and Related Fields, 80(1), 37-49.

[36] Hall, P. (1990). On the bias of variable bandwidth curve estimators. Biometrika, 77(3), 529-535.

[37] Hall, P., & Park, B. U. (2002). New methods for bias correction at endpoints and boundaries. Annals of Statistics, 1460-1479.

[38] Härdle, W. (1990). Applied nonparametric regression (Vol. 27). Cambridge: Cambridge university press.

[39] Härdle, W., & Vieu, P. (1992). Kernel regression smoothing of time series. Journal of Time Series Analysis, 13(3), 209-232.

[40] Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. Biometrika, 69(3), 635-640.

[41] Jones, M. C. (1993). Simple boundary correction for kernel density estimation. Statistics and Computing, 3(3), 135-146.

[42] Jones, M. C. (1990). Variable kernel density estimates and variable kernel density estimates. Australian Journal of Statistics, 32(3), 361-371.

[43] Jones, M. C., & Foster, P. J. (1993). Generalized jackknifing and higher order kernels. Journal of Nonparametric Statistics, 3(1), 81-94.

[44] Karunamuni, R. J., & Alberts, T. (2003). A locally adaptive generalized reflection method of boundary correction in kernel density estimation. Technical report.

[45] Karunamuni, R. J., & Alberts, T. (2005). On boundary correction in kernel density estimation. Statistical Methodology, 2(3), 191-212.

[46] Karunamuni, R. J., & Alberts, T. (2006). A locally adaptive transformation method of boundary correction in kernel density estimation. Journal of Statistical Planning and Inference, 136(9), 2936-2960.

[47] Lejeune, M. (1985). Estimation non-paramétrique par noyaux: régression polynomiale mobile. Revue de Statistique Appliquée, 33(3), 43-67.

[48] Lehmann, E. L. (1991). Theory of Point Estimation Wadsworth. Monterey, CA.

[49] Marron, J. S., & Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), 653-671.

[50] Müller, H. G. (1984). Boundary effects in nonparametric curve estimation models. In Compstat 1984 (pp. 84-89). Physica-Verlag HD.

[51] Müller, H. G. (1984a). Smooth optimum kernel estimators of densities, regression curves and modes. The Annals of Statistics, 766-774.

[52] Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. Journal of the American Statistical Association, 82(397), 231-238.

[53] Müller, H. G. (1988). Nonparametric Analysis of Longitudinal Data (Lecture Notes in Statistics 46).

[54] Müller, H. G. (1991). Smooth optimum kernel estimators near endpoints. Biometrika, 78(3), 521-530.

[55] Müller, H. G. (1993). On the boundary kernel method for non-parametric curve estimation near endpoints. Scandinavian Journal of Statistics, 313-328.

[56] Müller, H. G. (1993). [Local Regression: Automatic Kernel Carpentry]: Comment. Statistical Science, 134-139.

[57] Müller, H. G., & Wang, J. L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics, 61-76.

[58] Nadaraya, E. A. (1964). On estimating regression. Theory of Probability & Its Applications, 9(1), 141-142.

[59] Parzen, E. (1979). Nonparametric statistical data modeling. Journal of the American statistical association, 74(365), 105-121.

[60] Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 1065-1076.

[61] Priestley, M. B., & Chao, M. T. (1972). Non-parametric function fitting. Journal of the Royal Statistical Society. Series B (Methodological), 385-392.

[62] Ralescu, S. S., & Sun, S. (1993). Necessary and sufficient conditions for the asymptotic normality of perturbed sample quantiles. Journal of statistical planning and inference, 35(1), 55-64.

[63] Rice, J (1984). Boundary modification for kernel regression. Communications in Statistics-Theory and Methods, 13(7), 893-900.

[64] Rice, J., & Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. The annals of Statistics, 141-156.

[65] Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions. Scandinavian Journal of Statistics, 116-119.

[66] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3), 832-837.

[67] Sayah, A., Yahia, D. & Necir, A. (2010). Champernowne transformation in kernel quantile estimation for heavy-tailed distributions. Afrika Statistika, 5(1).

[68] Schucany, W. R., & Sommers, J. P. (1977). Improvement of kernel type density estimators. Journal of the American Statistical Association, 72(358), 420-423.

[69] Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics-Theory and methods, 14(5), 1123-1136.

[70] Schucany, W. R., Gray, H. L., & Owen, D. B. (1971). On bias reduction in estimation. Journal of the American Statistical Association, 66(335), 524-533.

[71] Serfling, R. J. (2009). Approximation theorems of mathematical statistics (Vol. 162). John Wiley & Sons.

[72] Sheather, S. J., & Marron, J. S. (1990). Kernel quantile estimators. Journal of the American Statistical Association, 85(410), 410-416.

[73] Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. The Annals of Statistics, 898-916.

[74] Silverman, B. W. (1986). Density Estimation London. UK: Chapman and Hall.

[75] Stone, C. J. (1977). Consistent nonparametric regression. The annals of statistics, 595-620.

[76] Tapia, R. A., & Thompson, J. R. (1978). Nonparametric probability density estimation.

[77] Wand, M.P. & Jones, M.C., (1995). Kernel Smoothing, London: Chapman and Hall.

[78] Wand, M. P., Marron, J. S., & Ruppert, D. (1991). Transformations in density estimation. Journal of the American Statistical Association, 86(414), 343-353.

[79] Wand, M. P., & Schucany, W. R. (1990). Gaussian-based kernels. Canadian Journal of Statistics, 18(3), 197-204.

[80] Watson, G. S. (1964). Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, 359-372.

[81] Yang, S. S. (1985). A smooth nonparametric estimator of a quantile function. Journal of the American Statistical Association, 80(392), 1004-1011.

[82] Zhang, S., & Karunamuni, R. J. (1998). On kernel density estimation near endpoints. Journal of Statistical Planning and Inference, 70(2), 301-316.

[83] Zhang, S., Karunamuni, R. J., & Jones, M. C. (1999). An improved estimator of the density function at the boundary. Journal of the American Statistical Association, 94(448), 1231-1240.

[84] Zhang, S., & Karunamuni, R. J. (2000). On nonparametric density estimation at the boundary. Journal of nonparametric statistics, 12(2), 197-221.

Symbols and Notations

We list the notations that will be used in this thesis

X	predictor variable
Y	variable of interest
$E(X)$	expectation or mean of X
F	distribution function
f	marginal density of X
$c \wedge 1$	$\min(c, 1)$
$f(0+)$	f right continuous at the point 0
$\hat{f}$	estimator of f
h	bandwidth
iid	independent and identically distributed
K	kernel function
$f^{(j)}$	the j^{th}-derivative
f', f'', f'''	the first, the second and the third derivatives of f.

$\hat{f}_g$	the reflection and transformation density estimator
$\hat{f}_{CN}$	Cut and Normalized density estimator
$\hat{f}_{grt}$	the generalized reflection and transformation density estimator
$\hat{f}_{refl}$	the reflection density estimator
$\hat{f}_{Tag}$	the transformation density estimator
$\hat{f}_{TAK}$	Translation in the Argument of the Kernel density
$\hat{f}_T$	The kernel density estimation for the transforms data
$\hat{f}_{TCh}$	Transformation Champernowne kernel density estimator of $f(x)$
$\bar{f}_{\alpha,h}$	Rice's boundary modification density estimator
g	transformation function
g^{-1}	the inverse function of g
$m(.)$	regression curve of Y on X
$\hat{m}(.)$	estimator of $m(.)$
$\hat{m}_h$	classical estimator
$\hat{m}_l$	local linear regression estimator
$\hat{m}_n$	kernel regression estimator of Gasser and Müller 1979
$\tilde{m}_n$	the generalized reflection and transformation regression estimator
$\hat{m}_{jh}^J$	the generalized Jackknifing regression estimator
$\hat{m}_{CN}$	Cut and Normalized kernel regression estimator
$\bar{m}_{\alpha,h}$	the Rice's modification kernel regression estimator

rv	random variable
MSE	Mean Squared Error
$AMSE$	Asymptotic Mean Squared Error
$AMISE$	Asymptotic Mean Integrated Squared Error
Q	quantile function
$Q_X(p)$	the pth quantile
$\tilde{Q}_{n,X}(p)$	The classical kernel quantile estimator
$\hat{Q}_{n,X}(p)$	the transformed estimator (CKQE) of $Q_X(p)$
$t_{\alpha,M}$	The original Champernowne density
$t_{\alpha,M,c}(x)$	The associated pdf
$T_{\alpha,M}$	The cumulative distribution function (cdf)
$T_{\alpha,M,c}(x)$	The modified Champernowne cdf
$\{(X_i, Y_i)\}_{i=1,..,n}$	sample of n observations
1_A	indicator function of set A
$\sigma^2(Y\|X=x)$	conditional variance of Y given $X=x$
$o(.)$	$f(x) = o(g(x))$ as $x \to x_0$: $f(x)/g(x)$ as $x \to x_0$
$O(.)$	$f(x) = O(g(x))$ as $x \to x_0$: $\exists M > 0, \|f(x)/g(x)\| \leq M$ as $x \to x_0$

$[0, \infty)$	positive interval
$Var\,(X)$	variance of X
Epa	Epanechnikov
biw	biweight
cdf	cumulative distribution function
CKQE	The classical kernel quantile estimator
CN	Cut-and-Normalized
CQ	the classical quantile estimator
CV	Cross Validation
e.g.	for example
gauss	gaussian
grt	generalized reflection and transformation
GTR	Generalized Transformation and Reflection
i.e.	that is to say
LL	Local Linear
opt	optimal
NW	Nadaraya and Watson
pdf	probability density function
refl	reflection
Tag	Tanslation in the argument of the kernel
TAK	Translation in the Argument of the Kernel

ملخـــــص

في هذه الرسالة نقوم بدراسة بعض طرق تصحيح الآثار الحدية لمقدرات النواة لدوال الكثافة والانحدار وخصائصهم الإحصائية. تظهر مقدرات النواة مشاكل في التقارب على مستوى حدود حاملها. بعبارة أخرى، فإن هذه الآثار الحدية تؤثر تأثيرا خطيرا على أداء هذه المقدرات .لتصحيح هذه الآثار، تم اقتراح وطرح مجموعة متنوعة من الأساليب ، الأكثر استخداما وعلى نطاق واسع هي الانعكاس ، التحويل والخطية المحلية... في هذه المذكرة قمنا بمزج طريقتي الانعكاس والتحويل من أجل تقديم طريقة جديدة وشاملة لتصحيح الآثار الحدية لدالة الانحدار. مشكلة الآثار الحدية لمقدرات النواة لدوال الكثافة والربيعيات في حالة التوزيعات ذات الذيل الثقيل تمت دراستها أيضا.

Résumé

Dans cette thèse nous étudions certaines méthodes de correction des effets de bord des estimateurs à noyaux des fonctions de densité et de la régression et leurs propriétés statistiques. Les estimateurs à noyau présentent des problèmes de convergence aux bords de leurs supports. En d'autre termes, ces effets de bord affectent sérieusement les performances de ces estimateurs. Pour corrigé ces effets de bord, une variété de méthodes ont été développées dans la littérature, la plus largement utilisée est la réflexion, la transformation et la linéaire locale... Dans cette thèse, nous combinons les méthodes de transformation et de réflexion, pour introduire une nouvelle méthode générale de correction de l'effet de bord lors de l'estimation de la régression. Le problème de l'effet de bord des estimateurs à noyau des fonctions de densité ou des quantiles en cas de distribution à queue lourde est également étudié.

Abstract

In this thesis we study some boundary correction methods for kernel estimators of both density and regression functions and their statistical properties. Kernel estimators are not consistent near the finite end points of their supports. In other words, these effects seriously affect the performance of these estimators. To remove the boundary effects, a variety of methods have been developed in the literature, the most widely used is the reflection, the transformation and the local linear methods... In this thesis, we combine the transformation and the reflection methods in order to introduce a new general method of boundary correction when estimating the regression function. Boundary problems for Kernel density or quantile functions estimators in heavy-tailed case are also studied.