2004

Smoothing Parameter Selection In Nonparametric Functional Estimation

Mohamed Amezziane
University of Central Florida

SMOOTHING PARAMETER SELECTION IN NONPARAMETRIC FUNCTIONAL ESTIMATION

by

MOHAMED AMEZZIANE
M.S. University of Central Florida, 2000

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Mathematics
in the College of Arts and Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2004

ABSTRACT

This study intends to build up new techniques for how to obtain completely data-driven choices of the smoothing parameter in functional estimation, within the confines of minimal assumptions. The focus of the study will be within the framework of the estimation of the distribution function, the density function and their multivariable extensions along with some of their functionals such as the location and the integrated squared derivatives.

ACKNOWLEDGEMENT

I would like to express my deep regard and profound indebtedness to Dr. Ibrahim Ahmad, under whose guidance I have worked for the last five years and whose constant support and skilful insight have been decidedly instrumental and profoundly fundamental in the completion of this study.

I am extremely grateful to Dr. James Schott for his continuous assistance and kind advice throughout this work.

I also take this privilege to express my heartfelt appreciation to Dr. Ram Mohapatra, Dr. Marianna Pensky and Dr. Gary Richardson for their careful analysis and valuable suggestions.

I also wish to extend my sincere thanks to the professors in the Department of Mathematics and the Department of Statistics and Actuarial Science for their invaluable education.

And finally I would like to both remember and thank everybody who provided support, encouragement, inspiration as well as the means for making this endeavour a reality.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

1.1. Summary

The last two decades have seen a remarkable amount of techniques that aim to find an adequate bandwidth selector or smoothing parameter for kernel curve estimation problems. Although this area enjoys a very rich literature, there still is much room for development. The purpose of the current work is to improve and enhance existing methods as well as to present new smoothing parameter estimating techniques superior (in term of rates of convergence) to earlier ones and also to address situations that are largely left untreated such as the multivariate cases.

1.2. General Kernel Density Estimation

In his fundamental paper, Rosenblatt (1956), defined the kernel density estimator as:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - X_i}{h}),$$

where the $X_1, \ldots, X_n$ form a random sample, k is a second order (has a second moment) symmetric kernel and h is the bandwidth, taken to depend on n such that $h \to 0$ as $n \to \infty$.

Several error criteria are used to evaluate the performance of such estimate. Among the common ones we may find:

- The integrated squared error or the L_2-distance and its expected value:

$$ISE(h) = \int (\hat{f}(x) - f(x))^2\, dx$$

$$MISE(h) = E\int (\hat{f}(x) - f(x))^2 dx$$

Sometimes it is more convenient to use the asymptotic mean integrated squared error:

$$AMISE = \frac{R(k)}{nh} + \frac{1}{4}\mu_2^2 R(f'')h^4 ,$$

where $R(g) = \int g^2(x)dx$, for some real valued function *g*.

- The integrated absolute error or L_1-distance and its expected value:

$$IAE(\hat{f}, f) = \int \left|\hat{f}(x) - f(x)\right| dx$$
$$MIAE(\hat{f}, f) = E\int \left|\hat{f}(x) - f(x)\right| dx$$

- The Hellinger distance and its expectation:

$$HD(h) = \{\int (\hat{f}^{\frac{1}{r}}(x) - f^{\frac{1}{r}}(x))^r dx\}^{\frac{1}{r}}$$
$$MHD(h) = E\{\int (\hat{f}^{\frac{1}{r}}(x) - f^{\frac{1}{r}}(x))^r dx\}^{\frac{1}{r}}$$

- The Kernel Contrast: Newly proposed by Ahmad and Ran (2004), this criterion, which is compatible with any kind of distance, leads to direct estimation of its mean. In the case of L_2-distance, the mean integrated squared contrast is defined as:

$$MISC(h) = E\int (\sum_{j=1}^{p} \alpha_j \hat{f}_j(x))^2 dx ,$$

where the α_i's are constants whose sum equals zero, and $\hat{f}_j$'s are kernel density estimates based on p different kernels and a common bandwidth. The above risk function can be estimated by the obviously unbiased estimator:

$$ISC(h) = \int (\sum_{j=1}^{p} \alpha_j \hat{f}_j(x))^2 dx$$

Though the outcome of the estimation clearly depend on which criterion is used, the choice of the kernel on the other hand is not crucial when the sample is i.i.d but may become critical for other sampling schemes (c.f. Wand and Jones 1995).

The following are some of main L_2-based bandwidth selection methodologies. For more details on some methodologies based on either L_1 or Hellinger distances, we refer to Devroye and Györfi (1985), and Kanazawa (1993) and Ahmad and Mugdadi (2005).

1.2.1. Least Squares Cross Validation:

A commonly used goodness-of-fit criterion for the kernel density estimate is the integrated squared error (ISE):

$$\int (\hat{f}(y) - f(y))^2 dy = \int \hat{f}(y)^2 dy - 2\int f(y)\hat{f}(y)dy + \int f(y)^2 dy$$

The last term on the right hand side does not involve h, the first term can be calculated based on the data, and the second term can be estimated from the data by:

$$-\frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i),$$

where $\hat{f}_{-i}(y)$ denotes “leave-one-out “ kernel density estimator, constructed from the data with the observation X_i removed. The obtained expression is called the least squares cross validation (LSCV) of $\hat{f}$.

Since $E\left[LSCV\left(h\right)\right] = MISE(h) - R\left(f\right)$, the value of h that minimizes LSCV(h), usually denoted by h_{LSCV}, provides an estimate of the optimal bandwidth.

This idea was first presented by Rudemo (1982) and Bowman (1984) in an attempt to find a data-based expression that “estimates” the stochastic terms in the ISE. Hall (1983) and Stone (1984) showed that this procedure leads to a consistent and asymptotically optimal smoothing parameter. However, although its asymptotic distribution is centered near h_{MISE}, h_{LSCV} has a relative error that suffers from a very slow order of convergence, which occurs at $n^{-1/10}$ as shown by Hall and Marron (1987). This obviously reflects on the variance of the asymptotic distribution.

Various simulation studies show that the distribution of h_{LSCV} has an intolerable high left tail (Jones et. al. 1996), which explains the tendency of the least squares cross validation estimates to undersmooth the estimating curves. For this reason, it is advisable to choose h_{LSCV} to be the largest local minimizer rather than the global minimizer of LSCV(h)(Hall and Marron 1991).

1.2.2. Biased Cross Validation:

This method (Scott and Terrell 1987) intends to estimate the AMISE of $\hat{f}$ as follows:

$$BCV(h) = \frac{R(k)}{nh} + \frac{\mu_2^2}{2n^2 h} \sum_{i<j} k'' * k''\left(\frac{X_i - X_j}{h}\right),$$

where R(k) and μ_2^2 are respectively, the integrated square and second moment of the kernel k(**.**).

The rate of convergence of h_{BCV}, the minimizer of BCV(h) is the same as that of h_{LSCV}, though it benefits from a noticeably less variable distribution than h_{LSCV}. But h_{BCV} tends to overestimate h_{MISE}, leading in its turn to a quite often oversmoothed estimate of f. Therefore, it would be worthwhile to choose h_{BCV} to be the local minimizer of BCV(h). Another drawback of this method is its poor small sample behavior due obviously to BCV(h) being an estimate of AMISE.

1.2.3. Plug-In Techniques:

There are many variations of the Plug-In methodology in kernel density estimation. Some go back to the early 1970's. The common idea is to plug an estimate of f'' in the criterion of interest. Woodroofe (1970) showed that $\hat{f}''$ is a consistent estimate, in squared mean, of f and its other asymptotic properties were studied by Ahmad (1976).

In their survey Jones, Marron and Sheather (1996) argue that $\hat{f}''$ "is asymptotically inconsistent for f" and instead present the commonly used refinement of the Plug-In due to Sheather and

Jones (1991). This version of the Solve-She-Equation Plug-In consists of finding h_{STE}, the solution of the fixed-point equation:

$$h = \left[\frac{R(k)}{R(\hat{f}''_{g(h)})\mu_2^2 n} \right]^{1/5}$$

where *g(h) = A(f'',f''') B(k)h*$^{5/7}$*,* for some functionals A and B, and f'' and f''' are replaced by their "inconsistent" kernel estimates. The starting *h* is chosen to correspond to a 'good' parametric guess.

This h_{STE} has a much faster rate of convergence (*5/14)* than cross validation estimators and if subjected to improvements (Chiu 1992, Engel, Herrmann, and Gasser1994 etc..) such as higher stage functional pilot estimation (Park and Marron 1992) , would enjoy rates as fast as *1/2* making the bandwidth selectors less variable than one obtainable via LSCV.

However, excellent asymptotics do not always agree with simulation studies and it is often the case when Plug-In techniques are applied. Except for what is described by Jones, Marron and Sheather (1996) "easy-to-estimate densities", the h_{PI} tends to perform as poorly as h_{BCV} for multimodal and relatively rough densities. It also behaves as badly for small samples. Examples provided by Loader (1999) show that in general the LSCV estimates are the most reliable.

1.2.4. Kernel Contrast Density Estimation:

The idea behind this methodology originates from the fact that all criteria previously considered involve the unknown, to be estimated density *f*. Therefore, instead of the usual measures of deviations Ahmad and Ran (2004) considered the following criterion:

$$ISC(h) = \int (\sum_{j=1}^{p} \alpha_j \hat{f}_j(x))^2 dx,$$

where the α_i's are constants whose sum equals zero, and $\hat{f}_j$'s are kernel density estimates based on p different kernels.

$\hat{h}_{con}$ the minimizer of the above expression is shown to converge in probability to h_{MISE} / Λ where Λ is known constant.

The rate of convergence of $\hat{h}_{con}$ is *1/5*, which is the best that can be obtained for density estimation under a two derivatives assumption (Stone 1980). The small sample simulations displayed in the same work shows that the estimator has a small variability even for multimodal densities. Though it shows some tendency for oversmoothing, this method still seems to catch humps and troughs in small-sampled mixtures.

1.3. Kernel Distribution Function Estimation

Nadaraya (1964) defined the kernel distribution function estimate by:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) ,$$

where the $X_1, \ldots, X_n$ form a random sample, K is the distribution function of a second order kernel and h is the bandwidth.

The above estimate is usually evaluated through two measures of performance:

$$ISE(h) = \int (\hat{F}(x) - F(x))^2 W(x) dF(x) \text{, and}$$

$$MISE(h) = E\int (\hat{F}(x) - F(x))^2 W(x) dF(x) ,$$

where W is a nonnegative weight function.

The unweighted version corresponds to taking $W(x) = \frac{1}{f(x)}$.

Unlike kernel density estimation, the kernel distribution function estimation did not receive much attention of the researchers, and the few existing bandwidth techniques require a large deal of improvement.

1.3.1. Leave-One-Out and Leave-None-Out Weighted Cross Validation:

Sarda (1993) presented two estimators for the MISE(h), namely the leave-one-out and leave-none-out criteria:

$$L1O(h) = n^{-1}\sum_{i=1}^{n} [\hat{F}_{-i}(X_i) - F_n(X_i)]^2 W(X_i)$$

$$LNO(h) = n^{-1}\sum_{i=1}^{n} [\hat{F}(X_i) - F_n(X_i)]^2 W(X_i)$$

Where $\hat{F}_{-i}(x)$ is the kernel estimator computed by leaving X_i out and $F_n(x)$ is the empirical distribution function.

Sarda argues that while the latter produce a very small bandwidth, the former leads to an asymptotically optimal smoothing parameter. Altman and Léger (1995) proved that the two quadratic measures of error are asymptotically equivalent. They also went on showing, both theoretically and empirically (for samples of size as large 1000), that neither method work and therefore challenging Sarda's optimality result.

1.3.2. Plug-In Estimation:

Altman and Léger (1995) introduce a Plug-In estimator for the distribution function analogous to Sheather and Jones's with a pilot bandwidth equal to $n^{-0.3}$. Though asymptotic analysis of the estimator was not done, the authors state that, based on simulations, their estimator seems to display a good behavior when the weight function is the indicator function of the interval [0.1,3].

1.3.3. Unweighted Cross Validation:

To estimate the unweighted mean integrated squared error, Hall, Bowman, and Prvan (1998) recommend the use of:

$$CV(h) = \frac{1}{n}\int \{I(x - X_i) - \hat{F}_{-i}(x)\}^2\, dx \,,$$

where I(.) is the indicator function of the positive half line.

The above expression, added to a vanishing constant, is proved to converge almost surely to the MISE(h), thus making h_{CV} converging with the same mode to h_{MISE}.

Bowman, Hall and Prvan (1998) assert that the leave-one-out cross-validatory estimate they suggest, $CV(h) = \frac{1}{n}\sum_{i=1}^{n}\int \{I(x - X_i) - \hat{F}_{-i}(x)\}^2 dx$, has nice theoretical properties among which the following asymptotic results:

Theorem: For each $\delta, \varepsilon, C > 0$,

$$CV(h) - \frac{1}{n}\sum_{i=1}^{n}\int [I(x - X_i) - F(x)]^2 dx + \int \{[F_n(x) - F(x)]^2 - E[F_n(x) - F(x)]^2\} dx$$

$$= MISE(h) + O\{(n^{-1/2}h^3 + n^{-1}h^{3/2} + n^{-3/2})n^{\delta}\} \quad (1.1)$$

with probability one, uniformly in $h \in [0, Cn^{-\varepsilon}]$ as $n \to \infty$.

The authors go on arguing that since the second and third terms in the left hand side are independent of the smoothing *h,* it would only be natural that minimization of CV*(h)* produces a bandwidth h_{CV} that is asymptotically equivalent to the optimal h_{MISE}, as stated next:

Corollary: For every $C > 0$ and every $\varepsilon \in (0, 1/3)$, if $h_{\mathrm{CV}} = \sup_{h \in [0, Cn^{-\varepsilon}]} CV(h)$ then $h_{CV}/h_{MISE} \xrightarrow[n\to\infty]{w.p.1} 1$

Nevertheless, the second and third terms in question form an unknown random variable, say -*A*, whose variance is of order n^{-1}, thus to subjecting *CV(h)* to a less attractive order of convergence while (1.1) becomes:

$$CV(h) = MISE(h) + E(A) + O\{(n^{-1} + n^{-1/2}h^3)n^{\delta}\}$$

MISE(h) being of order: $O(n^{-1}+h^4)$ makes the integrated variance encompassed by *A* leading to the following asymptotic behavior:

$$CV(h) = IB^2(\hat{f}) + E(A) + O\{(n^{-1} + n^{-1/2}h^3)n^{\delta}\},$$

where $IB^2(\hat{f})$ is the integrated squared bias of $\hat{f}$. *E(A)* being a constant not depending on *h*, the minimizer of *CV(h)* converges to the minimizer of the sum of $IB^2(\hat{f})$ and an unknown function of order n^{-1}.

Alongside the asymptotic deficiency, one may notice some flaws in the demonstration presented by the authors, notably in step 5 where the order of the L_2 convergence should be $O(n^{-3/2} + n^{-3/2}h^2)$ instead of $O(n^{-3/2} + n^{-1}h^4)$, and in step 6 where the second moment of the martingale differences does not seem to concur with any of those calculated based on known kernels.

Such shortcomings are incitements for finding better estimates, as we will see in the results section.

1.4. Current Work

The work at hand develops a new method of bandwidth estimation that has the attractive property of controlled rate of convergence regardless of sample size. This approach is applied in different contexts of functional estimation leading to bandwidth selectors that enjoys minimal assumptions and easy computations in addition to fast order of convergence. The method is then used in multivariate density and distribution function estimation of general bandwidth setting.

CHAPTER TWO: A NEW FAST-CONVERGING AND DATA-BASED BANDWIDTH SELETOR IN FUNCTION KERNEL ESTIMATION WITH APPLICATIONS

2.1. Summary

In this chapter, we introduce a new data based method for bandwidth selection and we present it to jointly apply to the density, its derivatives and the distribution functions all in one setting. We then apply the method to density derivative functional estimation and to estimation of the location parameter.

2.2. Bandwidth Selection In Function Estimation

2.2.1. Introduction

Given a random sample $X_1,\ldots,X_n$ from a distribution with density f , let

$$\hat{f}^{(s)}(x)=\frac{1}{nh^{s+1}}\sum_{i=1}^{n}W^{(s)}\left(\frac{x-X_i}{h}\right) \tag{2.1}$$

denote a kernel estimator of $f^{(s)}$, $s=-1,0,1,2,\ldots$.

With s set to -1 when we estimate the distribution function and $W^{(s)}(.)$ is the kernel that corresponds to estimating $f^{(s)}$.

The most widely accepted means that evaluate the performance of such estimator is in terms of its mean square error and mean integrated square error,

$$mse_h(x) = E\left(\hat{f}^{(s)}(x) - f^{(s)}(x)\right)^2$$

$$MISE_h = \int E\left(\hat{f}^{(s)}(x) - f^{(s)}(x)\right)^2 dx$$

2.2.2. A New Bandwidth Estimator

We introduce two statistics that allow us to estimate the bandwidths minimizing the above two risk functions as follows:

$$D_{h,s}(x) = \frac{\alpha}{n^2 h}\sum_{i=1}^{n} K_h(x - X_i) + \beta \binom{n}{2}^{-1} \sum\sum_{i<j} L_{s,h}(x - X_i) L_{s,h}(x - X_j)$$

$$D_{h,s} = \frac{\alpha S(K)}{nh} + \beta \binom{n}{2}^{-1} \sum\sum_{i<j} L_{s,h} * L_{s,h}(X_i - X_j)$$

Where $g_h(\cdot) = \frac{1}{h} g\left(\frac{\cdot}{h}\right)$ and $S(g) = \int g(x) dx$ when g is integrable.

K and L_s are symmetric functions that verify:

1. $\int u^i L_s(u) du = 0$, $i = 0,\dots,s+1$

2. $0 < \int u^{s+2} L_s(u) du < \infty$

3. $0 < \int K(u) du < \infty$

Remarks:

1. $D_{h,s} = \int D_{h,s}(x)\,dx$

2. When $\alpha = \beta = 1$, $L_s = W$ and $K = W^2$; $h^{-2s}\left[D_{h,s} - S\left(f^2\right)\right]$ is the cross-validatory criterion of Bowman (1982) and Rudemo (1982).

3. When $\alpha = \beta = 1$, $L_s = \sum_{i=1}^{p} c_i W_i$ and $K = L_s^{\ 2}$, W_i's are kernels; $h^{-2s} D_{h,s}$ is the kernel contrast of Ahmad and Ran (2004).

Theorem 2.1: Under the above conditions and if f is differential up to $(s+2)^{th}$ order, then:

$$ED_{h,s}(x) = \alpha S(K) f(x)(nh)^{-1} + \beta\left(\mu_{s+2}(L_s)/(s+2)!\right)^2 f^{(s+2)^2}(x) h^{2s+4} + o\left(h^{2s+4} + n^{-1}\right) \qquad (2.2)$$

$$ED_{h,s} = \alpha S(K)(nh)^{-1} + \beta\left(\mu_{s+2}(L_s)/(s+2)!\right)^2 R\left(f^{(s+2)}\right) h^{2s+4} + o\left(h^{2s+4}\right) \qquad (2.3)$$

Where $\mu_{s+2}(g) = \int u^{s+2} g(u)\,du$ and $R(g) = \int g^2(u)\,du$.

Proof:

$$EK_h(x - X_i) = K_h f(x)$$

$$= \int K(u) f(x - hu)\,du$$

$$= \left(\int K(u)\,du\right) f(x) + o(h)$$

$$E\left(L_{s,h}(x - X_i) L_{s,h}(x - X_j)\right) = \left(L_{s,h} * f(x)\right)^2$$

$$= \left\{\int L_s(u) f(x - hu)\,du\right\}^2$$

$$=\left\{\int L_s(u)\left(\sum_{i=0}^{s+2} f^{(i)}(x)(-h)^i\left(u^i/i!\right)+o\left(h^{s+2}\right)\right)du\right\}^2$$

$$=\left(\int \frac{u^{s+2}}{(s+2)!}L_s(u)du\right)^2 f^{(s+2)^2}h^{2s+4}+o\left(h^{2s+4}\right)$$

In the following discussion, we will use the notations below:

1. $h_m = \arg_h \min mse_h(x)$.

2. $h_M = \arg_h \min MISE_h$.

3. $AMISE_h$ is the asymptotic approximation of $MISE_h$.

4. $h_A = \arg_h \min AMISE_h$.

5. $AD_{h,s}(x)$ is the asymptotic approximation of $ED_{h,s}(x)$.

6. $AD_{h,s}$ is the asymptotic approximation of $ED_{h,s}$.

7. $h_{AD}(x) = \arg_h \min AD_{h,s}(x)$.

8. $h_{AD} = \arg_h \min AD_{h,s}$.

9. $h_D(x) = \arg_h \min D_{h,s}(x)$.

10. $h_D = \arg_h \min D_{h,s}$.

11. $amse_h(x)$ is the asymptotic approximation of $mse_h(x)$.

12. $h_A(x) = \arg_h \min amse_h(x)$

13. $h_o(x) = \arg_h \min D_{h,s}(x)$

14. $h_o(x) = \arg_h \min D_{h,s}$

Theorem 2.2:

1) $h_{AD}(x) = \left(\frac{\alpha}{\beta} \frac{(s+2)!^2 S(K)}{(2s+4)\mu_{s+2}^2(L_s)} \frac{f(x)}{f^{(s+2)^2}(x)} \frac{1}{n} \right)^{\frac{1}{2s+5}} = \Lambda^{-\frac{1}{2s+5}} h_A(x)$ (2.4)

2) $h_{AD} = \left(\frac{\alpha}{\beta} \frac{(s+2)!^2 S(K)}{(2s+4)\mu_{s+2}^2(L_s)} \frac{1}{R\left(f^{(s+2)}\right)} \right)^{\frac{1}{2s+5}} = \Lambda^{-\frac{1}{2s+5}}$ (2.5)

where $\Lambda = \left(\beta(2s+4)\mu_{s+2}^2(L_s) R(W) / \alpha\mu_2^2(W) S(K) \right)^{\frac{1}{2s+5}}$

Since finding estimators for h_m and h_M is achieved by optimizing the criteria functions $D_{h,s}(x)$ and $D_{h,s}$, respectively, the problem at hand becomes that of an M-estimation. Therefore, we will use results obtained for this methodology to prove the consistency and asymptotic normality for $h_D(x)$ and h_D .

2.2.3. Asymptotic Properties

First, we need to derive the weak convergence of the bandwidth estimator.

Lemma 1:

1) For any $h > 0$, $D_{h,s}(x) - ED_{h,s}(x) \xrightarrow{p} 0$ (2.6)

2) For any $h > 0$, $D_{h,s} - ED_{h,s} \xrightarrow{p} 0$ (2.7)

Proof:

$EK_h^2\left(x-X_1\right)=h^{-1}\int K^2\left(u\right)f\left(x-hu\right)du \le \underset{x}{Max}(f)R\left(K\right)h^{-1}<\infty$ since $h\neq 0$.

$E\left(L_{s,h}\left(x-X_1\right)L_{s,h}\left(x-X_2\right)\right)^2=\left\{EL_{s,h}^2\left(x-X_1\right)\right\}^2=\left\{h^{-1}\int L_s{}^2\left(u\right)f\left(x-hu\right)\right\}\le\left(\underset{x}{Max}\, f\left(x\right)\right)^2 R\left(L_s\right)h^{-2}<\infty$
since $h\neq 0$.

Hence

$$\frac{1}{n}\sum_{i=1}^{n}K_n\left(x-X_i\right)-EK_h\left(x-X_1\right)\xrightarrow{p}0$$

$$\binom{n}{2}^{-1}\sum\sum_{i<j}L_{s,h}\left(x-X_i\right)L_{s,h}\left(x-X_j\right)-\left(EL_{s,h}\left(x-X_1\right)\right)^2\xrightarrow{p}0$$

Therefore $D_{h,s}\left(x\right)-ED_{h,s}\left(x\right)\xrightarrow{p}0$

$$E\left(L_{s,h}*L_{s,h}\left(X_1-X_2\right)\right)^2=\iint\left\{L_{s,h}\left(x-y-u\right)L_{s,h}\left(u\right)du\right\}^2 f\left(x\right)f\left(y\right)dxdy$$

$$=\frac{1}{h}\iint\left\{L_s\left(t-v\right)L_s\left(v\right)dv\right\}^2 f\left(x\right)f\left(x-ht\right)dxdt$$

$$\le\frac{1}{h}\iint\left(L_s*L_s\left(t\right)\right)^2 f\left(x\right)\underset{x}{Max}\, f\left(x\right)dxdt<\underset{x}{Max}\, f\left(x\right)R\left(L_s*L_s\right)h^{-1}<\infty$$

since $h\neq 0$.

Hence $\binom{n}{2}^{-1}\sum\sum_{i<j}L_{s,h}*L_{s,h}\left(X_i-X_j\right)-EL_{s,h}*L_{s,h}\left(X_1-X_2\right)\xrightarrow{p}0$

The first term in $D_{s,h}$ is non-stochastic, then $D_{h,s}-ED_{h,s}\xrightarrow{p}0$

Lemma 2:

1. $h_D(x) \xrightarrow{p} 0$ (2.8)

2. $h_D \xrightarrow{p} h_o$ (2.9)

Proof:

By definition, $h_D(x), h_D, h_o(x)$ and h_o are respectively the minimizers of $D_{h,s}(x), D_{h,s}, ED_{h,s}(x)$ and $ED_{h,s}$. In addition, $D_{h,s}(x)$ and $D_{h,s}$ are uniformly consistent. Therefore we can use Theorem 5.7 of Van der Vaart (1998) to obtain the above consistency results.

Let $m_{h,x}(X_i, X_j)$ and $m_h(X_i, X_j)$ be defined such that: $D_{h,s}(x) = \binom{n}{2}^{-1} \sum\sum_{i<j} m_{h,x}(X_i, X_j)$ and $D_{h,s} = \binom{n}{2} \sum\sum_{i<j} m_h(X_i, X_j)$.

Lemma 3:

$$1)\quad E\left(\frac{\partial}{\partial h} m_{h,x}\left(X_i, X_j\right)\Big|_{h_o(x)}\right)^2 = \frac{1}{2}\left(\frac{\beta^{s+7/2}(2s+4)\mu_{s+2}^2(L_s)\left(f^{(s+2)}(x)\right)^2}{\alpha((s+2)!)^2 S(K)(f(x))^{-(s+3/2)}}\right)^{\frac{4}{2s+5}} \left(4R(L_s)R(N_s)+R^2(L_s)\right)+o\left(n^{\frac{3}{2s+5}}\right) \quad (2.10)$$

$$2)\quad E\left(\frac{\partial}{\partial h} m_h\left(X_i, X_j\right)\Big|_{h_o(x)}\right)^2 = \beta^{\frac{4s+13}{2s+5}}\left(\frac{(2s+4)\mu_{s+2}^2(L_s)R\left(f^{(s+2)}\right)}{\alpha((s+2)!)^2 S(K)} n\right)^{\frac{3}{2s+5}} R(L_s * N_s)R(f)+o\left(n^{\frac{2}{2s+5}}\right) \quad (2.11)$$

where $N(u) = -uL'(u) - L(u)$

3) $E\left(\frac{\partial^2}{\partial h^2} m_{h,x}\left(X_i, X_j\right)\Big|_{h_o(x)} \right)$

$$=\left(2+\left(2s+4\right)\left(2s+3\right)\right)\left(\frac{\beta\left(2s+4\right)\mu_{s+2}^2\left(L_s\right)\left(f^{(s+2)}\left(x\right)\right)^2}{\left(\left(s+2\right)!\right)^2}\right)^{\frac{3}{2s+5}}\left(\frac{\alpha S\left(K\right)f\left(x\right)}{n}\right)^{\frac{2s+2}{2s+5}}+o\left(n^{-\frac{2s+2}{2s+5}}\right) \qquad (2.12)$$

4) $E\left(\frac{\partial^2}{\partial h^2} m_{h}\left(X_i, X_j\right)\Big|_{h_o} \right)$

$$=\left(2+\left(2s+4\right)\left(2s+3\right)\right)\left(\frac{\beta\left(2s+4\right)\mu_{s+2}^2\left(L_s\right)R\left(f^{(s+7)}\right)}{\left(\left(s+2\right)!\right)^2}\right)^{\frac{3}{2s+5}}\left(\frac{\alpha S\left(K\right)f\left(x\right)}{n}\right)^{\frac{2s+2}{2s+5}}+o\left(n^{-\frac{2s+2}{2s+5}}\right) \qquad (2.13)$$

Proof:

$$\frac{\partial}{\partial h}\left(\frac{1}{h^2}K\left(\frac{x-X_i}{h}\right)\right)=\frac{1}{h^2}M_h\left(x-X_i\right), \qquad M\left(u\right)=-uK'\left(u\right)-2K\left(u\right)$$

$$\frac{\partial}{\partial h}\left(\frac{1}{h^2}L_s\left(\frac{x-X_i}{h}\right)L_s\left(\frac{x-X_j}{h}\right)\right)=\frac{1}{h}\left(L_{s,h}\left(x-X_i\right)N_{s,h}\left(x-X_j\right)+L_{s,h}\left(x-X_j\right)N_{s,h}\left(x-X_i\right)\right),$$

$$E\left(\frac{\partial}{\partial h}m_{h,x}\left(X_i,X_j\right)\right)^2=h^{-2}E\left(\frac{\alpha}{2nh}\left(M_h\left(x-X_i\right)+M_h\left(x-X_j\right)\right)\right.$$
$$\left.+\beta\left(L_{s,h}\left(x-X_i\right)N_{s,h}\left(x-X_j\right)+L_{s,h}\left(x-X_j\right)N_{s,h}\left(x-X_i\right)\right)\right)^2$$

$$
\begin{aligned}
&= h^{-2}\left\{2\left(\frac{\alpha}{nh}\right)^2\left(EM_h^2(x-X_i)+\left(EM_h(x-X_i)\right)^2\right)\right.\\
&\quad +2\beta^2\left(EL_{s,h}^2(x-X_i)EN_{s,h}^2(x-X_i)+\left(EL_{s,h}(x-X_i)N_{s,h}(x-X_i)\right)^2\right)\\
&\quad +4\frac{\alpha\beta}{nh}\left(EM_h(x-X_i)L_{s,h}(x-X_i)EN_{s,h}(x-X_i)\right.\\
&\qquad\left.\left. +EM_h(x-X_i)N_{s,h}(x-X_i)EL_{s,h}(x-X_i)\right)\right\}
\end{aligned}
$$

$$
Eg_h^2(x-X_i)=R(g)f(x)h^{-1}+o(1)
$$

$$
\begin{aligned}
EM_h(x-X_i) &= -\int\left(uK'(u)+2K(u)\right)f(x+hu)\,du\\
&= -f(x)\left(\int uK'(u)+2K(u)\,du\right)+o(h)\\
&= -f(x)\left(uK(u)\Big|_{-\infty}^{+\infty}-\int K(u)\,du+2\int K(u)\,du\right)+o(h)\\
&= -S(K)f(x)+o(h)
\end{aligned}
$$

$$
\begin{aligned}
EN_{s,h}(x-X_i) &= -\int\left(uL_s'(u)+L_s(u)\right)f(x+hu)\,du\\
&= -\sum_{i=0}^{s+2} f^{(i)}(x)\frac{(-h)^i}{i!}\int\left(u^{i+1}L_s'(u)+L_s(u)\right)du+o\left(h^{s+2}\right)\\
&= \frac{\mu_{s+2}(L_s)}{(s+1)!}f^{(s+2)}(x)\left(-h^{s+2}\right)+o\left(h^{s+2}\right)
\end{aligned}
$$

$$
\begin{aligned}
EL_{s,h}(x-X_i)N_{s,h}(x-X_i) &= -h^{-1}\int\left(uL_s'(u)L_s(u)+L_s^{\,2}(u)\right)f(x-hu)\,du\\
&= -h^{-1}f(x)\left(\int\frac{1}{2}u\,dL_s^{\,2}(u)+R(L_s)\right)+o(1)\\
&= \frac{f(x)R(L_s)}{2h}+o(1)
\end{aligned}
$$

$$EM_h(x-X_i)L_{s,h}(x-X_i)=h^{-1}\int\ M(u)L_s(u)f(x-hu)du$$
$$=S(ML_s)f(x)h^{-1}+o(1)$$

$$EM_h(x-X_i)N_{s,h}(x-X_i)=h^{-1}\int\ M(u)N_s(u)f(x-hu)$$
$$=S(MN_s)f(x)h^{-1}+o(1)$$

Hence

$$E\left(\frac{\partial}{\partial h}m_{h,x}(X_i,X_j)\right)^2=h^{-2}\{2\frac{\alpha^2}{n^2h^2}\left(\frac{R(M)f(x)}{h^2}+s^2(K)f^2(x)+o(1+h)\right)$$

$$+2\beta^2\left(\frac{R(L_s)R(N_s)f^2(x)}{h^2}+\frac{R^2(L_s)f^2(x)}{4h^2}+o\left(\frac{1}{h}\right)\right)$$
$$+\frac{4\alpha\beta}{n}(-h)^s\ \frac{\mu_{s+2}(L_s)f^{(s+2)}(x)f(x)}{(s+2)!}\left((s+2)S(ML_s)+S(MN_s)\right)+o\left(\frac{h^s}{n}\right)\}$$

$$=\frac{2\beta^2}{h^4}\left(R(L_s)R(N_s)+\frac{R^2(L_s)}{4}\right)f^2(x)+o\left(\frac{1}{h^3}\right)$$
$$=\frac{\beta^2}{2h^4}\left(4R(L_s)R(N_s)+R^2(L_s)\right)+f^2(x)+o\left(h^{-3}\right)$$

$$\frac{\partial}{\partial h}\left(L_{s,h}*L_{s,h}(X_i-X_j)\right)=-\frac{1}{h^2}\int L_s\left(\frac{X_i-X_j}{h}-x\right)L_s(x)dx+\frac{1}{h}\int-\frac{X_i-X_j}{h^2}L_s'\left(\frac{X_i-X_j}{h}-x\right)L_s(x)dx$$
$$=\frac{1}{h}P_h(X_i-X_j)$$

where $P(u)=L_s*N_s(u)$

$$E\left(\frac{\partial}{\partial h}m_h(X_i,X_j)\right)^2=\left(-\frac{\alpha S(K)}{nh^2}\right)^2+\frac{\beta^2}{h^2}EP_h^2(X_i-X_j)-\frac{2\alpha\beta S(K)}{nh^3}EP_h(X_i-X_j)$$

$$EP_h^2\left(X_i - X_j\right) = \iint \left\{ \int \frac{1}{h} N_s\left(\frac{x-y}{h} - u\right) L_s(u)\,du \right\}^2 f(x) f(y)\,dxdy$$

$$= \frac{1}{h} \iint \left\{ \int N_s(v-u) L_s(u)\,du \right\}^2 f(x) f(h-hv)\,dxdv$$

$$= \frac{R(L_s * N_s) R(f)}{h} + o(1)$$

Hence

$$E\left(\frac{\partial}{\partial h} m_h\left(x_i, x_j\right)\right)^2 = \frac{\beta^2}{h^3} R(L_s * N_s) R(f) + o\left(\frac{1}{h^2}\right)$$

$$E\left(\frac{\partial^2}{\partial h^2} m_{h,x}\left(X_i, X_j\right)\right) = \frac{\partial^2}{\partial h^2} E\left(m_{h,x}\left(X_i, X_j\right)\right)$$

$$= \frac{\partial^2}{\partial h^2} ED_{h,s}(x)$$

$$= \frac{2\alpha S(K)}{nh^3} f(x) + \beta \frac{(2S+4)(2S+3)\mu_{s+2}^2(L_s)}{(S+2)!^2} f^{(s+2)^2}(x) h^{2s+2} + o\left(h^{2s+2}\right)$$

$$E\left(\frac{\partial^2}{\partial h^2} m_h\left(X_i, X_j\right)\right) = \frac{\partial^2}{\partial h^2} E\left(m_h\left(X_i, X_j\right)\right)$$

$$= \frac{\partial^2}{\partial h^2} ED_h$$

$$= \frac{2\alpha S(K)}{nh^3} + \beta \frac{(2S+4)(2S+3)\mu_{s+2}^2(L_s)}{(S+2)!^2} R\left(f^{(s+2)}\right) h^{2s+2} + o\left(h^{2s+2}\right)$$

Let V_x and V be defined as follows:

$$\frac{n^{\frac{4s+8}{2s+5}}}{2} V_x = E\left(\frac{\partial}{\partial h} m_{h,x}\left(X_i, X_j\right)\Big|_{h_0(x)}\right)^2 \Big/ \left[E\left(\frac{\partial^2}{\partial h^2} m_{h,x}\left(X_i, X_j\right)\Big|_{h_0(x)}\right)\right]^2$$

$$\frac{n^{\frac{4s+7}{2s+5}}}{2} V = E\left(\frac{\partial}{\partial h} m_h\left(X_i, X_j\right)\Big|_{h_0}\right)^2 \Big/ \left[E\left(\frac{\partial^2}{\partial h^2} m_h\left(X_i, X_j\right)\Big|_{h_0}\right)\right]^2$$

$$\frac{n^{2-\frac{2}{2s+5}}}{2}V_x = \frac{1}{8\left(1+(s+2)(2s+3)\right)^2}\left(\frac{\beta^{4s+8}(s+2)!^4 f^2(x)}{\alpha^{4s+8}(2s+4)^2\mu_{s+2}^4(L_s)\left(f^{(s+2)}(x)\right)^4\left(S(K)\right)^{4s+8}}\right)^{\frac{1}{2s+5}}\left(4R(L_s)R(N_s)+R^2(L_s)\right)n^{\frac{4s+7}{2s+5}}\left(1+o(1)\right)$$

$$\frac{n^{2-\frac{3}{2s+5}}}{2}V = \frac{1}{4\left(1+(s+2)(2s+3)\right)^2}\left(\frac{\beta^{4s+7}(s+2)!^6}{\alpha^{4s+7}(2s+4)^3\mu_{s+2}^6(L_s)R^6(f^{(s+2)})\left(S(K)\right)^{4s+7}}\right)^{\frac{1}{2s+5}} R(L_s * N_s)R(f)n^{\frac{4s+7}{2s+5}}\left(1+o(1)\right)$$

Lemma 4:

1. $n^{\frac{1}{2s+5}}(h_D(x)-h_0(x)) \sim AN(0,V_x)$ and $n^{\frac{1}{2s+5}}\frac{(h_D(x)-h_0(x))}{h_0(x)} \sim AN(0,C_x^2)$ (2.14)

2. $n^{\frac{3}{4s+10}}(h_D-h_0) \sim AN(0,V)$ and $n^{\frac{1}{4s+10}}\frac{(h_D-h_0)}{h_0} \sim AN(0,C^2)$ (2.15)

where $C_x^2 = \left(\frac{\beta}{\alpha}\right)^{\frac{4s+7}{2s+5}}\left(\frac{(s+2)!^6 f^3(x)}{(2s+4)^3\mu_{s+2}^6(L_s)\left(f^{(s+2)}(x)\right)^6\left(S(K)\right)^{4s+7}}\right)^{\frac{1}{2s+5}}$

and $C^2 = \left(\frac{\beta}{\alpha}\right)^{\frac{4s+6}{2s+5}}\left(\frac{(s+2)!^8}{(2s+4)^4\mu_{s+2}^8(L_s)R^7\left(f^{(s+2)}\right)\left(S(K)\right)^{4s+6}}\right)^{\frac{1}{2s+5}}$

Proof:

The two results follow from Theorem 5.23 Van der Vaart (1998) along with the consistency of the estimators and the second degree differentiability of the loss functions.

Remark: The above results can be proven via the usual method of Hall (1984), but we choose to proceed differently since this approach is shorter and can be extended to the multivariate case.

The above result leads us to the relative rate of convergence of the bandwidth estimator as shown next.

Theorem 2.3:

1) $$\frac{(\Lambda h_D(x) - h_m(x))}{h_M(x)} \sim AN(0, C_x^2) \quad (2.16)$$

2) $$n^{\frac{1}{4s+10}} \frac{(\Lambda h_D - h_M)}{h_M} \sim AN(0, C^2) \quad (2.17)$$

We can also show as a consequence of Lemma 4 that the following holds.

Theorem 2.4:

1) $$n^{\frac{2}{2s+5}}((\Lambda h_D(x))^2 - h_m^2(x)) \sim AN(0, W_x) \quad (2.18)$$

2) $$n^{\frac{5}{4s+10}}((\Lambda h_D)^2 - h_m^2) \sim AN(0, W) \quad (2.19)$$

Where $$W_x = 4\left(\frac{\beta}{\alpha}\right)^{\frac{4s+7}{2s+5}} \left(\frac{(s+2)!^{10} (R(W)) f^7(x)}{(2s+4)^3 \mu_{s+2}^{10} (L_s) (S(K))^{4s+7} \left(f^{(s+2)}(x)\right)^{10}} \right)^{\frac{1}{2s+5}} \quad (2.20)$$

And $$W = 4\left(\frac{\beta}{\alpha}\right)^{\frac{4s+6}{2s+5}} \left(\frac{(s+2)!^{12} (R(\Phi))^4}{(2s+4)^4 \mu_{s+2}^{12} (L_s) R^{11}\left(f^{(s+2)}\right) (S(K))^{4s+6}} \right)^{\frac{1}{2s+5}} \quad (2.21)$$

Proof: The delta method around $h_m(x)$ and h_m applied to the function $f(h) = h^2$ leads to the above results.

Corollary: If h_m is the global choice of the bandwidth in the kernel estimator of a probability density function, then $\left(\Lambda h_D\right)^2$ converges to h_m^2 with an order of $n^{1/2}$:

$$n^{\frac{1}{2}}(\left(\Lambda h_D\right)^2 - h_m^2) \sim AN(0, W)$$

Where

$$W = \left(\frac{\beta}{\alpha}\right)^{\frac{6}{5}} \left(\frac{2^{14}\left(R(W)\right)^4}{\mu_2^{12}\left(L_s\right) R^{11}\left(f^{(2)}\right)\left(S\left(K\right)\right)^6}\right)^{\frac{1}{5}}$$

2.2.4. Conclusion

The superiority of the method discussed over cross-validatory techniques (least squares and biased cross-validation, bootstrap, etc.) is quite obvious given each of them corresponds to a particular choice of α and β.

If compared to $\sqrt{n}$ -consistent methodologies such as the Plug-In techniques and the smoothed bootstrap, the new approach have the following advantages:

- Minimal requirements about the smoothness of the function to be estimated.
- Needlessness of a pilot estimator unlike the other methods.

- Plug-In techniques require generally a moderately large sample so that the risk function can be replaced by its approximation thus making the $\sqrt{n}$ -consistency questionable when dealing with small sample while the new approach enjoys the unbiasedness of the least squares and biased cross-validation.
- Smaller variance and relative rate of convergence of the bandwidth estimator regardless of the sample size since the parameters α and β can be chosen by the user which does more than compensate for the $\sqrt{n}$ -consistency.

2.3. Applications

Two applications of the above method are now discussed.

2.3.1. Estimation Of The Square Density Derivative Functional

2.3.1.1. Introduction

Consider the problem of estimating the integrated squared density derivative functionals $R\left(f^{(s)}\right)=\int f^{(s)^2}\left(x\right)dx$, $s \geq 0$; based on a random sample $X_1,\ldots,X_n$ from an unknown density f .

Numerous studies addressed this problem by substituting $f\left(x\right)$ in the above integral by its kernel estimate (Hall and Marron (1987), Bickel and Ritov (1988), Jones and Sheather (1991), ...etc.). They succeeded in developing bandwidth selection methodologies that lead to $\sqrt{n}$ consistent estimates for $R\left(f^{(s)}\right)$ by putting assumptions on the degree of smoothness of the density and the order of the kernel. When the kernel condition is relaxed, these authors were able to derive MSE optimal bandwidth, although the one suggested by Jones and Sheather (1991) leads to a significantly smaller estimate. The only drawback of these procedures is their need for a pilot estimate at their initial stages.

To avoid this inconvenience, Chiu (1991, 1992) proceeds differently by working with the Fourier transform of the kernel density of f and was able to get an estimate for $R\left(f^{(s)}\right)$ with the same order consistency. However, this approach performs adequately, only when f has smoothness beyond a certain degree and an unbounded characteristic function.

Since these studies rely on some knowledge of the smoothness of unknown density f, Wu (1995) proposed a procedure that does not require such information. However, since the estimation of $R\left(f^{(s)}\right), s \geq 1$ is mainly needed to obtain a plug-in estimate for f through a bandwidth with a relative rate of convergence of order $n^{-1/2}$, the conditions on the smoothness of f become necessary as expressed by Theorem 2.1(ii) in Wu (1995).

Besides, the proposed estimator, though not requiring a pilot estimate of the bandwidth as is the case for those considered by Hall et. al. (1987, 1991) and Jones and Sheather (1991), does use the functionals of the Normal distribution at some stages of the estimation.

The intent of the current study is to develop a bandwidth selection methodology that yields a fast converging bandwidth and a better performing estimate of $R\left(f^{(s)}\right)$.

2.3.1.2 Bandwidth Estimation

Let $X_1, \ldots, X_n$ be a random sample selected from a population with an unknown density f differentiable up to the p^{th} order. Define the statistic $\hat{\Psi}_s(h)$ as follows:

$$\hat{\Psi}_s(h) = \alpha K^{(2s)}(0) n^{-1} h^{-2s-1} + \beta \binom{n}{2}^{-1} \sum_{i<j}\sum L_h^{(2s)}\left(X_i - X_j\right), \quad \alpha, \beta > 0 \qquad (2.22)$$

Where K is a symmetric kernel of order $2k$ with $(2k)^{th}$ moment $\mu_{2k} < \infty$ and $L(x) = \left(uK(u)\right)' / 2k$.

K is chosen such that $L^{(2s)}(x)$ exists $\forall x \in IR$, and is square-integrable with respect to $F(x) = \int_{-\infty}^{x} f(u)\,du$.

Let $\Psi_s(h)$ be the expectation of $\hat{\Psi}_s(h)$, then:

$$\Psi_s(h) = \alpha \frac{K^{(2s)}}{nh^{2s+1}} + \beta E L_h^{(2s)}\left(X_i - X_j\right)$$

$$= \alpha K^{(2s)}(0) n^{-1} h^{-2s-1} + \beta \iint L_h^{(2s)}(x-y) f(y) f(x)\,dxdy$$

$$= \begin{cases} \alpha K^{(2s)}(0) n^{-1} h^{-2s-1} + \beta \iint (-1)^{(s-k)} L_h(x-y) f^{(s+k)}(y) f^{(s-k)}(x)\,dxdy, & p \le 2s \\ \alpha K^{(2s)}(0) n^{-1} h^{-2s-1} + \beta \iint L_h(x-y) f^{(2s)}(y) f(x)\,dxdy, & p > 2s \end{cases}$$

$$= \begin{cases} \alpha K^{(2s)}(0) n^{-1} h^{-2s-1} + \beta \iint (-1)^{(s-k)} L(u) f^{(s+k)}(y) f^{(s-k)}(y+hu)\,dudy, & p \le 2s \\ \alpha K^{(2s)}(0) n^{-1} h^{-2s-1} + \beta \iint L(u) f^{(2s)}(y) f(y+hu)\,dudy, & p > 2s \end{cases}$$

k is chosen such that $s+k<p$ when $p\le 2s$ and $\max(2,2s)\le 2k<p$ when $p>2s$ thus:

$$\Psi_s(h)=\begin{cases}\alpha K^{(2s)}(0)n^{-1}h^{-2s-1}-\beta\dfrac{(-1)^{(s-k)}}{(2k)!}\mu_{2k}R\left(f^{(s+k)}\right)h^{2k}+\circ\left(h^{2k+1}\right), p\le 2s\\ \alpha K^{(2s)}(0)n^{-1}h^{-2s-1}-\beta\dfrac{\mu_{2k}}{(2k)!}R\left(f^{(s+k)}\right)h^{2k}+\circ\left(h^{2k+1}\right), p>2s\end{cases}$$

Let $a_s(h)$ be defined such that $\Psi_s(h)=a_s(h)+\circ\left(h^{2k+1}\right)$ and let h_a be the solution to the equation $a_s(h)=0$, then:

$$h_a=\begin{cases}\left(\dfrac{\alpha(2k)!K_{(\circ)}^{(2s)}}{\beta(-1)^{(s+k)}\mu_{2k}R\left(f^{(s+k)}\right)n}\right)^{\frac{1}{2s+2k+1}}, p\le 2s\\ \left(\dfrac{\alpha(2k)!K_{(\circ)}^{(2s)}}{\beta\mu_{2k}R\left(f^{(s+k)}\right)n}\right)^{\frac{1}{2s+2k+1}}, p>2s\end{cases} \tag{2.23}$$

Remarks:

1) results obtained under the condition $p\le 2s$ are valid only when $s\ge 2$.

2) The minimum requirement to obtain h_a is that p has to be greater than $\min((s+k),2s)$.

If we let h_{AMSE} be the bandwidth that minimizes $AMSE\left(R\left(\hat{f}\right)\right)$, where $\hat{f}(\cdot)$ is the kernel estimate for $f(\cdot)$ based on kernel $K(\cdot)$ and sample $X_1,...,X_n$, then if we proceed similarly as Jones and Wand (1995) did, we can easily prove the following relation:

Theorem 2.5: $\forall p \geq 2$, $h_{AMSE} = \Lambda h_a$ where $\Lambda = \left(\frac{\beta}{\alpha}\right)^{1/(2k+2s+1)}$ (2.24)

Define h_Ψ and $h_\circ$ to be respectively the zeros of $\Psi_s(h) = 0$ and $B\left(R\left(\hat{f}\right)\right) = 0$, where $B(g)$ is the bias of g. then, We get the following relation:

Theorem 2.6: $\forall p \geq 2$, $h_0 = \Lambda h_\Psi + o(h_\Psi)$ (2.25)

2.3.1.3. The Asymptotic Behavior Of The Bandwidth Estimator

Based on the last result, the study of the asymptotic behavior of $\hat{h}_\circ$, the bandwidth that estimates $h_\circ : \hat{h}_\circ = \Lambda \hat{h}_\Psi$ relies on that of $\hat{h}_\Psi$, the solution to $\hat{\Psi}_s(h) = \circ$.

This reduces our problem to that of studying the properties of a Z-estimator.

For that we define some parameters and functionals as follows:

$\theta_\Psi = n^{\frac{1}{2s+2k+1}} h_\Psi, \hat{\theta}_n = n^{\frac{1}{2s+2k+1}} \hat{h}_\Psi, \hat{\Psi}_\theta = n^{\frac{1}{2s+2k+1}} \hat{\Psi}_s\left(\theta n^{-\frac{1}{2s+2k+1}}\right)$, and $\Psi_\theta = E\hat{\Psi}_\theta$.

Lemma 1: $\sup_{\theta>0}\left|\hat{\Psi}_\theta - \Psi_\theta\right| \xrightarrow{a.s.} 0$ (2.26)

Proof: This follows strict forward from the fact that $\hat{\Psi}_\theta$ is a U-statistic with a square integrable kernel.

Lemma 2: $\hat{\theta}_n \xrightarrow{a.s.} \theta_\Psi$ (2.27)

Proof: Since Ψ_θ is continuous in $\theta \in (0,\infty)$ and θ_Ψ is a unique zero then:

$$\forall \varepsilon > 0 \quad \inf_{\theta:|\theta-\theta_\Psi|\geq\varepsilon} |\Psi_\theta| > |\Psi_{\theta_\Psi}|$$

This means that:

$$\forall \varepsilon > 0 \ \ \exists \eta > 0 \ \ \ni |\hat{\theta}_n - \theta_\Psi| \geq \varepsilon \Rightarrow |\Psi_\theta| > |\Psi_{\theta_\Psi}| + \eta$$

Thus, the event $\{|\hat{\theta}_n - \theta_\Psi| \geq \varepsilon\}$ is contained in the event $\{|\Psi_{\hat{\theta}_n}| - |\Psi_{\theta_\Psi}| > \eta\}$.

On the other hand, by Lemma 1: $|\Psi_{\hat{\theta}_\Psi}| = |\Psi_{\theta_\Psi}| + \circ_{a.s.}(1)$

Since $|\hat{\Psi}_{\hat{\theta}_n}| = 0$, then $|\Psi_{\theta_\Psi}| + \circ_{a.s.}(1) > |\hat{\Psi}_{\hat{\theta}_n}|$ and hence:

$$|\Psi_{\hat{\theta}_n}| - |\Psi_{\theta_\Psi}| < |\Psi_{\hat{\theta}_n}| - |\hat{\Psi}_{\hat{\theta}_n}| + \circ_{a.s.}(1)$$

$$< |\Psi_{\hat{\theta}_n} - \hat{\Psi}_{\hat{\theta}_n}| + \circ_{a.s.}(1) \xrightarrow{a.s.} 0 \text{ by Lemma 1.}$$

So, $P\left(\bigcup_{m\geq n}\left\{|\Psi_{\hat{\theta}_m}| - |\Psi_{\theta_\Psi}| > \eta\right\}\right) \xrightarrow[n\to\infty]{} 0$ and therefore, $P\left(\bigcup_{m\geq n}\left\{|\hat{\theta}_m - \theta_\Psi| > \varepsilon\right\}\right) \xrightarrow[n\to\infty]{} 0, \forall \varepsilon > 0$.

Let $\dot{\Psi}_\theta$ denote the derivative of Ψ_θ with respect to θ. Then we get the following result:

Lemma3:

$$\dot{\Psi}_{\theta_\Psi} = -(2k+2s+1)\left(\alpha K^{2s}(\circ)\right)^{\frac{2k-1}{2s+2k+1}}\left(\beta\mu_{2k}R\left(f^{(s+k)}\right)/(2k!)\right)^{\frac{2s+2}{2s+2k+1}} + \circ\left(n^{-\frac{1}{2s+2k+1}}\right) \qquad (2.28)$$

If we let V_θ denote the asymptotic variance of $\hat{\Psi}_\theta$, we obtain the following:

Lemma 4: $$V_{\theta_\Psi} = 2R(L)^{2s}\, R(f)\,\beta^2 \left(\frac{\beta\mu_{2k} R\left(f^{(s+k)}\right)}{\alpha(2k)!\left|K_{(0)}^{(2s)}\right|} \right)^{\frac{4s+1}{2s+2k+1}} n^{-\frac{3}{2s+2k+1}} \tag{2.29}$$

Proof: $$\operatorname{var}\left(\hat{\Psi}_\theta\right) = \beta^2 n \frac{4k}{2s+2k+1} \operatorname{var}\left(\binom{n}{2}^{-1} \sum\sum_{i<j} L_h^{(2s)}\left(X_i - X_j\right) \right)$$

$$= \beta^2 n \frac{4k}{2s+2k+1} \left(\frac{4}{n} \operatorname{cov}\left(L_h^{(2s)}\left(X_1 - X_2\right), L_h^{(2s)}\left(X_1 - X_3\right) \right) + \frac{2}{n^2} \operatorname{var}\left(L_h^{(2s)}\left(X_1 - X_2\right) \right) \right)$$

$$EL_h^{(2s)}\left(x - X_2\right) = \int L_h^{(2s)}(x-y)\, f(y)\, dy = h^{-2s} \int L^{(2s)}(u)\, f(x-hu)\, du$$

$$= h^{-2s} \sum_{i=0}^{m-1} f^{(i)}(x) \frac{h^i}{i!} (-1)^i \int u^i L^{(2s)}(u) du + h^{-2s} \frac{h^m}{m!} (-1)^m \int f^{(m)}(x - \gamma hu) L^{(2s)}(u)\, du,\ 0 \le \gamma \le 1$$

Where $m = \min(p, 2s+2k)$

If $i \le 2s$ then $\int u^i L^{(2s)}(u)\, du = (-1)^i\, i! \int L^{(2s-i)}(u)\, du = 0$

If $2s < i < 2s+2k$ then $\int u^i L^{(2s)}(u)\, du = (-1)^{2s} (2s)! \int u^{i-2s} L(u)\, du = 0$

If $i = 2s+2k$ then $\int u^i L^{(2s)}(u)\, du = (-1)^{2s} (2s)! \int u^{2k} L(u) = -(2s)!\, \mu_{2k}$

Hence, $$EL_h^{(2s)}\left(x - X_2\right) = \begin{cases} \dfrac{(-1)^p}{p!} h^{p-2s} \displaystyle\int f^{(p)}(x - \gamma hu) L^{(2s)}(u)\, du & p < 2s+2k \\ h^{2k} \dfrac{(2s)!\,\mu_{2k}}{(2s+2k)!} f^{(2s+2k)}(x) + \circ\left(h^{2k}\right) & p \ge 2s+2k \end{cases}$$

$$E\left(L_h^{(2s)}\left(X_1-X_2\right)L_h^{(2s)}\left(X_1-X_3\right)\right)=\begin{cases}\dfrac{h^{2p-4s}}{\left(p!\right)^2}\int\left(\int L^{(2s)}\left(u\right)f^{(p)}\left(x-\gamma hu\right)du\right)^2 f\left(x\right)dx & p<2s+2k\\ h^{4k}\left(\dfrac{\mu_{2k}\left(2s\right)!}{\left(2s+2k\right)!}\right)^2\int f^{(2s+2k)^2}\left(x\right)f\left(x\right)+\circ\left(h^{4k}\right) & p\geq 2s+2k\end{cases}$$

$$\operatorname{cov}\left(L_h^{(2s)}\left(X_1-X_2\right),L_h^{(2s)}\left(X_1-X_3\right)\right)$$

$$=\begin{cases}h^{2p-4s}\left\{\int\left(\int L^{(2s)}\left(u\right)f^{(p)}\left(x-\gamma hu\right)du\right)^2 f\left(x\right)dx\right\}/\left(p!\right)^2 & p<2s+2k\\ h^{4k}\mu_{2k}^2\left\{\left(\dfrac{\left(2s\right)!}{\left(2s+2k\right)!}\right)^2\int f^{(2s+2k)^2}\left(x\right)f\left(x\right)dx-\left(\dfrac{R\left(f^{(s+k)}\right)}{\left(2k\right)!}\right)^2\right\} & p\geq 2s+2k\end{cases}$$

$$E\left(L_h^{(2s)^2}\left(X_1-X_2\right)\right)=\int\int L_h^{(2s)^2}\left(x-y\right)f\left(x\right)f\left(y\right)dxdy=h^{-4s-1}\int L_h^{(2s)^2}(u)f\left(x\right)f\left(x-hu\right)dudx$$

$$=h^{-4s-1}R\left(f\right)R\left(L^{(25)}\right)+\circ\left(h^{-4s}\right)$$

$$\operatorname{var}\left(L_h^{(2s)^2}\left(X_1-X_2\right)\right)=h^{-4s-1}R\left(f\right)R\left(L^{(25)}\right)+\circ\left(h^{-4s}\right)$$

When $\theta=\theta_\Psi$, the covariance term becomes negligible in front of the variance, so:

$$V_{\theta_\Psi}=\beta^2 n^{\frac{4k}{2k+2s+1}}\frac{2}{n^2}\theta_\Psi^{-4s-1}n^{\frac{4s+1}{2s+2k+1}}R\left(f\right)R\left(L^{2s}\right)$$

$$=2R\left(f\right)R\left(L^{2s}\right)\beta^2\theta_\Psi^{-4s-1}n^{-\frac{3}{2s+2k+1}}$$

Lemma 5:

$$n^{\frac{3}{4s+4k+2}}\left(\hat{\theta}_n-\theta_\Psi\right)\sim AN\left(0,\frac{2R\left(f\right)R\left(L^{(2s)}\right)}{\left(2k+2s+1\right)^2K^{(2s)^2}\left(0\right)}\left(\frac{\left(2k\right)!\left|K^{(2s)}\left(0\right)\right|}{\mu_{2k}R\left(f^{(s+k)}\right)}\right)^{\frac{3}{2k+2s+1}}\left(\frac{\beta}{\alpha}\right)^{2-\frac{3}{2k+2s+1}}\right)\quad(2.30)$$

Proof: By Lemma 1: $\hat{\Psi}_{\hat{\theta}_n} = \Psi_{\hat{\theta}_n} + \circ_{a.s.}(1) = \Psi_{\hat{\theta}_n} \circ_p (1)$

$$= \Psi_{\theta_\Psi} + \dot{\Psi}_{\theta_\Psi}\left(\hat{\theta}_n - \theta_\Psi\right) + \circ p\left(\hat{\theta}_n - \theta_\Psi\right) + \circ p(1)$$

By Lemma 2: $\hat{\Psi}_{\hat{\theta}_n} = \Psi_{\theta_\Psi} + \dot{\Psi}_{\theta_\Psi}\left(\hat{\theta}_n - \theta_\Psi\right) + \circ p(1)$

Since $\hat{\Psi}_\theta$ is continuous in θ and $\hat{\theta}_n \xrightarrow{a.s.} \theta_\Psi$ then $\hat{\Psi}_{\hat{\theta}_n} \xrightarrow{a.s.} \hat{\Psi}_{\theta_\Psi}$

Hence the above equation becomes: $\hat{\Psi}_{\hat{\theta}_\Psi} + \circ_p(1) = \Psi_{\theta_\Psi} + \dot{\Psi}_{\theta_\Psi}\left(\hat{\theta}_n - \theta_\Psi\right) + \circ_p(1)$

$\Psi_{\theta_\Psi} = 0$ by definition leads to $\dot{\Psi}_{\theta_\Psi}\left(\hat{\theta}_n - \theta_\Psi\right) = \hat{\Psi}_{\theta_\Psi} + \circ p(1)$

Therefore $\dot{\Psi}_{\theta_\Psi}\left(\hat{\theta}_n - \theta_\Psi\right)$ converges to the same distribution as $\hat{\Psi}_{\theta_\Psi}$ does, i.e,. $N\left(0, V_{\theta_\Psi}\right)$.

So, $\hat{\theta}_n - \theta_\Psi \sim AN\left(0, \left|\dot{\Psi}_{\theta_\Psi}\right|^{-2} V_{\theta_\Psi}\right)$

By multiplying the formulas in Lemma 2 by $n^{-\frac{1}{2s+2k+1}}$ and the one in Lemma 5 by θ_Ψ^{-1}, we get the following:

Theorem 2.7: $\hat{h}_\circ \xrightarrow{a.s.} h_\circ$ (2.31)

Theorem 2.8: $n^{\frac{3}{4s+4k+2}}\left(\dfrac{\hat{h}_\circ - h_\circ}{h_\circ}\right)$

$$\sim AN\left(0, \frac{2R(f)R\left(L^{(2s)}\right)}{(2k+2s+1)^2 K^{(2s)^2}(0)}\left(\frac{(2k)!\left|K^{(2s)}(0)\right|}{\mu_{2k}R\left(f^{(s+k)}\right)}\right)^{\frac{1}{2k+2s+1}}\left(\frac{\beta}{\alpha}\right)^{2-\frac{1}{2k+2s+1}}\right) \quad (2.32)$$

2.3.2. Estimation Of The Location Parameter

2.3.2.1. Presenting The Estimator

Consider the following estimator of the location μ for a population with pdf f_μ, such that $f_\mu(x) = f_0(x-\mu)$:

$$\tilde{\mu} = \frac{1}{A_0(0)} \frac{1}{n(n-1)} \Big(\sum\sum_{i \neq j} X_i K_h(X_i - X_j) - A_1(0) \Big) \tag{2.33}$$

where $A_i(t) = \int u^i f_0(t) K_t * f_0(u) du, \qquad i \in \mathbb{N}, t \in \mathbb{R}_+$

$$\begin{aligned} EX_1K_h(X_1 - X_2) &= \iint xK_h(x-y) f_\mu(x) f_\mu(y) dxdy \\ &= \iint (u+\mu) K_h(u-v) f_0(u) f_0(v) dudv \\ &= \int f_0(u) K_h * f_0(u) du + \int u f_0(u) K_h * f_0(u) du \\ &= \mu A_0(h) + A_1(h) \end{aligned}$$

Hence $E\tilde{\mu} = \dfrac{A_0(h)}{A_0(0)}\mu + \dfrac{A_1(h)}{A_0(0)} - \dfrac{A_1(0)}{A_0(0)}$

The above estimator was introduced by Ahmad (1982) who showed that it possesses a nice asymptotic behavior in addition to its robustness against dependence in the sample.

To improve on such properties, we study the following unbiased estimator of μ :

$$\hat{\mu} = \frac{1}{A_0(h)} \frac{1}{n(n-1)} \sum\sum_{i \neq j} X_i K_h(X_i - X_j) - \frac{A_1(h)}{A_0(h)} \tag{2.34}$$

2.3.2.2. Properties Of The Estimator

Notice that $\hat{\mu}$ can be written in the form of a U-statistic as follows:

$$\hat{\mu} = \binom{n}{2}^{-1} \sum_{i \neq j}\sum \left(\frac{1}{A_0(h)} \left(\frac{X_i + X_j}{2} \right) K_h(X_i - X_j) - \frac{A_1(h)}{A_0(h)} \right) \qquad (2.35)$$

Let $\varphi(x, y) = (x + y)K_h(x - y)$, then:

$$E\varphi^2(X_1, X_2) = \iint (x + y)^2 K_h^2(x - y) f_\mu(x) f_\mu(y) dxdy$$

$$= \iint (u + v + 2\mu)^2 K_h^2(u - v) f_0(u) f_0(v) dudv$$

$$= \iint 4\mu^2 K_h^2(u - v) f_0(u) f_0(v) dudv$$
$$+ \iint 4\mu(u + v) K_h^2(u - v) f_0(u) f_0(v) dudv$$
$$+ \iint (u + v)^2 K_h^2(u - v) f_0(u) f_0(v) dudv$$

$$= 4\mu^2 \int f_0(u) K_h^2 * f_0(u) du + 8\mu \int u f_0(u) K_h^2 * f_0(u) du$$
$$+ 2\int u^2 f_0(u) K_h^2 * f_0(u) du + \iint 2uv K_h^2(u - v) f_0(u) f_0(v) dudv$$

$$= 4\int (u + \mu)^2 f_0(u) K_h^2 * f_0(u) du + 2\int u f_0(u) J_h * f_0(u) du$$

where $J_h(u) = uK_h^2(u)$

$$E(\varphi(X_1, X_2)\varphi(X_1, X_3)) = \iint (x + y)(x + y) K_h(x - y) K_h(x - z) f_\mu(x) f_\mu(y) f_\mu(z) dxdydz$$

$$= \iint (u + v + 2\mu)(u + w + 2\mu) K_h(u - v) K_h(u - w) f_0(u) f_0(v) f_0(w) dudvdw$$

$$
\begin{aligned}
&= \int 4\mu^2 f_0(u) \int K_h(u-v) f_0(v) dv \int K_h(u-w) f_0(w) dw du \\
&+ \int 4\mu u f_0(u) \int K_h(u-v) f_0(v) dv \int K_h(u-w) f_0(w) dw du \\
&+ \int 4\mu v f_0(v) \int K_h(u-v) f_0(u) \int K_h(u-w) f_0(w) dw du dv \\
&+ \int u^2 f_0(u) \int K_h(u-v) f_0(v) dv \int K_h(u-w) f_0(w) dw du \\
&+ \iint 2uv K_h(u-v) f_0(u) f_0(v) \int K_h(u-w) f_0(w) dw du dv \\
&+ \iint vw f_0(v) f_0(w) \int K_h(u-v) K_h(u-w) f_0(u) du dv dw
\end{aligned}
$$

$$
\begin{aligned}
&= 4\mu^2 \int f_0(u) \left(\int K_h * f_0(u) \right)^2 du + 4\mu \int u f_0(u) \left(\int K_h * f_0(u) \right)^2 du \\
&+ \int u^2 f_0(u) \left(\int K_h * f_0(u) \right)^2 du + 4\mu \int u f_0(u) \left(\int K_h * f_0(u) \right)^2 du \\
&- 4h\mu \int f_0(u) K_h * f_0(u) L_h * f_0(u) du - 2h \int u f_0(u) K_h * f_0(u) L_h * f_0(u) du \\
&+ 2\int u^2 f_0(u) \left(\int K_h * f_0(u) \right)^2 du + h^2 \int f_0(u) \left(\int L_h * f_0(u) \right)^2 du \\
&- \int u^2 f_0(u) \left(\int K_h * f_0(u) \right)^2 du - 2h \int u f_0(u) K_h * f_0(u) L_h * f_0(u) du \\
&+ 2\int u^2 f_0(u) \left(\int K_h * f_0(u) \right)^2 du
\end{aligned}
$$

where $L(u) = uK(u)$, then

$$
\begin{aligned}
E(\varphi(X_1, X_2)\varphi(X_1, X_3)) = 4\int f_0(u) \Big[&(u+\mu)^2 (K_h * f_0(u))^2 \\
&- h(u+\mu) K_h * f_0(u) L_h * f_0(u) du + \frac{h^2}{4} \left(L_h * f_0(u) \right)^2 \Big] du
\end{aligned}
$$

Let σ^2 be the variance of $\hat{\mu}$, then:

$$
\begin{aligned}
\sigma^2 &= \frac{1}{4A_0^2(h)} \left[\binom{n}{2}^{-1} \operatorname{var}(\varphi(X_1, X_2)) + 4\binom{n}{1}^{-1} \operatorname{cov}(\varphi(X_1, X_2), \varphi(X_1, X_3)) \right] \\
&= \frac{1}{A_0^2(h)} \Bigg[\binom{n}{2}^{-1} \left(\int f_0(u) \left[(u+\mu)^2 K_h^2 * f_0(u) + 2uJ_h * f_0(u) - \mu^2 \right] du \right) \\
&\quad + 4\binom{n}{1}^{-1} \left(\int f_0(u) \Big[(u+\mu)^2 (K_h * f_0(u))^2 - h(u+\mu) K_h * f_0(u) L_h * f_0(u) + \frac{h^2}{4} \left(L_h * f_0(u) \right)^2 - \mu^2 \Big] du \right) \Bigg]
\end{aligned}
$$

The asymptotic behavior of $\hat{\mu}$ is expressed in the following result:

Theorem 2.9

$$\forall h \geq 0: \qquad \sqrt{n}\left(\hat{\mu}-\mu\right) \sim AN\left(0,\sigma^2\right) \qquad (2.36)$$

Proof: Since $\varphi^2\left(X_1, X_2\right) < \infty$, the above follows from the asymptotic theory of U-statistics.

2.3.2.3. Example: *f* is the *pdf* of a normal population

To simplify the calculations we will choose K to be Φ the standard normal *pdf* .

f_o will also be chosen to be Φ because of the scale invariance of the problem:

$$K_h * f_o(u) = \Phi_{\left(1+h^2\right)^{1/2}}(u)$$

$$L_h * f_o(u) = \int \frac{u-v}{h}\,\Phi_h\left(\frac{u-v}{h}\right)\Phi(v)\,dv = -h\int \Phi'\left(\frac{u-v}{h}\right)\Phi(v)\,dv = -h\int \Phi'_h * \Phi(u) = -h\,\Phi'_{\left(1+h^2\right)^{1/2}}(u)$$

$$f_o(u)\left(K_h * f_o(u)\right)^2 = \Phi(u)\Phi^2_{\left(1+h^2\right)^{1/2}}(u) = \frac{-\left[\left(1+h^2\right)/\left(3+h^2\right)\right]^{1/2}}{2\pi\left(1+h^2\right)\sqrt{2\pi}\left(\dfrac{1+h^2}{3+h^2}\right)^{1/2}}\exp\left(-\frac{u^2}{2}\frac{3+h^2}{1+h^2}\right)$$

$$= \frac{1}{2\pi\left(\left(1+h^2\right)/\left(3+h^2\right)\right)^{1/2}}\Phi_{\left(\frac{1+h^2}{3+h^2}\right)^{1/2}}(u)$$

$$f_o(u)\left(L_h * f_o(u)\right)^2 = h^2\Phi(u)\Phi'^2_{\left(1+h^2\right)^{1/2}}(u) = \frac{h^2u^2}{2\pi\left(1+h^2\right)^{5/2}\left(3+h^2\right)^{1/2}}\Phi_{\left(\frac{1+h^2}{3+h^2}\right)^{1/2}}(u)$$

$$f_o(u)\left(K_h * f_o(u)\right)\left(L_h * f_o(u)\right) = \frac{hu}{2\pi\left(1+h^2\right)^{3/2}\left(3+h^2\right)^{1/2}}\Phi_{\left(\frac{1+h^2}{3+h^2}\right)^{1/2}}(u)$$

$$E\left(\varphi\left(X_1,X_2\right)\varphi\left(X_1,X_3\right)\right)=4\{\frac{1}{\left(1+h^2\right)^{1/2}\left(3+h^2\right)^{1/2}}\left[\frac{1+h^2}{3+h^2}+\mu^2\right]-\frac{h^2}{2\pi\left(1+h^2\right)^{3/2}\left(3+h^2\right)^{1/2}}\frac{1+h^2}{3+h^2}$$

$$+\frac{h^4/4}{2\pi\left(1+h^2\right)^{5/2}\left(3+h^2\right)^{1/2}}\frac{1+h^2}{3+h^2}\}$$

$$=\frac{4}{2\pi\left(1+h^2\right)^{1/2}\left(3+h^2\right)^{1/2}}\left(\frac{\left(2+h^2\right)^2}{4\left(1+h^2\right)\left(3+h^2\right)}+\mu^2\right)$$

$$f_o\left(u\right)K_h * f_o\left(u\right)=\Phi\left(u\right)\Phi_{\left(1+h^2\right)^{1/2}}=\frac{1}{\sqrt{2\pi}\left(2+h^2\right)^{1/2}}\Phi_{\left(\frac{2+h^2}{1+h^2}\right)^{1/2}}\left(u\right)$$

$$E\left(\varphi\left(X_1,X_2\right)\right)^2=4\mu^2\frac{1}{2\pi\left(2+h^2\right)}$$

$$cov\left(\varphi\left(X_1,X_2\right)\varphi\left(X_1,X_3\right)\right)=\frac{4}{2\pi}\left\{\mu^2\left(\frac{1}{\left(\left(2+h^2\right)-1\right)^{1/2}}-\frac{1}{\left(2+h^2\right)}\right)+\frac{\left(2+h^2\right)}{4\left(\left(2+h^2\right)^2-1\right)^{3/2}}\right\}$$

$$f_o\left(u\right)K_h^2 * f_o\left(u\right)=\Phi\left(u\right)\frac{1}{\sqrt{2\pi h}}\Phi_{h/\sqrt{2}} * \Phi\left(u\right)=\frac{1}{\sqrt{2\pi h}}\Phi_{\left(1+h^2/2\right)^{1/2}}\left(u\right)\Phi\left(u\right)=\frac{1}{2\pi h\left(4+h^2\right)^{1/2}}\Phi_{\left(\frac{2+h^2}{4+h^2}\right)^{1/2}}\left(u\right)$$

$$f_o\left(u\right)J_h * f_o\left(u\right)du=\frac{hu}{2\pi h\left(2+h^2\right)\left(4+h^2\right)^{1/2}}\Phi_{\left(\frac{2+h^2}{4+h^2}\right)^{1/2}}\left(u\right)$$

$$E\left(\varphi^2\left(X_1,X_2\right)\right)=\frac{4}{2\pi h\left(4+h^2\right)^{1/2}}\left\{\mu^2+\frac{2+h^2}{4+h^2}+\frac{h}{2\left(2+h^2\right)}\frac{2+h^2}{4+h^2}\right\}=\frac{4}{2\pi h\left(4+h^2\right)^{1/2}}\left\{\mu^2+\frac{3}{2}\frac{4/3+h^2}{4+h^2}\right\}$$

$$\operatorname{var}\left(\varphi\left(X_1,X_2\right)\right)=\frac{4}{2\pi}\left\{\mu^2\left(\frac{1}{\left(\left(2+h^2\right)-4\right)^{1/2}}-\frac{1}{\left(2+h^2\right)}\right)+\frac{\left(4/3+h^2\right)h^2}{\left(\left(2+h^2\right)^2-4\right)^{3/2}}\right\}$$

$$\operatorname{var}(\hat{\mu}) = \frac{1}{n}\{\mu^2\left[\left(\frac{2+h^2}{\left(\left(2+h^2\right)^2-1\right)^{1/2}}-1\right)+\frac{1}{n-1}\left(\frac{2+h^2}{\left(\left(2+h^2\right)^2-4\right)^{1/2}}-1\right)\right]$$

$$+\frac{1}{4}\frac{\left(2+h^2\right)^3}{\left(\left(2+h^2\right)^2-1\right)^{3/2}}+\frac{3}{2}\frac{1}{n-1}\frac{h^2\left(2+h^2\right)\left(4/3+h^2\right)}{\left(\left(2+h^2\right)^2-4\right)^{3/2}}\}$$

To obtain a minimal variance we choose a bandwidth that minimizes the formula:

$$\operatorname{var}(\hat{\mu}) = \frac{1}{n}\left(\mu^2\left(\frac{2+h^2}{\left(\left(2+h^2\right)^2-1\right)^{1/2}}-1\right)+\frac{\left(2+h^2\right)^3}{4\left(\left(2+h^2\right)^2-1\right)^{3/2}}\right)+O(n^{-2}) \qquad (2.37)$$

CHAPTER THREE: AUTOMATIC BANDWIDTH SELECTOR FOR NONPARAMETRIC MULTIVARIATE FUNCTION ESTIMATION

3.1 Summary

Unlike in the univariate case, bandwidth selection has not benefited of must interest in multivariate function estimation. The existing methodologies either use a single smoothing parameter as are the cases of the cross-validatory techniques or suffers the curse of dimensionality and subjective assumptions as do plug-in methods. This chapter studies the properties of optimal bandwidths in multivariate density and distribution functions estimation and extends last section's methodology to produce a new bandwidth selector in the multivariate setting.

3.2. Multivariate Density Estimation

3.2.1 Introduction

Consider a sample of n iid random variates $X_1, \ldots, X_n$ from a d-dimensional population with pdf f. Define the kernel density estimator of f by:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} w_{\mathrm{H}}(x - X_i), \qquad x \in IR^d \tag{3.1}$$

where H, the bandwidth matrix is a symmetric positive definite $d \times d$ matrix such that verifies the usual two conditions:

$$\mathrm{H} \xrightarrow[n\to\infty]{} 0_d \text{ , where } 0_d \text{ is the null (d×d) matrix}$$

$$n^{-1} |\mathrm{H}|^{-1/2} \xrightarrow[n\to\infty]{} 0$$

$w(.)$, the kernel, is a multivariate density function that satisfy:

$$\int_{IR^d} z w(z) dz = 0$$

$$\int_{IR^d} z z' w(z) dz = \mu_2(w) \mathrm{I}_d \text{ .}$$

and for every function g: $g_{\mathrm{H}}(x) = |\mathrm{H}|^{-1/2} g\left(\mathrm{H}^{-1/2} x\right)$.

The mean square error $\left(mse_{\mathrm{H}}(x)\right)$ and the mean integrated square error $\left(MISE_{\mathrm{H}}\right)$ have often been used as measures of performance of the above estimator in multivariate kernel estimates as well as in the univariate case. They are defined as follows:

$$mse_{\mathrm{H}}(x) = \mathrm{E}\left(\hat{f}(x) - f(x)\right)^2$$

$$MISE_{\mathrm{H}} = \mathrm{E}\int\left(\hat{f}(x) - f(x)\right)^2 dx$$

Selecting a good smoothing parameter H is very crucial for a good behavior of $\hat{f}(x)$. Therefore, bandwidths that minimize the above two risk functions have been widely considered to be quite appropriate.

3.2.2. Optimal Bandwidth Calculation

Now let's define $D_f(x)$ and $H_f(x)$ to be the gradient and the Hessian matrix of $f(x)$ and $\varphi_f(x) = vec(H_f(x))vec'(H_f(x))$, $\Psi_f = \int_{IR^d} \varphi_f(x)dx$ and $R(g) = \int_{IR^d} g^2(x)dx$. Following the calculations in Wand (1992), we can obtain the following:

Theorem 3.1

$$mse_{\mathrm{H}}(x) = amse_{\mathrm{H}}(x) + o\left(n^{-1}|\mathrm{H}|^{-1/2} + tr^2(\mathrm{H})\right), \ \forall x \in IR^d \tag{3.2}$$

$$MISE_{\mathrm{H}} = \mathrm{A}MISE_{\mathrm{H}} + o\left(n^{-1}|\mathrm{H}|^{-1/2} + tr^2(\mathrm{H})\right) \tag{3.3}$$

where

$$amse_{\mathrm{H}}(x) = n^{-1}|\mathrm{H}|^{-1/2} R(w) f(x) + \frac{1}{4}\mu_2^2(w)vec'(\mathrm{H})\varphi_f(x)vec(\mathrm{H}) \tag{3.4}$$

$$\mathrm{A}MISE_{\mathrm{H}} = n^{-1}|\mathrm{H}|^{-1/2} R(w) + \frac{1}{4}\mu_2^2(w)vec'(\mathrm{H})\Psi_f vec(\mathrm{H}) \tag{3.5}$$

By simplifying the structure of H, Wand and Jones (1995) developed an algorithm that converges to the bandwidth matrix $\mathrm{H}_{\mathrm{A}MISE}$ that minimizes the $\mathrm{A}MISE_{\mathrm{H}}$. However, one can obtain the orders of $H_{amse}(x)$ and H_{AMISE} the $amse_{\mathrm{H}}(x)$-optimal and $AMISE_{\mathrm{H}}$-optimal matrices. Explicit expressions of $H_{amse}(x)$ can be obtained when x is in some specific regions. These regions happen to be the ones in which the different structures of the population are more likely to appear (mounds and valleys).

Theorem 3.2:

(i) For every $x \in IR^d$ such that $H_f(x)$ is positive or negative definite,

$$H_{amse}(x) = \left(\frac{R(w) f(x) \| \mathrm{H}_f(x) \|^{\frac{1}{2}}}{\mu_2^2(w) dn} \right)^{\frac{2}{d+4}} S_f(x), \tag{3.6}$$

where $S_f(x)$ is the positive definite square root of $\left(H_f^{-1}(x)\right)^2$

(ii) $$H_{AMISE} = \mathrm{A}(f) \left(\frac{R(w)}{\mu_2^2(w) n} \right)^{\frac{2}{4+d}} \tag{3.7}$$

where *A(f)* is a positive definite matrix that verifies $\Psi_f vec(\mathrm{A}) = |\mathrm{A}|^{-1/2} vec(\mathrm{A})^{-1}$

(iii) For every $x \in IR^d$ such that $|H_f(x)| \neq 0$,

$$\mathrm{H}_{amse}(x) = \mathrm{A}(x, f) \left(\frac{R(w) f(x)}{\mu_2^2(w) n} \right)^{\frac{2}{4+d}}, \tag{3.8}$$

where *A(x,f)* is a positive definite matrix that verifies $\varphi_f vec(\mathrm{A}) = |\mathrm{A}|^{-1/2} vec(\mathrm{A})^{-1}$

Proof:

(i) $amse_{\mathrm{H}}(x)$ can be written as:

$$amse_{\mathrm{H}}(x) = n^{-1}|\mathrm{H}|^{-1/2} R(w) f(x) + \frac{1}{4}\mu_2^2(w) vec'(\mathrm{H})\varphi_f(x) vec(\mathrm{H}),$$

hence $\dfrac{\partial amse_{\mathrm{H}}(x)}{\partial vec'(\mathrm{H})} = \dfrac{-n^{-1}}{2} R(w) f(x)|\mathrm{H}|^{-3/2} vec'(\mathrm{H}_{\#}) + \dfrac{1}{2}\mu_2^2(w) vec'(\mathrm{H})\varphi_f(x)$

where $\mathrm{H}_{\#}^{1/2}$ is adjoint matrix of $\mathrm{H}^{1/2}$ (see e.g. Schott (1997) ch. 8).

Then $\dfrac{\partial amse_H(x)}{\partial vec'(\mathrm{H})} = -n^{-1} R(w) f(x)|\mathrm{H}|^{-1/2} vec'(\mathrm{H}^{-1}) + \mu_2^2(w) vec'(\mathrm{H})\varphi_f(x)$

where I_d is the ($d \times d$) identity matrix.

We can write the second half of the RHS as: $\mu_2^2(w) tr(\mathrm{H}\mathsf{H}_f(x)) vec'(\mathsf{H}_f(x))$ (see e.g. Schott (1997) ch. 7)

Hence $\dfrac{\partial amse_{\mathrm{H}}(x)}{\partial vec'(\mathrm{H})} = 0$ implies that:

$\mu_2^2(\omega) tr(\mathsf{H}_f(x)\mathrm{H}) vec(\mathsf{H}_f(x)) = \dfrac{R(w) f(x)}{n}|\mathrm{H}|^{-1/2} vec((\mathrm{H})^{-1})$, which is equivalent to:

$$\mu_2^2(w) tr(\mathrm{H}\mathsf{H}_f(x)) vec(\mathsf{H}_f(x)) = \frac{R(w) f(x)}{n|\mathrm{H}|^{1/2}} vec((\mathrm{H})^{-1})$$

Therefore H can be written as $\mathrm{H} = \alpha \mathsf{H}_f^{-1}(x)$ for some $\alpha \in IR$.

Replacing in the preceding formula leads to:

$$\alpha d\mu_2^2(w)vec(\mathsf{H}_f(x)) = \frac{R(w)f(x)}{\alpha n}\frac{\|\mathsf{H}_f(x)\|^{1/2}}{|\alpha^d|^{1/2}}vec(\mathsf{H}_f(x))$$

Hence: $|\alpha| = \left(\frac{R(w)f(x)\|\mathsf{H}_f(x)\|^{1/2}}{d\mu_2^2(w)n}\right)^{\frac{2}{d+4}}$

Then $\mathrm{H} = \alpha H_f^{-1}(x)$, can be expressed, since H is positive definite, in the form:

$$\mathrm{H} = \left(\frac{R(w)f(x)\|H_f(x)\|^{1/2}}{d\mu_2^2(w)n}\right)^{\frac{2}{d+4}} S_f(x)$$

(ii) Just like in (i), one can show that the bandwidth matrix that minimizes the *AMISE* verifies the equation: $\mu_2^2(w)\Psi_f vec(\mathrm{H}) = \frac{R(w)}{n}|\mathrm{H}|^{-1/2} vec(\mathrm{H}^{-1})$

Let's write H as $\left(\frac{R(w)}{\mu_2^2}n^{-1}\right)^{\gamma}$ A ,for some $\gamma \in \mathbb{R}$ and some (d$\otimes$d) matrix A . This leads to:

$$\Psi_f vec(\mathrm{A}) = \left(\frac{\mu_2^2(w)}{R(w)}n\right)^{\left(\gamma\left(\frac{4+d}{2}\right)-1\right)} |\mathrm{A}|^{-1/2} vec(\mathrm{A}^{-1})$$

If we choose γ to be $\frac{2}{4+d}$ then A is the solution to the equation: $\Psi_f vec(\mathrm{A}) = |\mathrm{A}|^{-1/2} vec(\mathrm{A})^{-1}$

Such solution exists as a shown by Wand (1992) and depends only on f .

(iii) Similar approach as in (ii) yields the result.

3.2.3. Bandwidth Estimation

To estimate the bandwidth matrices $\mathrm{H}_a(x)$ and H_A we introduce the following statistics:

$$\hat{a}_\mathrm{H}(x) = \alpha n^{-2} |\mathrm{H}|^{-\frac{1}{2}n} \sum_{i=1}^{n} K_\mathrm{H}(x - X_i) + \beta \binom{n}{2}^{-1} \sum_{i<j}\sum L_\mathrm{H}(x - X_i) L_\mathrm{H}(x - X_j) \tag{3.9}$$

$$\hat{\mathrm{A}}_\mathrm{H} = \alpha S(\mathrm{K}) n^{-1} |\mathrm{H}|^{-\frac{1}{2}} + \beta \binom{n}{2}^{-1} \sum_{i<j}\sum L_\mathrm{H} * L_\mathrm{H}(x_i - X_j) \tag{3.10}$$

where $S(g) = \int_{IR^d} g(x)\,dx$ and * is the convolution sign.

K and L are symmetric functions that verify:

1. $\int_{IR^d} L(u)du = 0$

2. $\int_{IR^d} uL(u)du = 0$

3. $\int_{IR^d} uu'L(u)du = \mu_2(L)\mathrm{I}_d$

4. $0 < \int_{IR^d} \mathrm{K}(u)du < \infty$

Remarks:

1- When $\alpha = \beta = 1$ and H is a dxd diagonal matrix, $\hat{\mathrm{A}}_\mathrm{H}$ is equal to the least-squares cross-validation, the biased cross-validation and the bootstrap estimators depending on the choice of K and L. We will see later in that such a choice does not influence the rate of convergence as α and β do.

2- $\hat{A}_H = \int_{IR^d} \hat{a}_H(x)dx$

If we define $a_H(x) = E\hat{a}_H(x)$ and $A_H = E\hat{A}_H$, then we get:

Theorem 3.3:

$$a_H(x) = \alpha S(K) f(x)\left(n|H|^{1/2}\right)^{-1} + \beta\frac{\mu_2^2(L)}{4} vec'(H)\varphi_f(x) vec(H) + \circ\left(\left(n|H|^{1/2}\right)^{-1} tr\left(H^{1/2}\right) + tr^2(H)\right) \tag{3.11}$$

$$A_H = \alpha S(K)\left(n|H|^{1/2}\right)^{-1} + \beta\frac{\mu_2^2(L)}{4} vec'(H)\psi_f vec(H) + \circ\left(\left(n|H|^{1/2}\right)^{-1} tr\left(H^{1/2}\right) + tr^2(H)\right) \tag{3.12}$$

Proof:

The two terms in $a_H(x)$ are obtained as follows:

$$EK_H(x - X_i) = \int_{IR_d} K(u) f\left(x - H^{1/2}u\right) = f(x)\int_{IR^d} K(u) - D'_f(x) H^{1/2}\int_{IR^d} uK(u)du + \circ\left(tr\left(H^{1/2}\right)\right)$$

$$= S(K) f(x) + \circ\left(tr\left(H^{1/2}\right)\right) \text{ since } \mathrm{K}(u) \text{ is symmetric.}$$

$$E\left(L_H(x - x_1) L_H(x - x_2)\right) = \left\{\int_{IR^d} L(u) f\left(x - H^{1/2}u\right) du\right\}^2$$

$$= \{ f(x)\int_{IR^d} L(u)du - D'_f(x) H^{1/2}\int_{IR^d} uL(u)du$$

$$+ \frac{1}{2} tr\left(H^{1/2}\mathrm{H}_f(x) H^{1/2}\int_{IR^d} uu'L(u)du\right) + \circ\left(tr(H)\right)\}^2$$

$$= \frac{1}{4}\mu_2^2(L) tr^2\left(H\mathrm{H}_f(x)\right) + \circ\left(tr^2(H)\right)$$

$$= \frac{\mu_2^2(L)}{4} vec'(H)\varphi_f(x)vec(H) + \circ\left(tr^2(H)\right)$$

Integrating both expressions will lead to the terms in $\mathrm{A_H}$.

If we let $H_a(x)$ and $\mathrm{H_A}$ be the bandwidth matrices that minimize $a_H(x)$ and $\mathrm{A_H}$, then it is easy to prove the following relationships.

Theorem 3.4:

1) For every $x \in IR^d$ such that $\left|H_f(x)\right| \neq 0$, $\quad H_{mse}(x) = \Lambda H_a(x)$

2) $H_{MISE} = \Lambda H_A$,

where $\Lambda = \left(\dfrac{\beta R(\omega)\mu_2^2(L)}{\alpha S(K)\mu_2^2(\omega)}\right)^{\frac{1}{d+4}}$

3.2.4. Asymptotic Properties Of Bandwidth Selectors

Let $\mathrm{H}_n(x)$ and H_n be the bandwidth matrices that minimize $\hat{a}_{\mathrm{H}}(x)$ and $\hat{\mathrm{A}}_{\mathrm{H}}$.

Based on the previous result, we define $\hat{\mathrm{H}}(x) = \Lambda \mathrm{H}_n(x)$ and $\hat{\mathrm{H}} = \Lambda \mathrm{H}_n$ as potential estimators of $\mathrm{H}_{mse}(x)$ and H_{MISE}, the bandwidth matrices that minimize $mse\left(\hat{f}(x)\right)$ and $MISE\left(\hat{f}\right)$.

So, the problem at hand is that of minimizing two objective functions, so we use available results from the theory of M-estimation to find asymptotic behaviors of $\mathrm{H}_n(x)$ and H_n and hence those of $\hat{\mathrm{H}}(x)$ and $\hat{\mathrm{H}}$.

First, let P be the set of all ($d \times d$) positive definite matrices.

Lemma 1:

(i) $\sup_{\mathrm{H} \in \mathrm{P}} \left| \hat{a}_{\mathrm{H}}(x) - a_{\mathrm{H}}(x) \right| \xrightarrow{a.s.} 0, \qquad \forall x \in IR^d$

(ii) $\sup_{\mathrm{H} \in \mathrm{P}} \left| \hat{\mathrm{A}}_{\mathrm{H}} - \mathrm{A}_{\mathrm{H}} \right| \xrightarrow{a.s.} 0$

Proof:

The above two results follow directly from the SLLN of U-Statistics.

Lemma 2:

(i) $\mathrm{H}_n(x) \xrightarrow{a.s.} \mathrm{H}_a(x)$

(ii) $\mathrm{H}_n \xrightarrow{a.s.} \mathrm{H}_{\mathrm{A}}$

Proof:

By definitions of $\mathrm{H}_a(x) : \forall \varepsilon > 0 \quad \inf_{\mathrm{H} \in \mathrm{P} : \left\| \mathrm{H} - \mathrm{H}_a(x) \right\| \geq \varepsilon} a_{\mathrm{H}}(x) \geq a_{\mathrm{H}_a(x)}(x)$

Then $\forall \varepsilon > 0 \ \ \exists \eta > 0$ s.t. $\left\| \mathrm{H} - \mathrm{H}_a(x) \right\| \ge \varepsilon \Rightarrow a_{\mathrm{H}}(x) - a_{\mathrm{H}_a(x)}(x) \ge \eta$

Thus the event $\left\{ \left\| \mathrm{H}_n(x) - \mathrm{H}_a(x) \right\| \ge \varepsilon \right\}$ is contained in $\left\{ a_{\mathrm{H}_n}(x) - a_{\mathrm{H}_a(x)}(x) \ge \eta \right\}$

Also, by Lemma 1, $\hat{a}_{\mathrm{H}_a(x)}(x) = a_{\mathrm{H}_a(x)} + o_{a.s.}(1)$

$\hat{a}_{\mathrm{H}_a(x)}(x) > \hat{a}_{\mathrm{H}_n(x)}(x)$ implies that $a_{\mathrm{H}_a(x)} + o_{a.s.}(1) > \hat{a}_{\mathrm{H}_n(x)}(x)$

Hence $a_{\mathrm{H}_n(x)}(x) - a_{\mathrm{H}_a(x)}(x) < a_{\mathrm{H}_n(x)}(x) - \hat{a}_{\mathrm{H}_n(x)}(x) + o_{a.s.}(1)$

$$< \left| \hat{a}_{\mathrm{H}_n(x)}(x) - a_{\mathrm{H}_n(x)}(x) \right| + o_{a.s.}(1) \xrightarrow{a.s.} o$$

Therefore $\mathrm{P}\left(\bigcup_{m \ge n} \left\{ \hat{a}_{\mathrm{H}_m(x)}(x) - a_{\mathrm{H}_a(x)}(x) \ge \eta \right\} \right) \xrightarrow[n \to \infty]{} 0$ and so

$$\mathrm{P}\left(\bigcup_{m \ge n} \left\{ \left\| \mathrm{H}_m(x) - \mathrm{H}_a(x) \right\| \ge \varepsilon \right\} \right) \xrightarrow[n \to \infty]{} 0$$

We proceed similarly to prove (2) by replacing a with A.

Let's define the following:

$$m_{\mathrm{H},x}\left(X_i, X_j\right) = \frac{\alpha}{2} n^{-1} \left|\mathrm{H}\right|^{-1/2} \left(\mathrm{K}_{\mathrm{H}}\left(x - X_i\right) + \mathrm{K}_{\mathrm{H}}\left(x - X_j\right) \right) + \beta L_{\mathrm{H}}\left(x - X_i\right) L_{\mathrm{H}}\left(x - X_j\right)$$

$$\mathrm{P}_{\mathrm{H}}\left(X_i - X_j\right) = L_{\mathrm{H}} * L_{\mathrm{H}}\left(X_i - X_j\right)$$

$$m_{\mathrm{H}}\left(X_i, X_j\right) = \alpha\, n^{-1} \left|\mathrm{H}\right|^{-1/2} S(\mathrm{K}) + \beta\, \mathrm{P}_{\mathrm{H}}\left(X_i - X_j\right)$$

$$V_H(u) = \left| \mathrm{H} \right|^{-1/2} vec((\mathrm{H}^{-1/2} u) \vec{\nabla} L(\mathrm{H}^{-1/2}(u))')$$

$$W_H(u) = |\mathrm{H}|^{-1/2}\, vec((\mathrm{H}^{-1/2}u)\vec{\nabla}P(\mathrm{H}^{-1/2}(u))')$$

Hence $\hat{a}_{\mathrm{H}}(x) = \binom{n}{2}^{-1} \sum_{i<j}\sum m_{\mathrm{H},x}(x_i, x_j)$ and $\hat{\mathrm{A}}_{\mathrm{H}}(x) = \binom{n}{2}^{-1} \sum_{i<j}\sum m_{\mathrm{H}}(x_i, x_j)$

Also let $\otimes$ denote Kroenecker product.

Lemma 3:

$$\mathrm{E}\left(\frac{\partial m_{\mathrm{H},X}(X_i, X_j)}{\partial vec(\mathrm{H})}\right)^2\Bigg|_{\mathrm{H}=\mathrm{H}_n(x)} = 2\beta^2\left(\frac{\beta\mu_2^2(L)n}{\alpha S(\mathrm{K})}\right)^{\frac{2d+4}{d+4}} \mathrm{B}(x,f)(1+o(1))$$

$$\mathrm{E}\left(\frac{\partial m_{\mathrm{H}}(X_i, X_j)}{\partial vec(\mathrm{H})}\right)^2\Bigg|_{\mathrm{H}=\mathrm{H}_n} = \beta^2\left(\frac{\beta\mu_2^2(L)n}{\alpha S(\mathrm{K})}\right)\mathrm{B}(f)(1+o(1))$$

where

$$B(x,f) = |A(x,f)|^{-1}\left(\left(\mathrm{I}\otimes A(x,f)\right)+\left(A(x,f)^{1/2}\otimes A(x,f)^{1/2}\right)\right)^{-1}[(\textstyle\int L^2(u)du)(\int V(u)(V(u))'du)$$
$$+(\textstyle\int L(u)V(u)du)(\int L(u)V(u)du)']\left(\left(\mathrm{I}\otimes A(x,f)\right)+\left(A(x,f)^{1/2}\otimes A(x,f)^{1/2}\right)\right)^{-1}$$

and

$$B(f) = |A(f)|^{-1}\left(\left(\mathrm{I}\otimes A(f)\right)+\left(A(f)^{1/2}\otimes A(f)^{1/2}\right)\right)^{-1}[(\textstyle\int P(u)(P(u))'du)f(x)]\left(\left(\mathrm{I}\otimes A(f)\right)+\left(A(f)^{1/2}\otimes A(f)^{1/2}\right)\right)^{-1}$$

Proof:

$$m_{\mathrm{H},x}\left(X_i,X_j\right)=\frac{\alpha}{2}n^{-1}\left|\mathrm{H}\right|^{-1}\left[\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)+\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right]+\beta\left|\mathrm{H}\right|^{-1}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)$$

$$\frac{\partial m_{\mathrm{H},x}\left(x_i,x_j\right)}{\partial vec\left(\mathrm{H}\right)}=\frac{\alpha}{2}n^{-1}\left\{-\left|\mathrm{H}\right|^{-2}vec\left(\mathrm{H}_{\#}\right)\left[\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)+\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right]\right.$$

$$+\left|\mathrm{H}\right|^{-1}\left[\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right.$$

$$\left.\left.+\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right]\right\}$$

$$+\beta\left\{-\left|\mathrm{H}\right|^{-2}vec\left(\mathrm{H}_{\#}\right)L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right.$$

$$+\left|\mathrm{H}\right|^{-1}\left[\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right.$$

$$\left.\left.+\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right)L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)\right\}$$

The above partial derivatives are computed as follows:

$$\frac{\partial vec\left(\mathrm{H}^{-1/2}u\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}=u'\otimes\mathrm{I}_d$$

$$\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec(\mathrm{H})}=\frac{\partial vec\left(\mathrm{H}^{1/2}\right)}{\partial vec(\mathrm{H})}\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}^{1/2}\right)}=-\left(\left(\mathrm{I}\otimes\mathrm{H}^{1/2}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{I}\right)\right)^{-1}\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)^{-1}$$

$$=-\left(\left(\mathrm{H}^{1/2}\otimes\mathrm{H}\right)+\left(\mathrm{H}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}=-\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left(\mathrm{H}^{-1/2}\otimes\mathrm{I}\right)$$

Since the first term in $\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})}$ is negligible compared to the second we get:

$$\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})}=-\beta\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\Big[L_{\mathrm{H}}\left(x-X_j\right)\left|\mathrm{H}\right|^{-1/2}vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)$$

$$+L_{\mathrm{H}}\left(x-X_i\right)\left|\mathrm{H}\right|^{-1/2}vec\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right)'\Big]$$

and then:

$$\mathrm{E}\left(\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})}\right)^2=2\beta^2\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\{\mathrm{E}\,L_{\mathrm{H}}^2\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\times$$

$$\mathrm{E}\left[\left|\mathrm{H}\right|^{-1}vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)\right]$$

$$+\mathrm{E}\left\{L_H\left(x-X_i\right)L_H\left(x-X_j\right)\left|\mathrm{H}\right|^{-1}vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)\right.$$

$$\left.\times vec'\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right)'\right)\right]\Big]\Big\}\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left(1+o(1)\right)$$

Let $V_H(u)=|\mathrm{H}|^{-1/2}\,vec((\mathrm{H}^{-1/2}u)\vec{\nabla}L(\mathrm{H}^{-1/2}(u))')$, then:

$$E(V_H(x-X_i)(V_H(x-X_i))')=|\mathrm{H}|^{-1/2}\,(\int V(u)(V(u))'du)\,f(x)(1+o(1))$$

and $E(L_H(x-X_i)V_H(x-X_i)) = |\mathrm{H}|^{-1/2} (\int L(u)V(u)du) f(x)(1+o(1))$

also $E(L^2{}_H(x-X_i)) = |\mathrm{H}|^{-1/2} (\int L^2(u)du) f(x)(1+o(1))$

Therefore we obtain the following:

$$\mathrm{E}\left(\left.\frac{\partial m_{\mathrm{H},X}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})}\right|_{\mathrm{H}=\mathrm{H}_n(x)}\right)^2 = 2\beta^2\left(\frac{\alpha S(\mathrm{K})}{\beta\mu_2^2(L)n}\right)^{\frac{-2(d+2)}{d+4}} \mathrm{B}(x,f)\big(1+o(1)\big)$$

$$= 2\beta^2\left(\frac{\alpha S(\mathrm{K})}{\beta\mu_2^2(L)n}\right)^{\frac{-2d-4}{d+4}} \mathrm{B}(x,f)\big(1+o(1)\big)$$

$$= 2\beta^2\left(\frac{\beta\mu_2^2(L)n}{\alpha S(\mathrm{K})}\right)^{\frac{2d+4}{d+4}} \mathrm{B}(x,f)\big(1+o(1)\big)$$

Let $\mathrm{P}_\mathrm{H} = L_\mathrm{H} * L_\mathrm{H}$ then $m_\mathrm{H}\left(X_i,X_j\right) = \alpha n^{-1}|\mathrm{H}|^{-1/2} S\mathrm{K} + \beta\mathrm{P}_\mathrm{H}\left(X_i - X_j\right)$

So $$\frac{\partial m_\mathrm{H}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})} = -\frac{1}{2}\alpha n^{-1}|\mathrm{H}|^{-3/2} vec(\mathrm{H}_\#)S(\mathrm{K}) + \beta\Big\{|\mathrm{H}|^{-3/2} vec(\mathrm{H}_\#)\mathrm{P}\Big(\mathrm{H}^{-1/2}\left(X_i - X_j\right)\Big)$$

$$+|\mathrm{H}|^{-1/2}\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec(\mathrm{H})}\frac{\partial vec\Big(\mathrm{H}^{-1/2}\left(X_i-X_j\right)\Big)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}\mathrm{P}\Big(\mathrm{H}^{-1/2}\left(X_i-X_j\right)\Big)\Big\}$$

Similarly

$$\mathrm{E}\left(\frac{\partial m_\mathrm{H}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})}\right)^2 = \beta^2\Big(\big(\mathrm{I}\otimes\mathrm{H}\big)+\big(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\big)\Big)^{-1}\mathrm{E}\Big\{vec\Big(|\mathrm{H}|^{-1/2}\mathrm{H}^{-1/2}\left(X_i-X_j\right)\Big(\vec{\nabla}\mathrm{P}\Big(\mathrm{H}^{-1/2}\left(X_i-X_j\right)\Big)\Big)'\Big)$$

$$vec'\Big(|\mathrm{H}|^{-1/2}\mathrm{H}^{-1/2}\left(X_i-X_j\right)\Big(\vec{\nabla}\mathrm{P}\Big(\mathrm{H}^{-1/2}\left(X_i-X_j\right)\Big)\Big)'\Big)\Big\}\Big(\big(\mathrm{I}\otimes\mathrm{H}\big)+\big(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\big)\Big)^{-1}\big(1+o(1)\big)$$

Since $E(P_H(X_i - X_j)(P_H(X_i - X_j))') = |\mathrm{H}|^{-1/2} (\int P(u)(P(u))'du) f(x)(1+o(1))$,

$$\mathrm{E}\left(\frac{\partial m_{\mathrm{H}}\left(X_i, X_j\right)}{\partial vec(\mathrm{H})}\bigg|_{\mathrm{H}=\mathrm{H}_n}\right)^2 = \beta^2\left(\frac{\alpha S(\mathrm{K})}{\beta\mu_2^2(L)n}\right)^{\frac{-d-4}{d+4}} \mathrm{B}(f)(1+o(1)) = \beta^2\left(\frac{\beta\mu_2^2(L)n}{\alpha S(\mathrm{K})}\right)\mathrm{B}(f)(1+o(1))$$

Lemma 4:

$$\frac{\partial a_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})} = \beta\mu_2^2(L)C(x,f)$$

$$\frac{\partial A_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})} = \beta\mu_2^2(L)C(f)$$

where

$$C(x,f) = \frac{1}{4}vec\left(A(x,f)^{-1}\right)vec'\left(A(x,f)^{-1}\right) + \frac{1}{2}A(x,f)^{-1}\otimes A(x,f)^{-1} + \varphi_f(x)$$

and

$$C(f) = \frac{1}{4}vec\left(A(f)^{-1}\right)vec'\left(A(f)^{-1}\right) + \frac{1}{2}A(f)^{-1}\otimes A(f)^{-1} + \Psi_f(x)$$

Proof:

$$\begin{aligned}\frac{\partial a_{\mathrm{H}}(x)}{\partial vec'(\mathrm{H})} &= -\frac{\alpha}{2}n^{-1}S(\mathrm{K})f(x)|\mathrm{H}|^{-3/2}vec'(\mathrm{H}_{\#}) + \frac{\beta}{2}\mu_2^2(L)vec'(\mathrm{H})\varphi_f(x)\\ &= -\frac{\alpha}{2}n^{-1}S(\mathrm{K})f(x)|\mathrm{H}|^{-1/2}vec'(\mathrm{H}^{-1}) + \frac{\beta}{2}\mu_2^2(L)vec'(\mathrm{H})\varphi_f(x)\end{aligned}$$

$$\begin{aligned}\frac{\partial a_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})} = &-\frac{\alpha}{2}n^{-1}S(\mathrm{K})f(x)\left(-\frac{1}{2}|\mathrm{H}|^{-1/2}vec(\mathrm{H}^{-1})vec'(\mathrm{H}^{-1}) + |\mathrm{H}|^{-1/2}\left(-\left(\mathrm{H}^{-1}\otimes\mathrm{H}^{-1}\right)\right)\right)\\ &+\beta\mu_2^2(L)\varphi_f(x)\end{aligned}$$

$$=\frac{\alpha}{4}n^{-1}S(\mathrm{K})f(x)|\mathrm{H}|^{-1/2}\left(vec\left(\mathrm{H}^{-1}\right)vec'\left(\mathrm{H}^{-1}\right)+2\left(\mathrm{H}^{-1}\otimes\mathrm{H}^{-1}\right)\right)+\beta\mu_2^2(L)\varphi_f(x)$$

$$=\frac{\alpha}{4}n^{-1}S(\mathrm{K})f(x)\left(\frac{\alpha S(\mathrm{K})f(x)}{\beta\mu_2^2(L)n}\right)^{\frac{-2(d/2+2)}{d+4}}(vec\left(A(x,f)^{-1}\right)vec'\left(A(x,f)^{-1}\right)$$
$$+2\left(A(x,f)^{-1}\otimes A(x,f)^{-1}\right))+\beta\mu_2^2(L)\varphi_f(x)$$

$$=\beta\left(\left(\frac{\mu_2^2(L)S(\mathrm{K})}{4S(\mathrm{K})}\right)\left(vec\left(A(x,f)^{-1}\right)vec'\left(A(x,f)^{-1}\right)\right.\right.$$
$$\left.\left.+2\left(A(x,f)^{-1}\otimes A(x,f)^{-1}\right)\right)+\mu_2^2(L)\varphi_f(x)\right)$$
$$=\beta\mu_2^2(L)C(x,f)$$

Similarly $\frac{\partial \mathrm{E}\hat{\mathrm{A}}_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})}=\beta\left[\frac{\mu_2^2(L)S(\mathrm{K})}{4S(\mathrm{K})}C_1(f)+\mu_2^2(L)\Psi_f\right]=\beta\mu_2^2(L)C(f)$

Lemma 5::

$$n^{\frac{2}{d+4}}\left(vec\left(\mathrm{H}_n(x)\right)-vec\left(\mathrm{H}_a(x)\right)\right)\sim \mathrm{A}N\left(0,\frac{4}{\mu_2^4(L)}\left(\frac{\beta}{\alpha}\right)^{\frac{2d+4}{d+4}}\left(\frac{\mu_2^2(L)}{S(\mathrm{K})}\right)^{\frac{2d+4}{d+4}}C^{-1}(x,f)\mathrm{B}(x,f)C^{-1}(x,f)\right)$$

$$n^{\frac{1}{2}}\left(vec\left(\mathrm{H}_n\right)-vec\left(\mathrm{H}_a\right)\right)\sim \mathrm{A}N\left(0,\frac{2}{\mu_2^4(L)}\left(\frac{\beta}{\alpha}\right)^{\frac{2d+4}{d+4}}\left(\frac{\mu_2^2(L)}{S(\mathrm{K})}\right)^{\frac{2d+4}{d+4}}C^{-1}(f)\mathrm{B}(f)C^{-1}(f)\right)$$

Proof: Since $m_{\mathrm{H},x}$ is continuously differentiable for any *H* in *P* then the Lipschitz condition in Theorem 5.23 (van der Vaart (1998)) is verified by taking the contraction factor to be the supremum of $m_{\mathrm{H},x}$ in a neighborhood of $H_{amse}(x)$. This along with Lemmas 3 and 4 enables us to use theorem 5.23 (van der Vaart (1998)) to demonstrate the above first result. A similar procedure should be used to prove the second.

Multiplying the formulas in Lemma 2 and Lemma 5 by Λ, leads to the following two results:

Theorem 3.5:

1) $\hat{\mathrm{H}}(x) \xrightarrow{a.s.} \mathrm{H}_{mse}(x)$

2) $\hat{\mathrm{H}} \xrightarrow{a.s.} \mathrm{H}_{MISE}$

Theorem 3.6:

$$n^{\frac{2}{d+4}}\left(vec\left(\mathrm{H}_n(x)\right)-vec\left(\mathrm{H}_a(x)\right)\right) \sim \mathrm{A}N\left(0,\frac{4}{\mu_2^4(L)}\left(\frac{\beta}{\alpha}\right)^{\frac{2d+6}{d+4}}\left(\frac{\mu_2^2(L)}{S(\mathrm{K})}\right)^{\frac{2d+6}{d+4}}\left(\frac{R(w)}{\mu_2^2(w)}\right)^{\frac{2}{d+4}}C^{-1}(x,f)\mathrm{B}(x,f)C^{-1}(x,f)\right) \quad (3.13)$$

$$n^{\frac{1}{2}}\left(vec\left(\mathrm{H}_n\right)-vec\left(\mathrm{H}_a\right)\right) \sim \mathrm{A}N\left(0,\frac{2}{\mu_2^4(L)}\left(\frac{\beta}{\alpha}\right)^{\frac{2d+6}{d+4}}\left(\frac{\mu_2^2(L)}{S(\mathrm{K})}\right)^{\frac{2d+6}{d+4}}\left(\frac{R(w)}{\mu_2^2(w)}\right)^{\frac{2}{d+4}}C^{-1}(f)\mathrm{B}(f)C^{-1}(f)\right) \quad (3.14)$$

3.3. Multivariate Distribution Estimation

3.3.1. Introduction

Given iid observations $X_1, \ldots, X_2$ from a d-dimensional population with a probability distribution function F that has continuous second partial derivatives.

The kernel distribution estimator of F at $x \in \mathbb{R}^d$ is defined as: $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} W_H(x - X_i)$, where:

- W(.), the kernel function, is a d-variate probability distribution function with the following characteristics:

1) $\int_{\mathbb{R}^d} x dW(x) = \text{O}_d$

2) $\int_{\mathbb{R}^d} xx' dW(x) = \mu_2(W) I_d$

3) $\int_{\mathbb{R}^d} x dW^2(x) = \mu_1(W^2) 1_d$

4) $\frac{\partial^d W}{\partial u_1 ... \partial u_1} = w(x)$

Where O_d and 1_d are d-dimensional vectors whose entries are respectively 0and 1, and I_d is the (d×d) identity matrix.

- H, the bandwidth matrix, is a positive definite (d×d) matrix.

The notation $W_H(x)$ denotes $W(H^{1/2}x)$ for $x \in \mathbb{R}^d$ where $H^{1/2}$ is a matrix that verifies $H = \left(H^{1/2}\right)' H^{1/2}$.

Just like in the case of univariate estimation, the mean square error (mse) and the mean integrated square error (MISE) will be used to assess the performance of the above estimator and obtain an appropriate smoothing parameter H.

In the remaining of this study we will be using the following functionals:

- $\vec{\nabla}F(x)$, the gradient of F at $x \in \mathbb{R}^d$

- $H_F(x)$, the Hessian of F at $x \in \mathbb{R}^d$

- $m_x(H) = mse(\hat{f}(x))$

3.3.2. Mse Calculations

Under the above conditions of regularity of F and moment condition of W(.) we obtain the following result:

Theorem 3.7:

$$m_x(H) = \frac{F(x)(1-F(x))}{n} - \frac{\mu_1(W^2)}{n} 1_d' H^{1/2} \vec{\nabla}F(x) + \frac{\mu_2^2(W)}{4} tr^2(H^{1/2} H^{1/2}{}' H_F(x)) + o(tr^2(H) + n^{-1} tr(H^{1/2})) \qquad (3.15)$$

Proof:

$$EW_H(x - X_i) = W_H * f(x) = W_H * \frac{\partial^d F}{\partial u_1 \ldots \partial u_d} = \frac{\partial^d W_H}{\partial u_1 \ldots \partial u_d} * F(x) = w_H * F(x)$$

$$= \int_{\mathbb{R}^d} |H|^{-1/2} w(H^{-1/2}(x - y)) F(y) dy$$

$$= \int_{\mathbb{R}^d} w(x) f(x - H^{1/2} z) dz$$

$$= \int_{\mathbb{R}^d} w(x)(F(x) - (H^{-1/2} z)' \vec{\nabla}F(x) \frac{1}{2}(H^{1/2} z)' H_F(x)(H^{1/2} z)) + o(tr(H))$$

$$= F(x) \int_{\mathbb{R}^d} w(z) dz - (\int_{\mathbb{R}^d} z' w(z) dz) H^{1/2} \vec{\nabla}F(x)$$

$$+\frac{1}{2}\int_{\mathbb{R}^d} tr(z'H^{\frac{1}{2}\prime}\ H_F(x)H^{\frac{1}{2}}zw(z))\ dz + o(tr(H))$$

$$= F(x) + \frac{1}{2}tr(H^{\frac{1}{2}\prime}H_F(x)H^{\frac{1}{2}}\int_{\mathbb{R}^d} zz'w(z)dz)\ + o(tr(H))$$

$$= F(x) + \frac{\mu_2(w)}{2}tr(H^{\frac{1}{2}}H^{\frac{1}{2}\prime}H_F(x))\ + o(tr(H))$$

$$EW_H^2(x - X_i) = W_H^2 * f(x) = \frac{\partial^d W_H^2}{\partial u_1 ... \partial u_d} * F(x) = (\frac{\partial^d W_H}{\partial u_1 ... \partial u_d}\frac{\partial W_H^2}{\partial W_H}) * F(x) = 2(w_H W_H) * F(x)$$

$$= 2\int_{\mathbb{R}^d} |H|^{-\frac{1}{2}}\ w(H^{-\frac{1}{2}}(x-y))W(H^{-\frac{1}{2}}(x-y))F(y)dy$$

$$= 2\int_{\mathbb{R}^d} w(x)W(x)F(x + H^{\frac{1}{2}}z)dz$$

$$= F(x)\int_{\mathbb{R}^d} 2w(z)W(z)dz - (\int_{\mathbb{R}^d} 2z'w(z)W(z)dz)H^{\frac{1}{2}}\vec{\nabla}F(x) + o(tr(H^{\frac{1}{2}}))$$

$$= F(x) - \mu_1(W^2)1_d' H^{\frac{1}{2}}\vec{\nabla}F(x) + o(tr(H^{\frac{1}{2}}))$$

By subtracting $E\hat{F}(x) = F(x) + o(tr(H^{\frac{1}{2}}))$ from this last formula, dividing by n and combining with the square of the bias obtained above, we get $m_x(H)$.

3.3.3. MSE-Optimal Bandwidth

We will try to find expressions for the smoothing that minimizes the use under different structure of H.

3.3.3.1 Case 1: $H = h^2 I_d$

In this case the expression for $m_x(H)$ becomes:

$$m_x(H) = \frac{F(x)(1-F(x))}{n} - \frac{\mu_1(W^2)1_d' \vec{\nabla} F(x)}{n} h + \frac{\mu_2^2(W) tr^2(H_F(x))}{4} h^4 + o(h^4 + n^{-1}h^{-1}) \quad (3.16)$$

Leading to an optimal bandwidth matrix:

$$H_{m,x} = \left(\frac{\mu_1(W^2) \sum_{i=1}^{d} \frac{\partial F(x)}{\partial x_i}}{\mu_2^2(W) \left(\sum_{i=1}^{d} \frac{\partial^2 F(x)}{\partial x_i^2} \right)^2} \right)^{2/3} \frac{1}{n^{2/3}} I_d \quad (3.17)$$

3.3.3.2. Case2: $H = diag(h_1^2, ..., h_d^2)$

$$m_x(H) = \frac{F(x)(1-F(x))}{n} - \frac{\mu_1(W^2)}{n} \underline{h_d'} \vec{\nabla} F(x) + \frac{\mu_2^2(W)}{4} \underline{h_d^{2\prime}} (h_F(x) h_F'(x)) \underline{h_d^2} + o(tr^2(\underline{h_d^2}) + n^{-1} tr(\underline{h_d})) \quad (3.18)$$

where $\underline{h_d^2} = [h_1^2, \ldots, h_d^2]'$, $\underline{h_d} = [h_1, \ldots, h_d]'$ and $\underline{h_F(x)} = [\frac{\partial^2 F}{\partial x_1^2}, ..., \frac{\partial^2 F}{\partial x_d^2}]'$

$$\frac{\partial m_x(H)}{\partial \underline{h_d'}} = -\frac{\mu_1(W^2)}{n} \vec{\nabla} F(x) + \mu_2^2(W) \underline{h_d^{2\prime}} (\underline{h_F(x)} \underline{h_F'(x)}) H^{1/2} + o(tr^2(H \underline{h_d^2}) + n^{-1})$$

$\frac{\partial m_x(H)}{\partial \underline{h_d'}} = \mathrm{O}_d$ implies that $\underline{h_F(x)} H^{1/2} = \frac{1}{n} \frac{\mu_1(W^2) \vec{\nabla} F(x)}{\mu_2^2(W) \underline{h_d^{2\prime}} h_F(x)}$

On the other hand, $\underline{h_F'(x)} H^{1/2} = \underline{h_d}' dg(H_F(x)) = \underline{h_d}' diag(\underline{h_F(x)})$ and hence $\underline{h_d}$ can be written as

$\underline{h_d} = \alpha \ dg^{-1}(H_F(x)) \vec{\nabla} F(x)$ which makes the last formula equivalent to:

$$\alpha \ \vec{\nabla}' F(x) = \frac{1}{n} \frac{\mu_1(W^2) \vec{\nabla}' F(x)}{\alpha^2 \mu_2^2(W) (\vec{\nabla}' F(x) dg^{-1}(H_F(x)) \vec{\nabla} F(x))}$$

Therefore

$$H_{m,x} = \left(\frac{\mu_1(W^2)}{\mu_2^2(W)(\vec{\nabla}' F(x) dg^{-1}(H_F(x))\vec{\nabla} F(x))} \right)^{2/3} \frac{1}{n^{2/3}} [diag^{-1}(H_F(x)\vec{\nabla} F(x))]^2 \qquad (3.19)$$

3.3.3.3. The General Case

$m_x(H)$ can be written as follows:

$$m_x(H) = \frac{F(x)(1-F(x))}{n} - \frac{\mu_1(W^2)}{2n}[vec'(\vec{\nabla}F(x)1_d') + vec'(1_d(\vec{\nabla}F(x))')]vec(H^{1/2})$$
$$+ \frac{\mu_2^2(W)}{4} vec'(H^{1/2}H^{1/2\prime})vec(H_F(x))vec'(H_F(x))vec(H^{1/2}H^{1/2\prime}) + o(tr^2(H) + n^{-1}tr(H^{1/2})) \qquad (3.20)$$

where vec symbolizes the vec operator (chapter 7, section 5 Schott(1997)),

$$\frac{\partial vec'(H)(vec(H_F(x)vec'(H_F(x))vec(H)}{\partial vec'(H^{1/2})} = 2vec'(H^{1/2}H^{1/2\prime})(vec(H_F(x)vec'(H_F(x)))\frac{\partial vec(H^{1/2}H^{1/2\prime})}{\partial vec'(H^{1/2})}$$

Let K_{dd} be the commutation matrix corresponding to $(d \times d)$ matrices, $N_d = \frac{1}{2}(I_{d^2} + K_{dd})$ and

$\otimes$ be the symbol of the Kroenecker product. Then $d(H^{1/2}H^{1/2\prime}) = H^{1/2}dH^{1/2\prime} + (dH^{1/2})H^{1/2\prime}$ and

Theorem 7.16 (Schott 1997) imply that:

$$\begin{aligned} dvec(H^{1/2}H^{1/2\prime}) &= (I \otimes H^{1/2})dvec(H^{1/2\prime}) + (H^{1/2} \otimes I)dvec(H^{1/2}) \\ &= (I \otimes H^{1/2})K_{dd}dvec(H^{1/2}) + (H^{1/2} \otimes I)dvec(H^{1/2}) \\ &= (I_{d^2} + K_{dd})(H^{1/2} \otimes I)dvec(H^{1/2}) \\ &= 2N_d(H^{1/2} \otimes I)dvec(H^{1/2}) \end{aligned}$$

$$\begin{aligned} \frac{\partial vec'(H^{1/2}H^{1/2\prime})(vec(H_F(x)vec'(H_F(x))vec(H^{1/2}H^{1/2\prime})}{\partial vec'(H^{1/2})} &= 4vec'(H^{1/2}H^{1/2\prime})(vec(H_F(x)vec'(H_F(x)))N_d(H^{1/2} \otimes I) \\ &= 4tr(H^{1/2}H^{1/2\prime}H_F(x))vec'(H_F(x)H^{1/2}) \end{aligned}$$

Hence, $tr(H^{1/2}H^{1/2\prime}H_F(x))vec(H_F(x)H^{1/2})=\frac{1}{n}\frac{\mu_1(W^2)}{\mu_2^2(W)}N_d vec(\vec{\nabla}F(x)1_d')$

And that is equivalent to: $H^{1/2}=\frac{1}{tr(H^{1/2}H^{1/2\prime}H_F(x))n}\frac{\mu_1(W^2)}{2\mu_2^2(W)}H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')$

And hence $tr(H^{1/2}H^{1/2\prime}H_F(x))=\left(\left(\frac{\mu_1(W^2)}{2\mu_2^2(W)n}\right)^2 tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')^2)\right)^{1/3}$.

In case $tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')^2)$ is negative, the above expression represents the real cubic root.

So the mse-optimal bandwidth matrix is:

$$H_{m,x}=\left(\frac{\mu_1(W^2)}{2\mu_2^2(W)tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')^2)}\right)^{2/3}\frac{1}{n^{2/3}}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')(H_F(x)^2)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)') \quad (3.21)$$

3.3.4. Bandwidth Estimation

Define the statistic $\hat{a}_H(x)$ as follows:

$$\hat{a}_H(x)=\alpha n^{-2}\sum_{i=1}^{n}K_H^2(x-X_i)+\beta\binom{n}{2}^{-1}\sum\sum_{i<j}L_H(x-X_i)L_H(x-X_j) \quad (3.22)$$

where K and L are symmetric functions that verify:

5. $\int_{IR^d}u\ dK(u)=\mu_1(K)1_d$

6. $\int_{IR^d}u\ dL(u)=O_d$

7. $\int_{IR^d}uu'dL(u)=\mu_2(L)I_d$

If we define $a_{\mathrm{H}}(x)$ to be $\mathrm{E}\,\hat{a}_{\mathrm{H}}(x)$ and proceed similarly as we did to calculate $m_x(H)$, we can easily obtain the following result:

Theorem 3.8:

$$a_H(x) = \frac{\alpha S(K)F(x)}{n} - \alpha\frac{\mu_1(K^2)}{n}1_d' H^{1/2}\vec{\nabla}F(x) + \beta\frac{\mu_2^2(L)}{4}tr^2(H^{1/2}H^{1/2\prime}H_F(x)) + o(tr^2(H) + n^{-1}tr(H^{1/2})) \quad (3.23)$$

Where $S(K) = \int dK^2(u)$

If $H_m(x)$ and $H_a(x)$ are the bandwidth matrices that minimize $m_x(H)$ and $a_{\mathrm{H}}(x)$ with respect to H, the following result is straight forward:

Theorem 3.9:

For any $x \in IR^d$ such that $\left|H_f(x)\right| \neq 0$, $\qquad H_m(x) = \Lambda H_a(x)$, (3.24)

where $\Lambda = \left(\dfrac{\beta}{\alpha}\dfrac{\mu_1(W^2)\mu_2^2(L)}{\mu_1(K^2)\mu_2^2(W)}\right)^{1/3}$

3.3.5. Asymptotic Properties:

Let $\mathrm{H}_n(x)$ be the bandwidth matrices that minimize $\hat{a}_{\mathrm{H}}(x)$, and based on the previous result, we define $\hat{\mathrm{H}}(x) = \Lambda \mathrm{H}_n(x)$ as a potential estimator of $\mathrm{H}_m(x)$. Also, let P be the set of all ($d \times d$) positive definite matrices.

So, the problem at hand is that of optimizing a risk function, so we use available results from the theory of M-estimation to find the asymptotic behavior of $\mathrm{H}_n(x)$ and hence that of $\hat{\mathrm{H}}(x)$.

Lemma 1: $\sup_{\mathrm{H}\in \mathrm{P}}\left|\hat{a}_{\mathrm{H}}(x)-a_{\mathrm{H}}(x)\right| \xrightarrow{a.s.} 0, \qquad \forall x \in IR^d$ (3.25)

Proof: The above result follows directly from the SLLN of U-Statistics.

Lemma 2: $\mathrm{H}_n(x) \xrightarrow{a.s.} \mathrm{H}_a(x), \qquad \forall x \in IR^d$ (3.26)

Proof: By definitions of $\mathrm{H}_a(x): \forall \varepsilon > 0 \quad \inf_{\mathrm{H}\in \mathrm{P}:\left\|\mathrm{H}-\mathrm{H}_a(x)\right\|\geq \varepsilon} a_{\mathrm{H}}(x) \geq a_{\mathrm{H}_a(x)}(x)$

Then $\forall \varepsilon > 0 \;\; \exists \eta > 0$ s.t. $\left\|\mathrm{H}-\mathrm{H}_a(x)\right\| \geq \varepsilon \Rightarrow a_{\mathrm{H}}(x) - a_{\mathrm{H}_a(x)}(x) \geq \eta$

Thus the event $\left\{\left\|\mathrm{H}_n(x)-\mathrm{H}_a(x)\right\| \geq \varepsilon\right\}$ is contained in $\left\{a_{\mathrm{H}_n}(x) - a_{\mathrm{H}_a(x)}(x) \geq \eta\right\}$

Also, by Lemma 1, $\hat{a}_{\mathrm{H}_a(x)}(x) = a_{\mathrm{H}_a(x)} + \circ_{a.s.}(1)$ which is greater than $\hat{a}_{\mathrm{H}_n(x)}(x)$

$$a_{\mathrm{H}_n(x)}(x) - a_{\mathrm{H}_a(x)}(x) < a_{\mathrm{H}_n(x)}(x) - \hat{a}_{\mathrm{H}_n(x)}(x) + \circ_{a.s.}(1)$$

$$\left|\hat{a}_{\mathrm{H}_n(x)}(x) - a_{\mathrm{H}_n(x)}(x)\right| + \circ_{a.s.}(1) \xrightarrow{a.s.} \circ$$

Hence $\mathrm{P}\left(\bigcup_{m\geq n}\left\{\hat{a}_{\mathrm{H}_m(x)}(x) - a_{\mathrm{H}_a(x)}(x) \geq \eta\right\}\right) \xrightarrow[n\to\infty]{} 0$ and so

$$\mathrm{P}\left(\bigcup_{m\geq n}\left\{\left\|\mathrm{H}_m(x)-\mathrm{H}_a(x)\right\| \geq \varepsilon\right\}\right) \xrightarrow[n\to\infty]{} 0$$

Let's define $m_{\mathrm{H},x}\left(X_i, X_j\right)$ as follows:

$$m_{\mathrm{H},x}\left(X_i, X_j\right) = \frac{\alpha}{2} n^{-1}\left(\mathrm{K}_{\mathrm{H}}\left(x - X_i\right) + \mathrm{K}_{\mathrm{H}}\left(x - X_j\right)\right) + \beta L_{\mathrm{H}}\left(x - X_i\right) L_{\mathrm{H}}\left(x - X_j\right)$$

Hence $\hat{a}_{\mathrm{H}}\left(x\right) = \binom{n}{2}^{-1} \sum\sum_{i<j} m_{\mathrm{H},x}\left(x_i, x_j\right)$

Also, let $V_H(u) = vec((\mathrm{H}^{-1/2}u)\vec{\nabla}L(\mathrm{H}^{-1/2}(u))')$.

In what remains, $H^{1/2}$ will denote the symmetric square root of H.

Lemma 3:

$$\mathrm{E}\left(\left.\frac{\partial m_{\mathrm{H},x}\left(X_i, X_j\right)}{\partial vec\left(\mathrm{H}\right)}\right|_{H=H_{m,x}}\right)^2 = 2\beta^2\left(\frac{\alpha}{\beta}\frac{\mu_1(K^2)}{2\mu_2^2(L)\,|\,tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d' + 1_d\vec{\nabla}F(x)')^2)\,|}\right)^{\frac{2d-4}{3}} \frac{1}{n^{\frac{2d-4}{3}}}\,|\,A(x,F)\,|\,f^2(x)B(x,F)B(L)B(x,F)\left(1+o\left(1\right)\right) \quad (3.27)$$

Where:

$$A(x,F) = (\vec{\nabla}F(x)1_d' + 1_d\vec{\nabla}F(x)')(H_F(x)^2)^{-1}(\vec{\nabla}F(x)1_d' + 1_d\vec{\nabla}F(x)')$$

$$B(x,F) = \left(\left(\mathrm{I} \otimes A(x,F)\right) + \left(A(x,F)^{1/2} \otimes A(x,F)^{1/2}\right)\right)^{-1}$$

$$B(L) = \left[\left(\int L^2(u)du\right)\left(\int V(u)(V(u))'du\right) + \left(\int L(u)V(u)du\right)\left(\int L(u)V(u)du\right)'\right]$$

Proof:

$$m_{\mathrm{H},x}\left(X_i, X_j\right) = \frac{\alpha}{2} n^{-1}\left[\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x - X_i\right)\right) + \mathrm{K}\left(\mathrm{H}^{-1/2}\left(x - X_j\right)\right)\right] + \beta L\left(\mathrm{H}^{-1/2}\left(x - X_i\right)\right) L\left(\mathrm{H}^{-1/2}\left(x - X_j\right)\right)$$

$$\frac{\partial m_{\mathrm{H},x}\left(x_i,x_j\right)}{\partial vec\left(\mathrm{H}\right)}=\frac{\alpha}{2}n^{-1}\Bigg[\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)$$

$$+\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}\mathrm{K}\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\Bigg]$$

$$+\beta\Bigg[\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)$$

$$+\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\Bigg]$$

$$\frac{\partial vec\left(\mathrm{H}^{-1/2}u\right)}{\partial vec\left(\mathrm{H}^{-1/2}\right)}=u'\otimes\mathrm{I}_d$$

$$\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}\right)}=\frac{\partial vec\left(\mathrm{H}^{1/2}\right)}{\partial vec\left(\mathrm{H}\right)}\frac{\partial vec\left(\mathrm{H}^{-1/2}\right)}{\partial vec\left(\mathrm{H}^{1/2}\right)}=-\left(\left(\mathrm{I}\otimes\mathrm{H}^{1/2}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{I}\right)\right)^{-1}\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)^{-1}$$

$$=-\left(\left(\mathrm{H}^{1/2}\otimes\mathrm{H}\right)+\left(\mathrm{H}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}=-\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left(\mathrm{H}^{-1/2}\otimes\mathrm{I}\right)$$

Since the first term of the sum in $\frac{\partial m_{\mathrm{H},x}\left(x_i,x_j\right)}{\partial vec\left(\mathrm{H}\right)}$ is negligible compared to the second, we get:

$$\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec\left(\mathrm{H}\right)}=-\beta\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left[L_{\mathrm{H}}(x-X_j)vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)\right.$$

$$\left.+L_{\mathrm{H}}(x-X_i)vec\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right)'\right)\right]$$

And hence:

$$\mathrm{E}\left(\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec\left(\mathrm{H}\right)}\right)^2 = 2\beta^2\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}$$

$$EL_{\mathrm{H}}^2\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\mathrm{E}\left[vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)vec'\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)\right]$$

$$+\mathrm{E}\left[L_H\left(x-X_i\right)L_H\left(x-X_j\right)vec\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_i\right)\right)\right)'\right)vec'\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\left(\vec{\nabla}L\left(\mathrm{H}^{-1/2}\left(x-X_j\right)\right)\right)'\right)\right]$$

$$\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left(1+o\left(1\right)\right)$$

$$E(V_H(x-X_i)(V_H(x-X_i))') = |\mathrm{H}|^{1/2}\left(\int V(u)(V(u))'du\right)f(x)(1+o(1))$$

and $E(L_H(x-X_i)V_H(x-X_i)) = |\mathrm{H}|^{1/2}\left(\int L(u)V(u)du\right)f(x)(1+o(1))$

also $E(L^2{}_H(x-X_i)) = |\mathrm{H}|^{1/2}\left(\int L^2(u)du\right)f(x)(1+o(1))$

therefore:

$$\mathrm{E}\left(\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec\left(\mathrm{H}\right)}\right)^2 = 2\beta^2\,|\mathrm{H}|\,f^2(x)\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left[\left(\int L^2(u)du\right)\left(\int V(u)(V(u))'du\right)\right.$$

$$\left.+\left(\int L(u)V(u)du\right)\left(\int L(u)V(u)du\right)'\right]\left(\left(\mathrm{I}\otimes\mathrm{H}\right)+\left(\mathrm{H}^{1/2}\otimes\mathrm{H}^{1/2}\right)\right)^{-1}\left(1+o\left(1\right)\right)$$

Since the optimal $\mathrm{H}^{1/2}$ is:

$$H_{m,x}^{1/2} = \left(\frac{\alpha}{\beta}\frac{\mu_1(K^2)}{2\mu_2^2(L)\,|\,tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')^2)\,|}\right)^{1/3}\frac{1}{n^{1/3}}[A(x,F)]^{1/2}$$

Then

$$\mathrm{E}\left(\left.\frac{\partial m_{\mathrm{H},x}\left(X_i,X_j\right)}{\partial vec(\mathrm{H})}\right|_{H=H_{m,x}}\right)^2=2\beta^2\left(\frac{\alpha}{\beta}\frac{\mu_1(K^2)}{2\mu_2^2(L)\,|\,tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')^2)\,|}\right)^{\frac{2d-4}{3}}\frac{1}{n^{\frac{2d-4}{3}}}|A(x,F)|\,f^2(x)$$

$$\left(\left(\mathrm{I}\otimes A(x,F)\right)+\left(A(x,F)^{1/2}\otimes A(x,F)^{1/2}\right)\right)^{-1}\left[\left(\int L^2(u)du\right)\left(\int V(u)(V(u))'du\right)\right.$$

$$\left.+\left(\int L(u)V(u)du\right)\left(\int L(u)V(u)du\right)'\right]\left(\left(\mathrm{I}\otimes A(x,F)\right)+\left(A(x,F)^{1/2}\otimes A(x,F)^{1/2}\right)\right)^{-1}\left(1+o(1)\right)$$

Lemma 4:

$$\left.\frac{\partial a_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})}\right|_{H=H_{m,x}}=\beta\mu_2^2(L)C(x,F)+o(1) \tag{3.28}$$

Where

$$C(x,F)=\left\{2\,|\,tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d'+1_d\vec{\nabla}F(x)')^2)\,|\left(I_{d^2}\otimes vec'(1_d'\vec{\nabla}F(x))\right)\right.$$

$$\left[\left(\left(I_d\otimes[A(x,F)]^{1/2}\right)+\left([A(x,F)]^{1/2}\otimes I_d\right)\right)\otimes\left(\left(I_d\otimes[A(x,F)]^{1/2}\right)+\left([A(x,F)]^{1/2}\otimes I_d\right)\right)\right]^{-1}$$

$$\left(I_d\otimes K_{dd}\otimes I_d\right)\left[(vec(I_d)\otimes I_{d^2})+(I_{d^2}\otimes vec(I_d))\right]\left(\left(I_d\otimes[A(x,F)]^{1/2}\right)+\left([A(x,F)]^{1/2}\otimes I_d\right)\right)^{-1}$$

$$\left.+\frac{1}{2}\varphi_f(x)\right\}$$

Proof:

$$a_{\mathrm{H}}(x)=\alpha\frac{F(x)}{n}-\alpha n^{-1}\mu_1\left(\mathrm{K}^2\right)vec'(1_d'\vec{\nabla}F(x))vec(H^{1/2})+\frac{\beta\mu_2^2(L)}{4}vec'(H)\varphi_f(x)vec(H)+o(tr^2(H)+n^{-1}tr(H))$$

$$\frac{\partial a_{\mathrm{H}}(x)}{\partial vec'(\mathrm{H})}=-\alpha n^{-1}\mu_1\left(\mathrm{K}^2\right)vec'(1_d'\vec{\nabla}F(x))\left(\left(I_d\otimes H^{1/2}\right)+\left(H^{1/2}\otimes I_d\right)\right)^{-1}+\frac{\beta\mu_2^2(L)}{2}vec'(H)\varphi_f(x)+o(tr(H)+n^{-1})$$

By theorem 7.32 (Schott (1997)):

$$vec\left(I_d \otimes \mathrm{H}^{1/2}\right) = (I_d \otimes K_{dd} \otimes I_d)(vec(I_d) \otimes vec(\mathrm{H}^{1/2}))$$

$$= (I_d \otimes K_{dd} \otimes I_d)(vec(I_d)1 \otimes I_{d^2} vec(\mathrm{H}^{1/2}))$$

So by theorem 7.7 (Schott (1997)):

$$vec\left(I_d \otimes \mathrm{H}^{1/2}\right) = (I_d \otimes K_{dd} \otimes I_d)(vec(I_d) \otimes vec(\mathrm{H}^{1/2}))$$

$$= (I_d \otimes K_{dd} \otimes I_d)(vec(I_d) \otimes I_{d^2}) vec(\mathrm{H}^{1/2})$$

Similarly,

$$vec\left(\mathrm{H}^{1/2} \otimes I_d\right) = (I_d \otimes K_{dd} \otimes I_d)(vec(\mathrm{H}^{1/2}) \otimes vec(I_d))$$

$$= (I_d \otimes K_{dd} \otimes I_d)(I_{d^2} vec(\mathrm{H}^{1/2}) \otimes vec(I_d)1)$$

$$= (I_d \otimes K_{dd} \otimes I_d)(I_{d^2} \otimes vec(I_d)) vec(\mathrm{H}^{1/2})$$

Therefore:

$$\frac{\partial a_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})} = \alpha n^{-1} \mu_1(\mathrm{K}^2)\left(I_{d^2} \otimes vec'(1_d' \vec{\nabla} F(x))\right)\left[\left(\left(I_d \otimes \mathrm{H}^{1/2}\right) + \left(\mathrm{H}^{1/2} \otimes I_d\right)\right) \otimes \left(\left(I_d \otimes \mathrm{H}^{1/2}\right) + \left(\mathrm{H}^{1/2} \otimes I_d\right)\right)\right]^{-1}$$

$$\left(I_d \otimes K_{dd} \otimes I_d\right)\left[(vec(I_d) \otimes I_{d^2}) + (I_{d^2} \otimes vec(I_d))\right]\left(\left(I_d \otimes \mathrm{H}^{1/2}\right) + \left(\mathrm{H}^{1/2} \otimes I_d\right)\right)^{-1}$$

$$+ \frac{\beta \mu_2^2(L)}{2} \varphi_f(x) + o(1)$$

The optimal $\mathrm{H}^{1/2}$ is: $H_{m,x}^{1/2} = \left(\frac{\alpha}{\beta} \frac{\mu_1(K^2)}{2\mu_2^2(L) \,|\, tr(H_F(x)^{-1}(\vec{\nabla} F(x) 1_d' + 1_d \vec{\nabla} F(x)')^2) \,|} \right)^{1/3} \frac{1}{n^{1/3}} [A(x,F)]^{1/2}$, so

$$\left.\frac{\partial a_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})}\right|_{H=H_{m,x}} = \alpha n^{-1} \mu_1(\mathrm{K}^2)\left(\frac{\beta}{\alpha}\frac{2\mu_2^2(L) \,|\, tr(H_F(x)^{-1}(\vec{\nabla} F(x) 1_d' + 1_d \vec{\nabla} F(x)')^2) \,|\, n}{\mu_1(K^2)}\right)\left(I_{d^2} \otimes vec'(1_d' \vec{\nabla} F(x))\right)$$

$$\left[\left(\left(I_d \otimes [A(x,F)]^{1/2}\right) + \left([A(x,F)]^{1/2} \otimes I_d\right)\right) \otimes \left(\left(I_d \otimes [A(x,F)]^{1/2}\right) + \left([A(x,F)]^{1/2} \otimes I_d\right)\right)\right]^{-1}$$

$$\left(I_d \otimes K_{dd} \otimes I_d\right)\left[(vec(I_d) \otimes I_{d^2}) + (I_{d^2} \otimes vec(I_d))\right]\left(\left(I_d \otimes [A(x,F)]^{1/2}\right) + \left([A(x,F)]^{1/2} \otimes I_d\right)\right)^{-1}$$

$$+ \frac{\beta \mu_2^2(L)}{2} \varphi_f(x) + o(1)$$

$$\left.\frac{\partial a_{\mathrm{H}}(x)}{\partial vec(\mathrm{H})\partial vec'(\mathrm{H})}\right|_{H=H_{m,x}} = \beta\mu_2^2(L)\{2|tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d' + 1_d\vec{\nabla}F(x)')^2)|\left(I_{d^2}\otimes vec'(1_d'\vec{\nabla}F(x))\right)$$

$$\left[\left(\left(I_d\otimes[A(x,F)]^{1/2}\right)+\left([A(x,F)]^{1/2}\otimes I_d\right)\right)\otimes\left(\left(I_d\otimes[A(x,F)]^{1/2}\right)+\left([A(x,F)]^{1/2}\otimes I_d\right)\right)\right]^{-1}$$

$$\left(I_d\otimes K_{dd}\otimes I_d\right)\left[(vec(I_d)\otimes I_{d^2})+(I_{d^2}\otimes vec(I_d))\right]\left(\left(I_d\otimes[A(x,F)]^{1/2}\right)+\left([A(x,F)]^{1/2}\otimes I_d\right)\right)^{-1}$$

$$+\frac{1}{2}\varphi_f(x)\} + o(1)$$

Lemma 5:

$$\left(\frac{\beta}{\alpha}\right)^{\frac{d-2}{3}} n^{\frac{d+1}{3}} vec(H_n(x)-H_a(x))$$

$$\sim \mathrm{AN}\left(0_{d^2}, 4\mu_2^{-4}(L)\left(\frac{\mu_1(K^2)}{2\mu_2^2(L)|tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d' + 1_d\vec{\nabla}F(x)')^2)|}\right)^{\frac{2d-4}{3}} D(x,F)\right) \tag{3.29}$$

Proof:

Since $m_{\mathrm{H},x}$ is continuously differentiable for any *H* in *P* then the Lipschitz condition in theorem 5.23 (van der Vaart (1998)) is verified by taking the contraction factor to be the supremum of $m_{\mathrm{H},x}$ in a neighborhood of $H_a(x)$. This along with Lemmas 3 and 4 enables us to use theorem 5.23 (van der Vaart (1998)) to demonstrate the above result.

Theorem 3.10:

For every $x \in IR^d$,

$$\hat{\mathrm{H}}(x) \xrightarrow{a.s.} \mathrm{H}_m(x), \tag{3.30}$$

Proof: Multiplying both sides of the limit in Lemma 1 by Λ , leads to the above result.

Theorem 3.11:

$$\left(\frac{\beta}{\alpha}\right)^{\frac{d-2}{3}} n^{\frac{d+1}{3}} vec(H_n(x)-H_a(x))$$
$$\sim \mathrm{AN}\left(0_{d^2}, 4\mu_2^{-4}(L)\left(\frac{\mu_1(W^2)\mu_2^2(L)}{\mu_1(K^2)\mu_2^2(W)}\right)^{2/3}\left(\frac{\mu_1(K^2)}{2\mu_2^2(L)\,|\,tr(H_F(x)^{-1}(\vec{\nabla}F(x)1_d' + 1_d\vec{\nabla}F(x)')^2)\,|}\right)^{\frac{2d-4}{3}} D(x,F)\right) \tag{3.31}$$

Proof:

Multiplying the right hand side of the formula in Lemma 5 by Λ, leads to the above result.

CHAPTER FOUR: CONCLUSION

4.1. Results Summary

As seen from the developments in the previous chapters, the current research was able to achieve the following:

- Build up a new methodology for bandwidth estimation that is as data-based as cross-validatory techniques and faster than the plug-in techniques.
- Extend some of the usual results to the multivariate case.
- Improve the plug-in technique through completely data-driven estimation of the integrated squared derivative.
- Come up with a new estimate for the location parameter via kernels.

4.2. Future Research

Based on the above findings, it is intended that future work will be directed towards the following:

- Extend the obtained results to regression, time series and dependent data.
- Improve the transformation-based estimation by using our distribution estimate.
- Elaborate a kernel based test of hypothesis for the one-sample and two-sample location problems.

LIST OF REFERENCES

(1) Ahmad, I. A. (1976), ``On asymptotic properties of an estimate of a functional of a probability density", **Scandinavian Actuarial Journal**, **0** , 176-182

(2) Ahmad, I. A. (1982), ``Nonparametric estimation of the location and scale parameters based on density estimation", **Annals of the Institute of Statistical Mathematics**, **34** , 39-53

(3) Ahmad, I. A. and Mugdadi, A. R. (To appear in 2005), "A weighted Hellinger distance as an error criterion for bandwidth selection in kernel estimation", **Journal of Nonparametric Statistics**

(4) Ahmad, I. A. and Ran, I. (2004), "Data based bandwidth selection in kernel density estimation with parametric start via kernel contrasts", **Journal of Nonparametric Statistics**, 16, 841-877

(5) Altman, Naomi and Léger, Christian (1995), ``Bandwidth selection for kernel distribution function estimation", **Journal of Statistical Planning and Inference**, **46** , 195-214

(6) Bickel, P. J., and Ritov, Y. (1988), ``Estimating integrated squared density derivatives: Sharp best order of convergence estimates", **Sankhya, Series A, Indian Journal of Statistics**, **50** , 381-393

(7) Bowman, Adrian W. (1984), ``An alternative method of cross-validation for the smoothing of density estimates", **Biometrika**, **71** , 353-360

(8) Bowman, Adrian , Hall, Peter , and Prvan, Tania (1998), ``Bandwidth selection for the smoothing of distribution functions", **Biometrika**, **85** , 799-808

(9) Chiu, Shean-Tsong (1991), ``Bandwidth selection for kernel density estimation", **The Annals of Statistics**, **19** , 1883-1905

(10) Chiu, Shean-Tsong (1992), ``An automatic bandwidth selector for kernel density estimation'', **Biometrika**, **79** , 771-782

(11) Devroye, Luc and Györfi, Laszlo (1985), ``Nonparametric density estimation: the L1 view'', John Wiley & Sons (New York; Chichester)

(12) Engel, J., Herrmann, E., and Gasser, T. (1994), ``An iterative bandwidth selector for kernel estimation of densities and their derivatives'', **Journal of Nonparametric Statistics**, **4** , 21-34

(13) Hall, Peter (1983), ``Large sample optimality of least squares cross-validation in density estimation'', **The Annals of Statistics**, **11** , 1156-1174

(14) Hall, Peter , and Marron, James S. (1987), ``Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation'', **Probability Theory and Related Fields, 74** , 567-581

(15) Hall, Peter, and Marron, J. S. (1987), ``Estimation of integrated squared density derivatives (Corr: V7 p87)'', **Statistics & Probability Letters**, **6** , 109-115

(16) Hall, Peter, and Marron, J. S. (1991), ``Local minima in cross-validation functions (Corr: V54 p907)'', **Journal of the Royal Statistical Society, Series B, Methodological**, **53** , 245-252

(17) Hall, Peter , Sheather, Simon J. , Jones, M. C. , and Marron, J. S. (1991), ``On optimal data-based bandwidth selection in kernel density estimation'', **Biometrika**, **78** , 263-269

(18) Jones, M. C. , and Sheather, S. J. (1991), ``Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives'', **Statistics & Probability Letters**, 11 , 511-514

(19) Jones, M. C., Marron, J. S. , and Sheather, S. J. (1996), ``A brief survey of bandwidth selection for density estimation'', **Journal of the American Statistical Association**, 91, 401-407

(20) Kanazawa, Y. (1993), ``Hellinger distance and Kullback-Leibler loss for the kernel density estimator'', **Statistics & Probability Letters**, **18** , 315-321

(21) Loader, Clive R. (1999), ``Bandwidth selection: Classical or plug-in?'', **The Annals of Statistics**, **27** , 415-438

(22) Nadaraya, È. A. (1964), ``Some new estimates for distribution functions'', **Theory of Probability and its Applications**, **9** , 497-500

(23) Park, Byeong U. , and Marron, J. S. (1992), ``On the use of pilot estimators in bandwidth selection'', **Journal of Nonparametric Statistics**, **1** , 231-240

(24) Rosenblatt, M. (1956). "Remarks on some non-parametric estimates of a density function. **Annals of** Mathematical **Statistics**, 27, 642-669.

(25) Rudemo, Mats (1982), ``Empirical choice of histograms and kernel density estimators'', **Scandinavian Journal of Statistics**, **9** , 65-78

(26) Sarda, P. (1993), ``Smoothing parameter selection for smooth distribution functions'', **Journal of Statistical Planning and Inference**, **35** , 65-75

(27) Scott, David W. , and Terrell, George R. (1987), ``Biased and unbiased cross-validation in density estimation'', **Journal of the American Statistical Association**, **82** , 1131-1146

(28) Schott, James R. (1997), ``Matrix analysis for statistics'', John Wiley & Sons (New York; Chichester)

(29) Stone, Charles J. (1980), ``Optimal rates of convergence for nonparametric estimators'', **The Annals of Statistics**, **8** , 1348-1360

(30) Stone, Charles J. (1984), ``An asymptotically optimal window selection rule for kernel density estimates'', **The Annals of Statistics**, **12** , 1285-1297

(31) Vaart, A. W. van der (1998), ``Asymptotic statistics'', Cambridge University Press (Cambridge, UK)

(32) Woodroofe, Michael (1970), ``On choosing a Delta-sequence'', **The Annals of Mathematical Statistics**, **41** , 1665-1671