A Review of Kernel Density Estimation with Applications to Econometrics

ADRIANO Z. ZAMBOM AND RONALDO DIAS

Universidade Estadual de Campinas

December, 2012

Abstract

Nonparametric density estimation is of great importance when econometricians want to model the probabilistic or stochastic structure of a data set. This comprehensive review summarizes the most important theoretical aspects of kernel density estimation and provides an extensive description of classical and modern data analytic methods to compute the smoothing parameter. Throughout the text, several references can be found to the most up-to-date and cut point research approaches in this area, while econometric data sets are analyzed as examples. Lastly, we present SIZer, a new approach introduced by Chaudhuri and Marron (2000), whose objective is to analyze the visible features representing important underlying structures for different bandwidths.

Keywords: nonparametric density estimation; SiZer; plug-in bandwidth selectors; cross-validation; smoothing parameter.

1 Introduction

The field of econometrics focuses on methods that address the probabilistic or stochastic phenomena involving economic data. Modeling the underlying probabilistic structure of the data, i.e., the uncertainty of the process, is a crucial task, for it can be used to describe the mechanism from which the data was generated. Thus, econometricians have widely explored density estimation, both the parametric and nonparametric approaches, to identify these structures and then make inferences about the unknown "true models". A parametric model assumes that the density is known up to a finite number of parameters, while a nonparametric model allows great flexibility in the possible form, usually assuming that it belongs to some infinite collection of curves (differentiable with square integrable second derivatives for example). The most used approach is kernel smoothing, which dates back to Rosenblatt (1956) and Parzen (1962). The aim of this paper is to review the most import aspects of kernel density estimation, both traditional approaches and modern ideas.

A large extent of econometric research concerning estimation of densities has shown that a well estimated density can be extremely useful for applied purposes. An interesting comprehensive review of kernel smoothing and its applications can be found in Bierens (1987). Silverman (1986) and Scott (1992) discuss kernel density estimation thoroughly, giving details about assumptions on the kernel weight, properties of the estimator such as bias and variance, and discusses how to choose the smoothness of the estimate. The choice of the smoothing parameter is a crucial issue in nonparametric estimation, and will be discussed in detail in Section 4.

The remainder of this paper is as follows. In Section 2 we describe the most basic and intuitive method of density estimation: the histogram. Then, in Section 3 we introduce kernel density estimation and the properties of estimators of this type, followed by an overview of old and new bandwidth selection approaches in Section 4. Finally, SiZer, a modern idea for

accessing features that represent important underlying structures through different levels of smoothing, is introduced in Section 5.

2   The Histogram

The grouping of data in the form of a frequency histogram is a classical methodology that is intrinsic to the foundations of a variety of estimation procedures. Providing useful visual information, it has served as a data presentation device, however, as a density estimation method, it has played a fundamental role in nonparametric statistics.

Basically, the histogram is a step function defined by bin heights, which equal the proportion of observations contained in each bin divided by the bin width. The construction of the histogram is very intuitive, and to formally describe this construction, we will now introduce some notation. Suppose we observe random variables $X_1, \ldots, X_n$ i.i.d. from the distribution function F_X, and that F_X is absolutely continuous with respect to a Lesbegue measure on $\mathbb{R}$. Assume that $x_1, \ldots, x_n$ are the data points observed from a realization of the random variables $X_1, \ldots, X_n$. Define the bins as $I_j = [x_0 + jh, x_0 + (j+1)h), j = 1, \ldots, k,$ for a starting point x_0. Note that

$$P(X \in I_j) = \int_{I_j} f(x)dx = f(\xi)h, \tag{1}$$

where $\xi \in I_j$ and the last equality follows from the mean value theorem for continuous bounded functions. Intuitively, we can approximate the probability of X falling into the interval I_j by the proportion of observations in I_j, i.e.,

$$P(X \in I_j) \approx \frac{\#\{x_i \in I_j\}}{n}. \tag{2}$$

Using the approximation in (2) and the equation in (1), the density function $f(x)$ can be

estimated by

$$\hat{f}_h(x) = \frac{\#\{x_i \in I_j\}}{nh} = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(x_i \in I_j) \text{ for } x \in I_j, \tag{3}$$

where

$$\mathbb{1}(x_i \in I_j) = \begin{cases} 1 & \text{if } x \in I_j \\ 0 & \text{otherwise.} \end{cases}$$

The smoothness of the histogram estimate is controlled by the smoothing parameter h, a characteristic shared by all nonparametric curve estimators. Choosing a small bandwidth leads to a jagged estimate, while larger bandwidths tend to produce over smoothed histogram estimates (see Hardle (1991)). Figure 1 shows an example of two histograms of the same randomly generated data: the histogram on the left hand side was estimated with a small bandwidth and consequently has many bins, while the histogram on the right hand side was computed with a large bandwidth, producing a smaller number of bins. The choice of the bandwidth is discussed in more detail in Section 4. Note that in practice, the choice of k will determine h or vice versa (a rule of thumb for the choice of k is the Sturges' rule: $k = 1 + \log_2 n$).

When building a histogram, not only the bandwidth needs to be chosen, but also the starting point of each bin edge. These choices can produce different impressions of the shape, and hence different estimates. The bin edge problem is a disadvantage of the histogram not shared by other estimators, such as the kernel density estimator. Another disadvantage is that the histogram estimators are usually not smooth, displaying bumps that may have been observed only due to noise.

3   Kernel Density Estimation

In econometrics, kernel density estimation is also known as the Parzen-Rosenblatt window method. It is an approach that is rooted in the histogram methodology. The basic idea is to
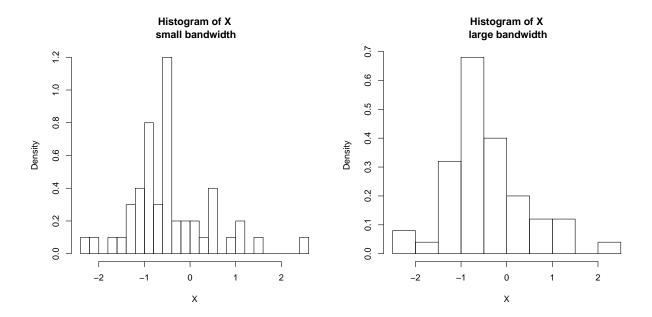
Figura 1: Histogram estimate with small bandwidth (left) and large bandwidth (right)

estimate the density function at a point x using neighboring observations. However, instead of building up the estimate according to bin edges, the naive kernel method (adaptively) uses each point of estimation x as the center of the bin of width $2h$. To express it more transparently, consider the weight function

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

called the kernel weight. Then, the kernel estimate (Rosenblatt (1956)) of $f(x)$ is defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{5}$$

This kernel density estimator is specifically called naive because the kernel weight used is simply a bin of width $2h$ centered at x. See Silverman (1986) for a deeper discussion about this kind of estimator.

Note that the estimator in (5) is an additive function of the kernel weight, inheriting properties such as continuity and differentiability. Hence, it is not continuous and has zero derivatives everywhere except on the jump points $X_i \pm h$. Moreover, even with a good

choice of h, estimators that use weights as in (4) most often do not produce reasonable estimates of smooth densities. This is because the discontinuity of the kernel weight gives the estimate function a ragged form, creating sometimes misleading impressions due to several bumps and constant estimates where few data points are observed. As an illustration, we consider the CEO compensation data in 2012, containing the 200 highest paid chief executives in the U.S. This data set can be obtained from the Forbes website http://www.forbes.com/lists/2012/12/ceo-compensation-12_rank.html. For a better visualization of the plot, we excluded the number 1 in the ranking, with an income of US$131.19 mil, as it was an outlier.
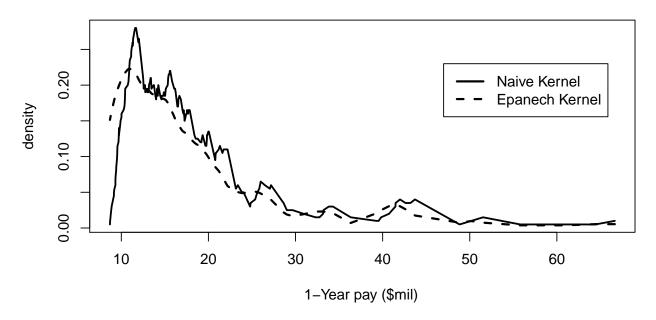

Figura 2: Estimated density of CEO compensation using the naive(solid line) and the Epanechnikov(dashed line) kernels.

Figure 2 shows two density estimators: the solid line represents the naive estimator, while the dashed line represents a more adequate kernel type, called Epanechnikov, which will be described later. The density estimated by the naive kernel appears to have several small

bumps, which are probably due to noise, not a characteristic of the true underlying density. On the other hand, the Epanechnikov kernel is smooth, avoiding this issue.

A usual choice for the kernel weight K is a function that satisfies $\int_{-\infty}^{\infty} K(x)dx = 1$. If moreover, it is assumed that K is a unimodal probability density function that is symmetric about 0, then the estimated density $\hat{f}(x)$ is guaranteed to be a density. Note that the weight in (4) is an example of such choice. Suitable weight functions help overcome problems with bumps and discontinuity of the estimated density. For example, if K is a gaussian distribution, the estimated density function $\hat{f}$ will be smooth and have derivatives of all orders. Table 1 presents some of the most used kernel functions and Figure 3 displays the format of the Epanechnikov, Uniform, Gaussian and Triweight kernels.

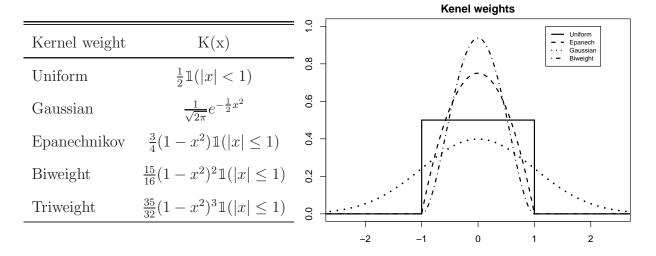| Kernel weight	K(x)
Uniform	$\frac{1}{2}\mathbb{1}(|x| < 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
Epanechnikov	$\frac{3}{4}(1 - x^2)\mathbb{1}(|x| \leq 1)$
Biweight	$\frac{15}{16}(1 - x^2)^2\mathbb{1}(|x| \leq 1)$
Triweight	$\frac{35}{32}(1 - x^2)^3\mathbb{1}(|x| \leq 1)$

Tabela 1: Most common Kernel weights

Figura 3: Kernel weight functions

One of the drawbacks of the kernel density estimation is that it is always biased, particularly near the boundaries (when the data is bounded). However, the main drawback of this approach happens when the underlying density has long tails. In this case, if the bandwidth is small, spurious noise appears in the tail of the estimates, or if the bandwidth is large enough to deal with the tails, important features of the main part in the distribution may be lost due to the over-smoothing. To avoid this problem, adaptive bandwidth methods have been proposed, where the size of the bandwidth depends on the location of the estimation.

See Section 4 for more details on bandwidth selection.

3.1 Properties of Kernel density estimators

In this section, some of the theoretical properties of the kernel density estimator are derived, yielding reliable practical use. Assume we have $X_1, \ldots, X_n$ i.i.d. random variables from a density f and let $K()$ be a Kernel weight function such that the following conditions hold

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \text{ and } \int u^2 K(u)du = \mu_2(K) > 0.$$

Then, for a non-random h, the expected value of $\hat{f}(x)$ is

$$E(\hat{f}(x)) = \frac{1}{nh}\sum_{i=1}^{n} E\left(K\left(\frac{x-X_i}{h}\right)\right) = \frac{1}{h}E\left(K\left(\frac{x-X_i}{h}\right)\right) \tag{6}$$

$$= \frac{1}{h}\int K\left(\frac{x-u}{h}\right)f(u)du = \int K(y)f(x-yh)dy \tag{7}$$

It is easy to see that $\hat{f}$ is an asymptotic unbiased estimator of the density, since $E(\hat{f}(x)) \to f(x)\int K(y)dy = f(x)$ when $h \to 0$. It is important to note that the bandwidth strongly depends on the sample size, so that when the sample size grows, the bandwidth tends to shrink.

Now, assume also that the second derivative f'' of the underlying density f is absolutely continuous and square integrable. Then, expanding $f(x+yh)$ in a Taylor series about x we have

$$f(x-yh) = f(x) - hyf'(x) + \frac{1}{2}h^2 y^2 f''(x) + o(h^2)$$

Then, using the conditions imposed on the Kernel, the bias of the density estimator is

$$Bias(\hat{f}(x)) = \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2) \tag{8}$$

The variance of the estimated function can be calculated using steps similar to those in (6):

$$
\begin{aligned}
Var(\hat{f}(x)) &= \frac{1}{nh} \int K^2(y) f(x - hy) dy - \frac{1}{n} \left(E(\hat{f}(x)) \right)^2 \\
&= \frac{1}{nh} \int K^2(y) \{ f(x) + o(1) \} dy = \frac{1}{n} \{ f(x) + o(1) \}^2 \\
&= \frac{1}{nh} \int K^2(y) dy f(x) + o \left(\frac{1}{nh} \right) \\
&= \frac{1}{nh} R(K) f(x) + o \left(\frac{1}{nh} \right),
\end{aligned}
$$

where $R(g) = \int g^2(y) dy$ for any square integrable function g. From the definition of Mean Square Error (MSE), we have

$$
\begin{aligned}
MSE(\hat{f}(x)) &= \int (\hat{f}(x) - f(x))^2 dx = Var(\hat{f}(x)) + Bias^2(\hat{f}(x)) \\
&= \frac{1}{nh} R(K) f(x) + \frac{h^4}{4} f''^2(x) \mu_2^2(K) + o \left(\frac{1}{nh} \right) + o(h^4)
\end{aligned}
$$

It is straightforward to see that, in order for the kernel density estimation to be consistent for the underlying density, two conditions on the bandwidth are needed as $n \to \infty$: $h \to 0$ and $nh \to \infty$. When these two conditions hold, $MSE(\hat{f}(x)) \to 0$, and we have consistency. Moreover, the trade-off between bias and variance is controlled by the MSE, where decreasing bias leads to a very noise (large variance) estimate and decreasing variance yields over-smoothed estimates (large bias). As has already been pointed out, the smoothness of the estimate depends on the smoothing parameter h, which is chosen as a function of n. For the optimal asymptotic choice of h, a closed form expression can be obtained from minimizing the Mean Integrated Square Error (MISE). Integrating the MSE over the entire line, we find (Parzen (1962))

$$
MISE(\hat{f}) = E \int (\hat{f}(x) - f(x))^2 dx = \frac{R(K)}{nh} + \frac{h^4 \mu_2^2(K) R(f'')}{4}, \tag{9}
$$

and the bandwidth h that minimizes MISE is then

$$
h_{\text{MISE}} = \left(\frac{R(K)}{\mu_2^2(K) R(f'')} \right)^{1/5} n^{-1/5}. \tag{10}
$$

Using this optimal bandwidth, we have

$$\inf_{h>0} MISE(\hat{f}) \approx \frac{5}{4} \left[\mu_2^2(K) R^4(K) R(f'') \right]^{1/5} n^{-4/5}. \tag{11}$$

A natural question is how to choose the kernel function K to minimize (11). Interestingly, if we restrict the choice to a proper density function, the minimizer is the Epanechnikov kernel, where $\mu_2^2(K) R^4(K) = 3^4/5^6$.

The problem with using the optimal bandwidth is that it depends on the unknown quantity f'', which measures the speed of fluctuations in the density f, i.e., the roughness of f. Many methods have been proposed to select a bandwidth that leads to good performance in the estimation, some of these are discussed in Section 4.

The asymptotic convergence of the kernel density estimator has been widely explored. Bickel and Rosenblatt (1973) showed that for sufficiently smooth f and K, $\sup_\ell |\hat{f}(x) - f(x)|/\sqrt{f(x)}$, when normalized properly, has an extreme value limit distribution. The strong uniform convergence of $\hat{f}$

$$\lim_{n\to\infty} \sup_x |\hat{f}(x) - f(x)| = 0 \ a.e. \tag{12}$$

has been studied extensively when the observations are independent or weakly dependent. Nadaraya (1965) showed that if K is of bounded variation and if f is uniformly continuous, then (12) holds as long as $\sum_{m\geq 1} e^{\gamma m h_n^2} < \infty$ for each $\gamma > 0$. Moreover, Stute (1982) derives a law of the logarithm for the maximal deviation between a kernel density estimator and the true underlying density function, Gine and Guillou (2002) find rates for the strong uniform consistency of kernel density estimators and Einmahl and Mason (2005) introduce a general method to prove uniform in bandwidth consistency of kernel-type function estimators. Other results on strong uniform convergence with different conditions can be found in several other papers, such as Parzen (1962), Bhattacharya (1967), Van Ryzin (1969), Moore and Yackel (1977),Silverman (1978) and Devroye and Wagner (1980) .

10

4 The choice of the smoothing parameter h

Selecting an appropriate bandwidth for a kernel density estimator is of crucial importance, and the purpose of the estimation may be an influential factor in the selection method. In many situations, it is sufficient to subjectively choose the smoothing parameter by looking at the density estimates produced by a range of bandwidths. One can start with a large bandwidth, and decrease the amount of smoothing until reaching a "reasonable" density estimate. However, there are situations where several estimations are needed, and such an approach is impractical. An automatic procedure is essential when a large number of estimations are required as part of a more global analysis.

The problem of selecting the smoothing parameter for kernel estimation has been explored by many authors, and no procedure has yet been considered the best in every situation. Automatic bandwidth selection methods can basically be divided in two categories: classical and plug-in. Plug-in methods refer to those that find a pilot estimate of f, sometimes using a pilot estimate of h, and "plug it in" the estimation of MISE, computing the optimal bandwidth as in (10). Classical methods, such as cross-validation, Mallow's C_p, AIC, etc, are basically extensions of methods used in parametric modeling. Loader (1999) discusses the advantages and disadvantages of the plug-in and classical methods in more detail. Besides these two approaches, it is possible to find an estimate of h based on a reference density. Next, we present in more detail the reference method and the most used automatic bandwidth selection procedures.

4.1 Reference to a Distribution

A natural way to overcome the problem of not knowing f'' is to choose a reference density for f, compute f'' and substitute it in (10). For example, assume that the reference density

is Gaussian, and a Gaussian kernel is used, then

$$h_{\text{MISE}} = \left(\frac{R(K)}{\mu_2^2(K)R(f'')} \right)^{1/5} n^{-1/5} = \left[\frac{(2\sqrt{\pi})^{-1}}{\frac{3}{8}\pi^{-1/2}\sigma^{-5}} \right]^{1/5} n^{-1/5} = 1.06\sigma n^{-1/5}.$$

By using an estimate of σ, one has a data-based estimate of the optimal bandwidth. In order to have an estimator that is more robust against outliers, the interquartile range R can be used as a measure of the spread. This modified version can be written as

$$\hat{h}_{\text{robust}} = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34} \right) n^{-1/5}.$$
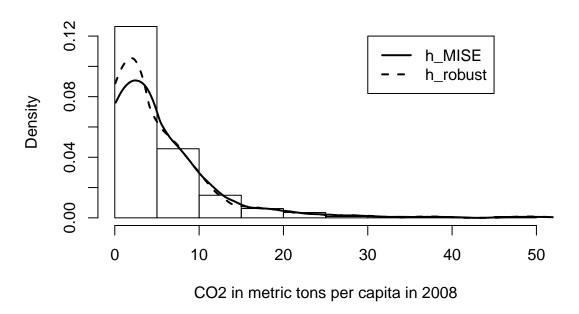
Histogram of CO2

Figura 4: Estimated density of CO2 per capita in 2008 using the bandwidth that minimizes MISE(solid line) and the robust bandwidth(dashed line).

Figure 4 shows the estimated density of CO2 per capita in the year of 2008. The data set can be found at http://data.worldbank.org/indicator/EN.ATM.CO2E.PC/countries. Note

that the estimated density that was computed with the robust bandwidth captures the peak that characterizes the mode, while the estimated density with the bandwidth that minimizes MISE smoothes out this peak. This happens because the outliers at the tail of the distribution contribute to h_{MISE} be larger than the robust bandwidth h_{robust}. For more details on this estimator, see Silverman (1986) or Hardle (1991).

These methods are of limited practical use, since they are restricted to situations where a pre-specified family of densities is correctly selected. Plug-in and classical methods, described below, do not suffer from this limitation.

4.2 Plug-in Methods

There are several papers that address the plug-in approach for bandwidth selection. Some of them study different ways to estimate $R(f'')$, others explore ideas on how to select a pilot bandwidth to better estimate $R(f'')$. The idea is that the only unknown part of (10) needs to be estimated, and hence the bandwidth estimator h_{MISE} can be obtained.

Scott, Tapia and Thompson (1977) proposed a sequential process: calculate $\hat{R}(f'') = R(\hat{f}_{h_2}(x))$, plug $\hat{R}(f'')$ into (10) to obtain h_3, and iterate until convergence of the bandwidth. Hall and Marron (1987) proposed estimating $\hat{R}(f^{(p)})$ by $\hat{R}(f_h^{(p)}) = R(\hat{f}_h^{(p)}) - \frac{R(K^{(p)})}{nh^{2p+1}}$. Parzen and Marron (1990) modified this idea, estimating $\hat{R}(f^{(p)}) = R(\hat{f}_g^{(p)}) - \frac{R(K^{(p)})}{ng^{2p+1}}$ with g having the optimal rate given in Hall and Marron (1987). An improvement of Parzen and Marron (1990) method can be found in Sheather and Jones (1991). Hall, Sheather, Jones and Marron (1991) proposed to use a kernel of order 2 and to take one extra term in the Taylor expansion of the integrated square bias, leading to

$$\text{MISE}_2(h) = \frac{R(K)}{nh} + \frac{h^4}{4}\mu_2^2(K)R(f'') - \frac{h^6}{24}\mu_2(K)\mu_4(K)R(f'''). \tag{13}$$

Since the minimizer of (13) is not analytically feasible, they proposed to estimate the

bandwidth by

$$h_{\text{HSJM}} = (\hat{J}_1/n)^{1/5} + \hat{J}_2(\hat{J}_1/n)^{3/5}$$

where $\hat{J}_1 = \frac{R(K)}{\mu_2^2(K)\hat{R}(f'')}$ and $\hat{J}_2 = \frac{\mu_4(K)\hat{R}(f''')}{20\mu_2(K)\hat{R}(f'')}$.

Several other plug-in methods have been proposed, and a review of the first procedures that address this type of methodology can be found in Turlach (1993). Modern research on plug-in methods have actually become somewhat hybrid, combining ideas of plug-in and classical approaches such as cross validation, see Biased Cross-Validation described below for example. More recently, inspired by developments in threshold selection, Chan, Lee and Peng (2010) propose to choose $h = O(n^{-1/5})$ as large as possible, so that the density estimator has a larger bias, but smaller variance than $\hat{f}_{h_{AMSE}(x)}$. The idea is to consider an alternative kernel density estimator $\bar{f} = \frac{1}{nh}\sum_{i=1}^{n}\bar{K}\left(\frac{x-X_i}{h}\right)$ and define $\Delta_n(x;h) = \frac{\sqrt{nh}\{\hat{f}(x;h)-\bar{f}(x;h)\}}{\hat{f}^{1/2}(x;h)\{\int(K(s)-\bar{K}(s))^2 ds\}^{1/2}}$. Then, the choice for the smoothing parameter is

$$\hat{h}_l = arg\min\{h : |\Delta_n(x;r)| > z_\alpha \text{ for all } r > h, r \in [cn^{-1/5}, n^{-\epsilon}]\},$$

where z_α denotes a critical point in $N(0,1)$, $c > 0$ and $0 < \epsilon < 1/5$. The intuition is that, when h is large $\Delta_n(x;h) > z_\alpha$, since $\Delta_n(x;r) \overset{d}{\to} N(0,1)$.

4.3   Classical Methods

4.3.1   Least Squares Cross-Validation

Cross-validation is a popular and readily implemented heuristic for selecting the smoothing parameter in kernel estimation. Introduced by Rudemo (1982) and Bowman (1984), least squares cross-validation is very intuitive and has been a fundamental device in recent research. The idea is to consider the expansion of the Integrated Square Error (ISE) in the following way

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

Note that the last term does not depend on $\hat{f}_h$, hence on h, so that we only need to consider the first two terms. The ideal choice of bandwidth is the one which minimizes

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

The principle of the least squares cross-validation method is to find an estimate of $L(h)$ from the data and minimize it over h. Consider the estimator

$$CV_{LS}(h) = \int \hat{f}_h^2(x)dx - 2\frac{1}{n}\sum_i \hat{f}_{h,-i}(X_i), \qquad (14)$$

where

$$\hat{f}_{h,-i}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x - X_j}{h}\right). \qquad (15)$$

The summation in (14) has expectation

$$E\frac{1}{n}\sum_i \hat{f}_{h,-i}(X_i) = E\hat{f}_{h,-n}(X_n) = E\int \hat{f}_{h,-n}(x)f(x)dx = E\int \hat{f}_h(x)f(x)dx,$$

because $E(\hat{f}_h)$ depends only on the kernel and bandwidth, not on the sample size. It follows that $E(CV_{LS}(h)) = E(L(h))$, and hence $CV_{LS}(h) + \int f^2(x)dx$ is an unbiased estimator of MISE (reason why this method is also called *unbiased cross-validation*). Assuming that the minimizer of $CV_{LS}(h)$ is close to the minimizer of $E(CV_{LS}(h))$, the bandwidth

$$h_{LSCV} = arg\min_h CV_{LS}(h)$$

is the natural choice. This method suffers from sample variation, that is, using different samples from the same distribution, the estimated bandwidths may have large variance. Further discussion on this method can be found in Bowman, Hall and Titterington (1984), Hall (1983) and Stone (1984).

4.3.2 Biased Cross-Validation

Biased cross-validation considers the asymptotic MISE

$$AMISE\{\hat{f}_h\} = \frac{R(K)}{nh} + \frac{h^4}{4}\mu_2^2(K)R(f'').$$

This method was suggested by Scott and Terrell (1987), and its main idea is to replace the unknown quantity $R(f'')$ by the estimator

$$
\begin{aligned}
\widetilde{R(f'')} &= R(\hat{f}_h'') - (nh^5)^{-1}R(K'') \\
&= n^{-2}\sum\sum_{i \neq j}(K_h'' * K_h'')(X_i - X_j),
\end{aligned}
$$

to give

$$
BCV(h) = (nh)^{-1}R(K) + \frac{h^4}{4}\mu_2^2(K)\widetilde{R(f'')}.
$$

Then, the bandwidth selected is $h_{BCV} = arg\min_h BCV(h)$. This selector is considered a hybrid of cross-validation and plug-in, since it replaces an unknown value in AMISE by a cross-validatoin kernel estimate $\widetilde{R(f'')}$.

4.3.3 Likelihood Cross-Validation

Suppose that in addition to the original data set $X_1, \ldots, X_n$, we have another independent observation X^* from f. Thinking of $\hat{f}_h$ as a parametric family depending on h, but with fixed data $X_1, \ldots, X_n$, we can view $\log \hat{f}(X^*)$ as the likelihood of the bandwidth h. Because in reality no additional observation is available, we can omit a randomly selected observation from the original data, say X_i, and compute $\hat{f}_{h,-i}(X_i)$, as in (15). Note that there is no pattern when choosing the observation to be omitted, so that the score function can be taken as the log likelihood average

$$
CV(h) = n^{-1}\sum_{i=1}^{n}\log \hat{f}_{h,-i}(X_i).
$$

Naturally, we choose the bandwidth the minimizes $CV(h)$, which is known to minimize the Kullback-Leibler distance between $\hat{f}_h(x)$ and $f(x)$. This method was proposed by Habbema, Hermans and van den Broek (1974) and Duin (1976), but other results can be found in Marron (1987), Marron (1989) and Cao, Cuevas and Gonzalez-Manteiga (1994).

In general, bandwidths chosen via cross validation methods in kernel density estimation are highly variable, and usually give undersmooth density estimates, causing undesired spurious bumpiness.

4.3.4 Indirect Cross-validation

The Indirect Cross-validation (ICV) method, proposed by Savchuk, Hart and Sheather (2010), slightly outperforms least squares cross-validation in terms of mean integrated squared error. The method can be described as follows. First define the family of kernels $\mathbb{L} = \{L(.; \alpha, \sigma) : \alpha \geq 0, \sigma > 0\}$ where, for all u, $L(u; \alpha, \sigma) = (1 + \alpha)\phi(u) - \frac{\alpha}{\sigma}\phi\left(\frac{u}{\sigma}\right)$. Note that this is a linear combination of two gaussian kernels. Then, select the bandwidth of an L-kernel estimator using least squares cross-validation, and call it $\hat{b}_{UCV}$. Under some regularity conditions on the underlying density f, h_n and b_n that asymptotically minimize the MISE of ϕ and L-kernel estimators, have the following relation

$$h_n = \left(\frac{R(\phi)\mu_2^2(L)}{R(L)\mu_2^2(\phi)}\right)^{1/5} b_n = Cb_n.$$

The indirect cross-validation bandwidth is chosen to be $\hat{h}_{ICV} = C\hat{b}_{UCV}$. Savchuk et al. (2010) show that the relative error of ICV bandwidths can converge to 0 at a rate of $n^{1/4}$, much better than the $n^{1/10}$ rate of LSCV.

4.4 Other Methods

4.4.1 Variable Bandwidth

Rather than using a single smoothing parameter h, some authors have considered the possibility of using a bandwidth $h(x)$ that varies according to the point x at which f is estimated. This is often referred as the *balloon* estimator and has the form

$$\hat{f}(x) = \frac{1}{nh(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(x)}\right). \tag{16}$$

The *balloon* estimator was introduced by Loftsgaarden and Quesenberry (1965) in the form of the k*th* nearest neighbor estimator. In Loftsgaarden and Quesenberry (1965), $h(x)$ was based on a suitable number k, so that it was a measure of the distance between x and the k*th* data point nearest to x. The optimal bandwidth for this case can be shown to be (analogue of (10) for asymptotic MSE)

$$h_{\text{AMSE}}(x) = \left(\frac{R(K)f(x)}{\mu_2^2(K)f''^2(x)} \right)^{1/5} n^{-1/5}. \tag{17}$$

Another variable bandwidth method is to have the bandwidth vary not with the point of estimation, but with each observed data point. This type of estimator, known as sample-point or variable kernel density estimator, was introduced by Breiman, Meisel and Purcell (1977) and has the form

$$\hat{f}(x) = \frac{1}{nh(X_i)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(X_i)} \right). \tag{18}$$

This type of estimator has one advantage over the *balloon* estimator: it will always integrate to 1, assuring that it is a density. Note that $h(X_i)$ is a function of random variables, and thus it is also random.

More results on the variable bandwidth approach can be found in Hall (1992), Taron, Paragios and Jolly (2005), Wu, Chen and Chen (2007) and Gine and Sang (2010).

4.4.2 Binning

An adaptive type of procedure is the binned kernel density estimation, studied by a few authors such as Scott (1981), Silverman (1982) and Jones (1989). The idea is to consider equally spaced bins B_i with centers at t_i and bin counts n_i, and define the estimator as

$$\hat{f}_{bin}(x) = \frac{1}{n} \sum_{i=-\infty}^{\infty} n_i K\left(\frac{x - t_i}{h} \right) = \frac{1}{n} \sum_{i=1}^{m} K\left(\frac{x - t_i}{h} \right), \tag{19}$$

where the sum over m means summing over the finite non-empty bins that exist in practice. It is also possible to use a variable bandwidth in (19), yielding the estimator

$$\tilde{f}_{bin}(x) = \frac{1}{n} \sum_{i=1}^{m} K\left(\frac{x - t_i}{h(X_i)}\right),\tag{20}$$

Examples of other approaches and discussion on this type of estimation can be found in Hall and Wand (1996), Cheng (1997), Minnotte (1999), Pawlak and Stadtmuller (1999), Holmstrom (2000).

4.4.3 Bootstrap

A methodology that has been recently explored is that of selecting the bandwidth using bootstrap. It focuses on replacing the MSE by MSE^*, a bootstrapped version of MSE, which can be minimized directly. Some authors resample from a subsample of the data $X_1, \ldots, X_n$ (see Hall (1990)), others replace from a pilot density based on the data (see Faraway and Jhun (1990), Hazelton (1996), Hazelton (1999)), more precisely, from $\tilde{f}_h^b(x) = \frac{1}{nb_n} \sum_{i=1}^{n} L\left(\frac{x-X_i}{b_n}\right)$, where L is another kernel and b_n is a pilot bandwidth. Since the bandwidth choice reduces to estimating s in $h = n^{-1/5}s$, Ziegler (2006) introduces $f_{n,s}^*(x) = \frac{1}{n^{4/5}s} \sum_{i=1}^{n} K\left(\frac{x-X_i^*}{n^{-1/5}s}\right)$, and obtain $MSE_{n,s}^*(x) = E^*((f_{n,s}^*(x) - \tilde{f}_h^b(x))^2)$. The proposed bandwidth is

$$h_n = n^{-1/5} arg \min_{s} MSE_{n,s}^*.$$

Applications of the bootstrap idea can be found in many different areas of estimation, see Delaigle and Gijbels (2004), Loh and Jang (2010) for example.

4.4.4 Estimating Densities on $\mathbb{R}_+$

It is known that kernel density estimators have larger bias on the boundaries. Many methods have been proposed to alleviate such problem, such as the use of gamma kernels or inverse and reciprocal inverse gaussian kernels, also known as varying kernel approach.

Chen (2000) proposes to replace the symmetric kernel by a gamma kernel, which has flexible shapes and locations on $\mathbb{R}_+$. Their estimator can be described in the following way. Suppose the underlying density f has support $[0, \infty)$ and consider the gamma kernel

$$K_{x/b+1,b}(t) = \frac{t^{x/b}e^{-t/b}}{b^{x/b+1}\Gamma(x/b+1)},$$

where b is a smoothing parameter such that $b \to 0$ and $nb \to \infty$. Then, the gamma kernel estimator is defined as

$$\hat{f}^G(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x/b+1,b}(X_i).$$

The expected value of this estimator is

$$E\hat{f}^G(x) = \int_0^{\infty} K_{x/b+1,b}(y)f(y)dy = Ef(\xi_x),$$

where ξ_x is a Gamma(x/b+1,b) random variable. Using Taylor Expansion and the fact that $E(\xi_x) = x + b$ and $Var(\xi_x) = xb + b^2$ we have that

$$
\begin{aligned}
Ef(\xi_x) &= f(x+b) + \frac{1}{2}f''(x)Var(\xi_x) + o(b) \\
&= f(x) + b\left[f'(x) + \frac{1}{2}xf''(x)\right] + o(b).
\end{aligned}
$$

It is clear then, that this estimator does not have bias problems on the boundaries, since the bias is $O(b)$ near the origin and in the interior. See Chen (2000) for further details. Other approaches on estimating the density on $\mathbb{R}_+$ can be found in Scaillet (2004), Mnatsakanov and Ruymgaart (2012), Mnatsakanov and Sarkisian (2012), Comte and V.Genon-Catalot (2012) and references therein.

Some interest on density estimation research is on bias reduction techniques, which can be found in Jones, Linton and Nielsen (1995), Choi and Hall (1999), Cheng, Choi, Fan and Hall (2000), Choi, Hall and Roussan (2000) and Hall and Minnotte (2002). Other recent improvements and interesting applications of the kernel estimate can be found in Hirukawa (2010),Liao, Wu and Lin (2010), Matuszyk, Cardew-Hall and Rolfe (2010), Miao, Rahimi

and Rao (2012), Chu, Liau, Lin and Su (2012), Golyandina, Pepelyshev and Steland (2012) and Cai, Rushton and Bhaduri (2012) among many others.

4.4.5 Estimating the distribution function $F(x)$

It is not uncommon to find situations where it is desirable to estimate the distribution function $F(x)$ instead of the density function $f(x)$. A whole methodology known as kernel distribution function estimation (KDFE) has been explored since Nadaraya (1964) introduced the estimator

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

where K is the distribution function of a positive kernel k, i.e $K(x) = \int_{-\infty}^{x} k(t)dt$. Authors have considered many alternatives for this estimation, but the basic measures of quality or this type of estimator are

$$
\begin{aligned}
ISE(h) &= \int [\hat{F}_h(x) - F(x)]^2 W(x) dF(x) \text{ and} \\
MISE(h) &= E \int [\hat{F}_h(x) - F(x)]^2 W(x) dF(x),
\end{aligned}
$$

where W is a non-negative weight function.

Sarda (1993) considered a discrete approximation to MISE, the average squared error

$$ASE(h) = \frac{1}{n} \sum_{i=1}^{n} [\hat{F}_h(X_i) - F(X_i)]^2 W(X_i).$$

He suggests replacing the unknown $F(X_i)$ by the empirical $F_n(X_i)$ and then selecting the bandwidth that minimizes the leave-one-out criterion

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

As an alternative to this cross-validation criterion, Altman and Leger (1995) introduce a plug-in estimator of the asymptotically optimal bandwidth. There is a vast literature

on estimating kernel distribution functions, for example Bowman, Hall and Prvan (1998), Tenreiro (2006), Ahmad and Amezziane (2007), Janssen, Swanepoel and Veraberbeke (2007), Berg and Politis (2009), just to cite a few.

4.5 Example of Bandwidth Selection Methods

It is well known that plug-in bandwidth estimators tend to select larger bandwidths when compared to the classical estimators. They are usually tuned by arbitrary specification of pilot estimates and most often produce over smoothed results when the smoothing problem is difficult. On the other hand, smaller bandwidths tend to be selected by classical methods, producing under smoothed results. The goal of a selector of the smoothing parameter is to make that decision purely from the data, finding automatically which features are important and which should be smoothed away.

Figure 5 shows an example of classical and plug-in bandwidth selectors for a real data set. The data corresponds to the exports of goods and services of countries in 2011, representing the value of all goods and other market services provided to the rest of the world. The data set can be downloaded from the world bank website (http://data.worldbank.org).

The plug-in estimators a) rule of thumb for Gaussian and b) Seather and Jones selector produced a very smooth fit, while unbiased cross-validation selects a small bandwidth, yielding a highly variable density estimate. The hybrid method biased cross-validation, is the one that selects the largest bandwidth, hence its corresponding density estimate is very smooth, smoothing away information of the peak (mode).

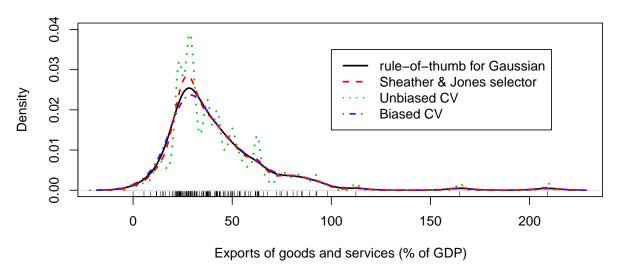**Density estimates of exports of goods and services**

Figura 5: Estimated densities for bandwidths chosen using different methods.

5   SiZer

In nonparametric estimation, the challenge of selecting the smoothing parameter that yields the best possible fit has been addressed through several methods, as described in previous sections. The challenge is to identify the features that are really there, but at the same time to avoid spurious noise. Marron and Chung (1997) and other authors noted that it may be worth to consider a family of smooths with a broad range of bandwidths, instead of a single estimated function. Figure 6 shows an example of a density generated from a mixture of a Gaussian variable with mean 0 and variance 1 and another Gaussian variable, with mean 8 and variance 2. The density was estimated with a Epanechnikov kernel using bandwidths that vary from 0.4 to 10. The wide range of smoothing considered, from a small bandwidth producing a wiggly estimate to a very large bandwidth yielding nearly the simple least squares fit, allows a contrast of estimated features at each level of smoothing. The two highlighted bandwidths are equal to 0.6209704 and 1.493644, corresponding to the choice of biased cross-validation (blue) and to Silverman's rule of thumb (red) (see Silverman, 1986)

respectively .

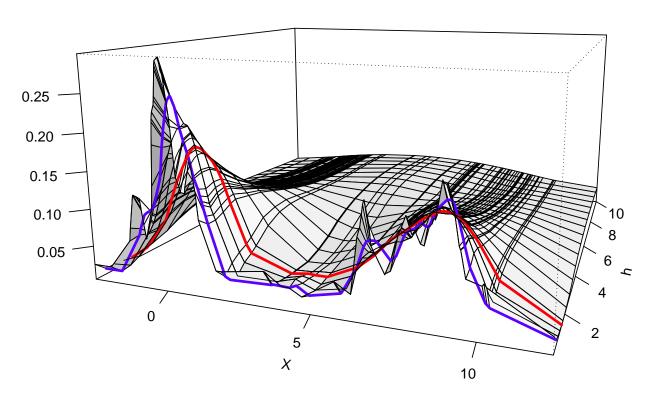**Mixture of Normals Density**

Figura 6: Estimated density with several bandwidths.

The idea of considering a family of smooths has its origins in scale space theory in computer science. A fundamental concept in such analysis is that it does not aim at estimating one true curve, but at recovering the significant aspects of the underlying function, since different levels of smoothing may reveal different intrinsic features. Exploring this concept in a statistical point of view, Chaudhuri and Marron (2000) introduced a procedure called **SI**gnificance **ZER**o crossings of smoothed estimates (SiZer), whose objective is to analyze the visible features representing important underlying structures for different bandwidths. Next, we briefly describe such method.

Suppose that $h \in H$, where H is a subinterval of $(0, \infty)$, and $x \in I$, where I is a subinterval of $(-\infty, \infty)$. Then the family of smooth curves $\{\hat{f}_h(x) | h \in H, x \in I\}$ can be represented by a surface called *scale space surface*, which captures different structures of the curve under different levels of smoothing. Hence, the focus is really on $E(\hat{f}_h(x))$ as h varies in H and x in I, which is called in Chaudhuri and Marron (2000) as "true curves viewed at different scales of resolution".

A smooth curve $\hat{f}_h(x)$ has derivatives equal to 0 at points of minimum (valleys), maximum (peaks) and points of inflection. Note that, before a peak (or valley), the sign of the derivative $\partial \hat{f}_h(x)/\partial x$ is positive (or negative), and after it the derivative is negative (or positive). In other words, peaks and valleys are determined by zero crossings of the derivative. Actually, we can identify structures in a smooth curve by zero crossings of the m*th* order of the derivative. Using a Gaussian kernel $K(x) = (1/\sqrt{2\pi})exp(-x^2/2)$, Silverman (1981) showed that the number of peaks in a kernel density estimate decreases monotonically with the increase of the bandwidth, and Chaudhuri and Marron (2000) extended this idea for the number of zero crossings of the mth order derivative $\partial^m \hat{f}_h(x)/\partial x^m$ in kernel regression.

The asymptotic theory of the scale space surfaces and their derivatives studied by Chaudhuri and Marron (2000), which hold even under bootstrapped or resampled distributions, provides tools for building bootstrap confidence intervals and tests of significance for their features (see Chaudhuri and Marron (1999)). SiZer basically considers the null hypothesis

$$H_0^{h,x} : \partial^m E(\hat{f}_h(x))/\partial x^m = 0,$$

for a fixed $x \in I$ and $h \in H$. If $H_0^{h,x}$ is rejected, there is evidence that $\partial^m E(\hat{f}_h(x))/\partial x^m$ is positive or negative, according to the sign of $\partial^m \hat{f}_h(x)/\partial x^m$.

The information is displayed in a color map of scale space, where the pixels represent the location of x (horizontally) and h (vertically). The regions are shaded blue for significant increasing curve, red for significantly decreasing, purple for unable to distinguish

and gray for insufficient data. Note that purple is displayed when the confidence interval for the derivative contains 0. There are a few options of software available, including java (http://www.wagner.com/SiZer/SiZerDownload.html), matlab (http://vm1.cas.unc.edu/stat-or/webspace/miscellaneous/marron/Matlab7Software/Smoothing/) and R (SiZer package).

Figure 7 shows an example of a color map obtained with SiZer. The data is the GDP per person employed in 2010 (downloadable at http://data.worldbank.org). It is easy to see that for large bandwidths, the density function significantly increases until about 16000, then after a small area that SiZer is unable to distinguish, it has a significant decrease, hence estimating a density with one mode at around 16000. Small bandwidths produce a map that is mostly gray, meaning that the wiggles in the estimate at that level of resolution can not be separated from spurious sampling noise. An interesting blue area appears, with a mid-level resolution, near 43000, indicating a slightly significant increase. This comes after and before a purple area, which SiZer is unable to distinguish if it is increasing or decreasing. Thus, with a mid-level bandwidth, the estimated density would suggest 2 modes, one somewhere near 10000 and another near 43000.
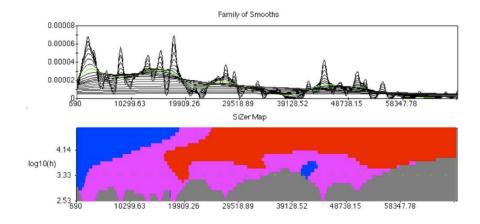

Figura 7: SiZer example

Referências

Ahmad, I. A. and Amezziane, M. (2007). A general and fast convergent bandwidth selection method of kernel estimator, *Journal of Nonparametric Statistics* **19**: 165–187.

Altman, N. and Leger, C. (1995). Bandwidth selection for kernel distribution function estimation, *Journal of Statistical Planning and Inference* **46**: 195–214.

Berg, A. and Politis, D. (2009). Cdf and survival function estimation with infinite order kernels, *Electronic Journal of Statistics* **3**: 1436–1454.

Bhattacharya, P. (1967). Estimation of a probability density function and its derivatives, *Sankhyii Ser. A* **29**: 373–382.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *The Annals of Statistics* pp. 1071–1095.

Bierens, H. J. (1987). Kernel estimators of regression functions, *Advances in Econometrics: Fifth World Congress, Vol.I, Cambridge University Press* pp. 99–144.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* **71**: 353–360.

Bowman, A. W., Hall, P. and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functio, *Biometrika* **85**: 799–808.

Bowman, A. W., Hall, P. and Titterington, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities, *Biometrika* **71**: 341–351.

Breiman, L., Meisel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities, *Technometrics* **19**: 135–144.

Cai, Q., Rushton, G. and Bhaduri, B. (2012). Validation tests of an improved kernel density estimation method for identifying disease clusters, *Journal of Geographical Systems* **14**(3): 243–264.

Cao, R., Cuevas, A. and Gonzalez-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Computational Statistics & Data Analysis* **17**(2): 153–176.

Chan, N.-H., Lee, T. C. and Peng, L. (2010). On nonparametric local inference for density estimation, *Computational Statistics & Data Analysis* **54**: 509–515.

Chaudhuri, P. and Marron, J. S. (1999). Sizer for exploration of structures in curves, *Journal of the American Statistical Association* **94**: 807–823.

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation, *The Annals of Statistics* **28**: 402–428.

Chen, S. (2000). Probability density function estimation using gamma kernels, *Annals of the Institute of Statistical Mathematics* **52**: 471–480.

Cheng, M.-Y. (1997). A bandwidth selector for local linear density estimators, *The Annals of Statistics* **25**: 1001–1013.

Cheng, M.-Y., Choi, E., Fan, J. and Hall, P. (2000). Skewing-methods for two parameter locally-parametric density estimation, *Bernoulli* **6**: 169–182.

Choi, E. and Hall, P. (1999). Data sharpening as prelude to density estimation, *Biometrika* **86**: 941–947.

Choi, E., Hall, P. and Roussan, V. (2000). Data sharpening methods for bias reduction in nonparametric regression, *Annals of Statistics* **28**: 1339–1355.

Chu, H.-J., Liau, C.-J., Lin, C.-H. and Su, B.-S. (2012). Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the taiwan region, *Expert Systems with Applications* **39**: 9451–9457.

Comte, F. and V.Genon-Catalot (2012). Convolution power kernels for density estimation, *Journal of Statistical Planning and Inference* **142**: 1698–1715.

Delaigle, A. and Gijbels, I. (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample, *Annals of the Institute of Statistical Mathematics* **56**(1): 19 – 47.

Devroye, L. and Wagner, T. (1980). The strong uniform consistency of kernel density estimates, *in: P.R. Krishnaiah, ed., Multivariate Analysis, Vol. V (North-Holland, Amsterdam)* pp. 59–77.

Duin, R. P. W. (1976). On the choice of smoothing parameters of parzen estimators of probability density functions, *IEEE Transactions on Computers* **C-25**: 1175 – 1179.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators, *The Annals of Statistics* **33**: 1380–1403.

Faraway, J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation, *Journal of the American Statistical Association* **85**: 1119 –1122.

Gine, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators, *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **38**: 907 –921.

Gine, E. and Sang, H. (2010). Uniform asymptotics for kernel density estimators with variable bandwidths, *Journal of Nonparametric Statistics* **22**: 773 –795.

Golyandina, N., Pepelyshev, A. and Steland, A. (2012). New approaches to nonparametric density estimation and selection of smoothing parameters, *Computational Statistics and Data Analysis* **56**: 2206 –2218.

Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation, *Proceedings in Computational Statistics, Physica Verlag, Vienna.*

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation, *Annals of Statistics* **11**: 1156 –1174.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smooth- ing parameter in nonparametric problems, *Journal of Multivariate Analysis* **32**: 177 –203.

Hall, P. (1992). On global properties of variable bandwidth density estimators, *The Annals of Statistics* **20**: 762 –778.

Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives, *Statistics & Probability Letters* **6**: 109 –115.

Hall, P. and Minnotte, M. (2002). High order data sharpening for density estimation, *Journal of the Royal Statistical Society Series B* **64**: 141 –157.

Hall, P. and Wand, M. (1996). On the accuracy of binned kernel density estimators, *Journal of Multivariate Analysis* **56**: 165 –184.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation, *Biometrika* **78**: 263 –269.

Hardle, W. (1991). *Smoothing Techniques, With Implementations in S*, Springer, New York.

Hazelton, M. (1996). Bandwidth selection for local density estimators, *Scandinavian Journal of Statistics* **23**: 221 –232.

Hazelton, M. (1999). An optimal local bandwidth selector for kernel density estimation, *Journal of Statistical Planning and Inference* **77**: 37 –50.

Hirukawa, M. (2010). Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval, *Computational Statistics and Data Analysis* **54**: 473–495.

Holmstrom, L. (2000). The accuracy and the computational complexity of a multivariate binned kernel density estimator, *Journal of Multivariate Analysis* **72**: 264–309.

Janssen, P., Swanepoel, J. and Veraberbeke, N. (2007). Modifying the kernel distribution function estimator towards reduced bias, *Statistics* **41**: 93 –103.

Jones, M. C. (1989). Discretized and interpolated kernel density estimates, *Journal of the American Statistical Association* **84**: 733 –741.

Jones, M., Linton, O. and Nielsen, J. (1995). A simple bias reduction method for density estimation, *Biometrika* **82**: 327 –328.

Liao, J., Wu, Y. and Lin, Y. (2010). Improving sheather and jones bandwidth selector for difficult densities in kernel density estimation, *Journal of Nonparametric Statistics* **22**: 105–114.

Loader, C. R. (1999). Bandwidth selection: Classical or plug-in?, *The Annals of Statistics* **27**(2): 415–438.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function, *The Annals of Mathematical Statistics* **36**: 1049–1051.

Loh, J. M. and Jang, W. (2010). Estimating a cosmological mass bias parameter with bootstrap bandwidth selection, *Journal of the Royal Statistical Society Series C* **59**: 761–779.

Marron, J. S. (1987). An asymptotically efficient solution to the bandwidth problem of kernel density estimation, *The Annals of Statistics* **13**: 1011–1023.

Marron, J. S. (1989). Comments on a data based bandwidth selector, *Computational Statistics & Data Analysis* **8**: 155–170.

Marron, J. S. and Chung, S. S. (1997). Presentation of smoothers: the family approach, *unpublished manuscript*.

Matuszyk, T. I., Cardew-Hall, M. J. and Rolfe, B. F. (2010). The kernel density estimate/point distribution model(kde-pdm)for statistical shape modeling of automotive stampings and assemblies, *Robotics and Computer-Integrated Manufacturing* **26**: 370–380.

Miao, X., Rahimi, A. and Rao, R. P. (2012). Complementary kernel density estimation, *Pattern Recognition Letters* **33**: 1381–1387.

Minnotte, M. C. (1999). Achieving higher-order convergence rates for density estimation with binned data, *Journal of the American Statistical Association* **93**: 663–672.

Mnatsakanov, R. and Ruymgaart, F. (2012). Moment-density estimation for positive random variables, *Statistics* **46**: 215–230.

Mnatsakanov, R. and Sarkisian, K. (2012). Varying kernel density estimation on r+, *Statistics and Probability Letters* **82**: 1337–1345.

Moore, D. and Yackel, J. (1977). Consistency properties of nearest neighbour density function estimators, *The Annals of Statistics* **5**: 143–154.

Nadaraya, E. A. (1964). On estimating regression, *Theory Probab. Appl.* **9**: 141–142.

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves, *Theory Probab. Appl.* **10**: 186–190.

Parzen, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors, *Journal of the American Statistical Association* **85**: 66–72.

Parzen, E. (1962). On estimation of a probability density function and mode, *Annals of Mathematical Statistics* **33**: 1065–1076.

Pawlak, M. and Stadtmuller, U. (1999). Kernel density estimation with generalized binning, *Scandinavian Journal of Statictics* **26**: 539–561.

Rosenblatt, M. (1956). Remarks on some nonparametrice stimates of a density function, *Annals of Mathematical Statistics* **27**: 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics* **9**: 65–78.

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

Savchuk, O., Hart, J. and Sheather, S. (2010). Indirect cross-validation for density estimation, *Journal of the American Statistical Association* **105**: 415–423.

Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse gaussian kernels, *Nonparametric Stat.* **16**: 217–226.

Scott, D. W. (1981). Using computer-binned data for density estimation, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. W. F. Eddy, Ed., Springer-Velag, New York* pp. 292–294.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*, John Wiley & Sons.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *Journal of American Statistical Association* **82**: 1131–1146.

Scott, D. W., Tapia, R. A. and Thompson, J. R. (1977). Kernel density estimation revisited, *Nonlinear Analysis, Theory, Methods and Applications* **1**: 339–372.

Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society - B* **53**: 683–690.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *The Annals of Statistics* **6**: 177–184.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society - B* **43**: 97–99.

Silverman, B. W. (1982). Kernel density estimation using the fast fourier transform, *Applied Statistics* **31**: 93–97.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics* **12**: 1285–1297.

Stute, W. (1982). A law of the logarithm for kernel density estimators, *The Annals of Probability* **10**(2): 414–422.

Taron, M., Paragios, N. and Jolly, M. P. (2005). Modelling shapes with uncertainties: Higher order polynomials, variable bandwidth kernels and non parametric density estimation, *10th IEEE International Conference on Computer Vision* pp. 1659–1666.

Tenreiro, C. (2006). Asymptotic behaviour of multistage plug-in bandwidth selections for kernel distribution function estimators, *Journal of Nonparametric Statistics* **18**: 101–116.

Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: A review, *CORE and Institut de Statistique.*

Van Ryzin, J. (1969). On strong consistency of density estimates, *The Annals of Mathematical Statistics* **40**(486): 1765–1772.

Wu, T.-J., Chen, C.-F. and Chen, H.-Y. (2007). A variable bandwidth selector in multivariate kernel density estimation, *Statistics & Probability Letters* **77**(4): 462–467.

Ziegler, K. (2006). On local bootstrap bandwidth choice in kernel density estimation, *Statistics & Decisions* **24**: 291–301.