Kernel estimation when density may not exist

Victoria Zinde-Walsh*

McGill University and CIREQ
victoria.zinde-walsh@mcgill.ca
(514) 398 4834

May 11, 2006

*The support of the Social Sciences and Humanities Research Council of Canada (SSHRC), the *Fonds québecois de la recherche sr la société et la culture* (FRQSC) is gratefully acknowledged. I thank participants of the COMPSTAT 2004, RSS 2004, CEA 2005 and UK Econometric Study Group meetings, participants of the seminars at Vanderbilt and Dalhousie Universities and Yanqin Fan, Jeffrey Racine, Oliver Linton, John W.Galbraith and anonymous referees for very useful comments and suggestions.

Running head: Kernel estimation: density may not exist

Victoria Zinde-Walsh
Department of Economics, McGill University
855 Sherbrooke Street West,
Montreal, Quebec, Canada
H3A 2T7

Abstract

Nonparametric kernel estimation of density and conditional mean is widely used, but many of the pointwise and global asymptotic results for the estimators are not available unless the density is continuous and appropriately smooth; in kernel estimation for discrete-continuous cases smoothness is required for the continuous variables. Non-smooth density and mass points in distributions arise in various situations that are examined in empirical studies; some examples and explanations are discussed in the paper. Generally, any distribution function consists of absolutely continuous, discrete and singular components but only a few special cases of nonparametric estimation involving singularity have been examined in the literature and asymptotic theory under the general set-up has not been developed. In this paper the asymptotic process for the kernel estimator is examined by means of the generalized functions and generalized random processes approach; it provides a unified theory since density and its derivatives can be defined as generalized functions for any distribution, including cases with singular components. The limit process for the kernel estimator of density is fully characterized in terms of a generalized Gaussian process. Asymptotic results for the Nadaraya-Watson conditional mean estimator are also provided.

1 Introduction

Many non- or semi-parametric estimators utilize nonparametric kernel estimators of density. For asymptotic results for such estimators existence of density and some smoothness properties of the density are routinely assumed. In much of the literature that develops asymptotic results for estimators of conditional mean, its derivative and average derivatives, assumptions on smoothness of the density function as well as about smoothness of conditional mean are made (see, e.g. Pagan and Ullah, 1999 for a review). However, while conditional mean could often be smooth (even linear) and satisfy various types of conditions on derivatives that follow from some theoretical model, there may be no theoretical basis for assuming density smoothness for many variables. Lumpiness and lack of smoothness in various processes and institutional set-ups can lead to singularities in distributions of observables.

Singularities may manifest themselves as mass points such as those occurring in various hazard functions in biomedical studies; in economics an example is the empirical study by Green and Riddell, 1997 where mass points in the hazard function for weeks worked occur due to unemployment insurance qualification rules. The following example illustrates why singularities may occur in distributions of labor supply as well as earned income, disposable income and the joint distribution of male, female labor supply as a result of a lump-sum transfer; they may be represented by mass points but may also be of a more complex nature. The example is highly simplified; more general varying tax schedules, benefits and transfers would lead to more complicated distributions with singularities.

Example 1. Suppose that each household chooses to supply labor amount y in the range $0 \leq y \leq \overline{y}$ where $\overline{y}$ is random in the population with some distribution function $F_{\overline{y}}(\cdot)$ (possibly absolutely continuous with density $f_{\overline{y}}(\cdot)$). Denote by y^* the supply of labor (= earned income) chosen and by c disposable income (=consumption). The household maximizes c. If $c(y) = y$ then *$y^* = \overline{y}$. Suppose that a lump-sum transfer occurs and as a result*

$$c(y) = \begin{cases} y + t & \text{if } y \leq y_t; \\ y & \text{if } y > y_t. \end{cases} \tag{1}$$

Next, consider two cases.

Case a: Unanticipated transfer. Labor supply is still $y^* = \overline{y}$, but $c(y^*)$ is given by (1); the distribution of disposable income then is such that the density is discontinuous:

$$f_c(c) = \begin{cases} f_{\overline{y}}(c - t) & \text{if } c \leq y_t; \\ f_{\overline{y}}(c - t) + f_{\overline{y}}(c) & \text{if } y_t < c \leq y_t + t; \\ f_{\overline{y}}(c) & \text{if } c > y_t + t. \end{cases}$$

Case b: Anticipated transfer. In this case supply of labor adjusts to maximize c in (1):

$$y^* = \begin{cases} \overline{y} & \text{if } \overline{y} \leq y_t \text{ or } > y_t + t; \\ y_t & \text{if } y_t < \overline{y} \leq y_t + t. \end{cases}$$

Then the distribution of labor supply (earned income) for the household is

$$\begin{aligned} F_{y^*}(y^*) &= F_{\overline{y}}(y^* - t)I(y^* < y_t) + \\ [F_{\overline{y}}(y_t) - F_{\overline{y}}(y_t - t)]I(y^* &= y_t) + F_{\overline{y}}(y^*)I(y^* > y_t), \end{aligned}$$

with indicator $I(a) = 1$ if a is true, 0 otherwise. This distribution has a mass point. Consider the joint distribution of male (y_m^*) and *female (y_f^*) labour supply where in a household we may have $y^* = y_m^* + y_f^*$. In the joint distribution $F_{y_m^*, y_f^*}(y_m^*, y_f^*)$ singularities can occur at isolated points as well as on one-dimensional subsets; to specify such subsets one would need extra information (e.g. on how decisions are made in the household).*

Mathematical examples of singular measures include self-similar and fractal measures (e.g. Lu, 1999, Frigyesi, 2004); some examples are presented in this paper. Even though in such situations the use of kernel estimators could be questionable, e.g. using a kernel estimator when density does not exist as an ordinary function, the estimators may have been used in applied work either because the problem was not taken into account (possibly as a result of it being obscure or because of neglect on the part of the researcher), or because an alternative is not available; thus determining the asymptotic properties could be helpful in interpreting existing empirical work. Of course, if the structure of the distribution is known in advance one can attempt to suit the estimator to account for mass points, discontinuities in the density, reduced dimension and the like. There are results on detecting singularities that occur in isolated interior points or on the boundary such as change points in distribution functions or conditional mean and peaks or cusps in density or hazard functions mostly starting with Muller's (1992) paper on detecting discontinuity in the conditional mean function. A general test for singularities using kernel density estimators was developed by Frigyesi and Hössjer (1998); thus kernel estimators provide useful information for inference when the distribution may have singularities.

Generally a distribution function (or a probability measure) can be represented as a mixture of absolutely continuous, discrete and singular components. Results on joint estimation in combined discrete-continuous cases (see, e.g. Ahmad and Cerrito, 1994, papers by Li and Racine, e.g. 2003) extend the area of coverage of kernel density estimation to cases when the set of variables can be partitioned into two subsets: one of variables in which the distribution is absolutely continuous (and the density is at least twice continuously differentiable), and the other of discrete variables. This leaves out cases where the density exists, but is not continuous for non-discrete variables (or even if continuous is not differentiable in some of the continuous variables); the possibility of a singular part is not considered. Lu (1999) discusses the Nadaraya-Watson estimator of the conditional mean for a special class of singular measures that are given by a local spherical measure of reduced dimension and derives the limiting Gaussian distribution. However, no general local asymptotic results for kernel density estimator or conditional mean estimator are available; this paper aims to provide a general method for obtaining such results by using the apparatus of generalized functions.

Generalized functions, or distributions as Schwartz (1950) called them, are useful in cases of non-differentiability (e.g. arising from the singularity of the distribution function) since they allow characterization of generalized derivatives when derivatives do not exist as ordinary functions. Some useful references are Halperin (1952) who provided an English introductory version of Schwartz's lectures and Gel'fand and Shilov (1964, volumes 1 and 2), where the main introductory results on generalized functions are collected. If the distribution function exists then density can always be defined as a generalized function even if it does not belong to any of the spaces of continuous/smooth functions, or any of the L_p spaces such as L_1 (Devroye and Györfi, 1985) or the usual L_2 or Sobolev or Besov spaces (considered e.g. by Härdle et.al., 1998) so the characterization as a generalized function provides the most generality.

Generalized functions are widely used in mathematics and physics for solving differential equations. Sometimes interest focuses on special generalized functions (e.g. the $\delta-$function), or on functions that form special spaces such as Sobolev or Besov spaces where some of the partial generalized derivatives exist as ordinary functions. Phillips (1991) proposed the generalized functions approach in application to the asymptotics of the LAD estimator using the fact that the generalized derivative of the *sign* function is proportional to the $\delta-$function and then in 1995 successfully applied it to derive the limit process for nonstationary LAD regression. These papers as well as others that use generalized functions[1] focus on cases where the final results are expressed through ordinary functions. This is not always possible. In Zinde-Walsh (2002) the limit process for the least median of squares estimator is described in terms of a generalized Gaussian process; Zinde-Walsh and Phillips (2003) derived the generalized Gaussian random process that represents the derivative of the fractional Wiener process; these are not expressible through ordinary functions. Gel'fand and Vilenkin (1964, volume 4) is the main reference for generalized random processes. This paper demonstrates that the limit process for the kernel density estimator may sometimes exist only as a generalized process.

Depending on the context it may be useful to think of a generalized function as an element in a variety of spaces. For example, if the interest is in the conditional mean we may view the density as a functional on the space of l times continuously differentiable functions but for the response (derivative of conditional mean) we may wish to view the density as a functional on the space of $l+1$ times differentiable functions. The introductory chapter of Sobolev's (1992) monograph (which makes use of generalized functions in approximation of multivariate integrals) provides useful diagrams of embedding mappings for different spaces of generalized functions.

The generalized derivative, f, of the distribution function, F, will be called "density" here whether or not it exists as an ordinary function. As a generalized function it can be interpreted as a linear continuous functional on a space D of special "test functions" so that for any $\psi \in D$ the value of the functional

[1] Another example is Schennach, 2004.

(f, ψ) is well defined[2]. We discuss here estimation based on random sampling; the issue of dependence is left for further development. The kernel density estimator considered as a generalized random function converges in probability as the sample size goes to ∞ and the bandwidth parameter goes to zero to the generalized derivative of the distribution function, which may or may not exist as an ordinary function; if density exists in the ordinary sense and is continuous at x the estimator converges to the value of the density function $f(x)$. The kernel estimator has a limit process that under the usual assumptions on the rate of the bandwidth is a generalized Gaussian process. A full characterization of this generalized Gaussian process is provided here.

For the Nadaraya-Watson estimator of conditional mean it is shown that the estimator, rescaled by the kernel density estimator, converges as a generalized random process to a generalized Gaussian process; a full characterization of this process is provided; this is the most general result obtainable without any assumptions on the marginal distribution. Under additional assumptions consistency and rate of convergence for the Nadaraya-Watson estimator is derived.

The paper is organized as follows. Section 2 provides interpretation of the distribution function as a generalized function and of the density as its generalized derivative; local properties are defined and the kernel estimator is interpreted as an estimator of the generalized density. In Section 3 the limit process for the kernel estimator is derived and shown to be a generalized Gaussian process. Section 4 derives asymptotic results for the Nadaraya-Watson conditional mean estimator. Section 5 concludes. Appendix A provides proofs of the results of the paper. Appendix B gives a collection of definitions and results about generalized functions and generalized random processes that are used here.

2 Distribution function and density as generalized functions and the kernel estimator

2.1 Distribution function, density and its derivatives as generalized functions

The definitions and results concerning generalized functions are collected in Appendix B; this subsection specializes these results to distribution functions.

Define for a random vector $x \in R^k$ its distribution function $F(x)$; it can be defined as a monotonic bounded function that has a left limit and is continuous from the right; thus it is an ordinary (locally summable) function on R^k as discussed here in Appendix B.

We start with the univariate case $k = 1$. As a generalized function $F(x)$ can be represented by a functional on the generic function space D. That space could be the space K of infinitely differentiable functions with finite support, or

[2] Unless F is absolutely continuous the generalized derivative may not coincide with the pointwise derivative even when it is an ordinary function.

the space S of infinitely differentiable functions that go to zero at infinity faster than any power, or any of the spaces D_m of m times continuously differentiable functions (with finite support) defined in Appendix B. For any $\psi \in D$ define the value of the functional:

$$(F, \psi) = \int F(x)\psi(x)dx. \tag{2}$$

Then as long as D contains continuously differentiable functions (e.g. coincides with K, S, or any D_m, $m \geqslant 1$) we can define density as a generalized derivative of F :

$$(f, \psi) = (F', \psi) = -(F, \psi') = -\int F(x)\psi'(x)dx. \tag{3}$$

A similar relation holds for multivariate densities. If the distribution function $F(x_1, ..., x_k)$ is absolutely continuous then the density can be defined as an ordinary function

$$f(x) = f(x_1, ..., x_k) = \frac{\partial^k F(x_1, ..., x_k)}{\partial x_1 ... \partial x_k}; \tag{4}$$

$f(x)$ integrates to $F(x)$. Whether (4) exists as an ordinary function or not, it can be defined as a generalized function. If functions in D are suitably differentiable: $D \subseteq D_k$, the space of k times continuously differentiable functions (with finite support), for any $\psi \in D$,

$$(f, \psi) = (-1)^k (F, \frac{\partial^k \psi(x_1, ..., x_k)}{\partial x_1 ... \partial x_k}). \tag{5}$$

Generalized derivatives of the density function are defined by formulas similar to (5), e.g. in the univariate case the generalized derivative of the density function, f', is given for any $\psi \in D$ (as long as $D \subseteq D_2$) by

$$(f', \psi) = -(f, \psi') = (F, \psi'').$$

2.2 Distribution function and density function locally at a point

For local properties of the distribution and density functions (generalized functions) we can consider the distribution around the point of interest x for $\tilde{x}$ in some small $h-$neighborhood of x. We introduce a kernel function, K.

Assumption A (kernel).

(a). $K(w)$ is an ordinary bounded function on R^k; $\int K(w)dw = 1$;

(b). Support of K belongs to $[-1, 1]^k$;

(c). $K(w)$ is an $l-$th order kernel: for $w = (w_1, ...w_k)$ the integral

$$\int w_1^{j_1} ... w_k^{j_k} K(w) dw_1 ... dw_k \begin{cases} = 0 & \text{if } j_1 + ... + j_k < l; \\ < \infty & \text{if } j_1 + ... + j_k = l. \end{cases}$$

If $l = 1$ Assumption 1 reduces to (a) and (b). The finite support and boundedness assumptions can be relaxed and are introduced to simplify assumptions and derivations; K is not restricted to be symmetric or non-negative.

For $k = 1$ define $K_h(\tilde{x}, x) = \frac{1}{h}K(\frac{\tilde{x}-x}{h})$. Note that $\int K_h(\tilde{x}, x)d\tilde{x} = \int K(w)dw$. Then for any h and any fixed x define

$$F_{hK}(x) \equiv \int F(\tilde{x})K_h(\tilde{x}, x)d\tilde{x} = \frac{1}{h}\int F(\tilde{x})K(\frac{\tilde{x}-x}{h})d\tilde{x}. \tag{6}$$

This provides the value of F_{hK} at x as a weighted average of values of the function $F(\tilde{x})$ in the h neighborhood[3]. Once x is allowed to vary $F_{hK}(x)$ can be viewed as a functional (generalized function) defined for functions $\psi(x) \in D$ (but it is also an ordinary function of x).

An analogous construction applies to the multivariate case. Denote for the multivariate case

$$\begin{aligned} \overline{h} &= \max\{h_1, ..., h_k\};\ x = (x_1, ..., x_k); \\ w &= (w_1, ..., w_k);\ dw = dw_1...dw_k; \qquad (7) \\ \frac{\tilde{x}-x}{h} &= \left(\frac{\tilde{x}_1 - x_1}{h_1}, ...\frac{\tilde{x}_k - x_k}{h_k}\right). \qquad (8) \end{aligned}$$

For the kernel function K define

$$K_h(\tilde{x}, x) = \frac{1}{\Pi h_i}K(\frac{\tilde{x}_1 - x_1}{h_1}, ..., \frac{\tilde{x}_k - x_k}{h_k}) = \frac{1}{\Pi h_i}K(\frac{\tilde{x}-x}{h});$$

then $\int K_h(\tilde{x}, x)d\tilde{x} = \int K(w)dw$. The functional $F_{hK}(x)$ is defined similarly to the univariate case:

$$F_{hK}(x) = \int F(\tilde{x})K_h(\tilde{x}, x)d\tilde{x}.$$

The following theorem establishes convergence of generalized functions $F_{hK}(x)$ to $F(x)$ as $\overline{h} \to 0$. To distinguish convergence of generalized functions (weak convergence of linear continuous functionals on the space D) from ordinary pointwise convergence we denote it by $\Rightarrow$ as opposed to $\to$. To distinguish between different spaces on which the functionals are defined we could subscript $\Rightarrow$ by the corresponding space, e.g. $\Rightarrow_{D_n}$. Usually it is clear for which spaces the convergence holds and the subscript is omitted.

Theorem 1 *For $\overline{h} \to 0$ and K that satisfies Assumption A*

$$F_{hK}(x) \Rightarrow F(x).$$

In other words, for any $\psi(x) \in D$ we have

$$(F_{hK}(x), \psi(x)) \to (F(x), \psi(x)),$$

and if F is continuous at x then $F_{hK}(x) \to F(x)$.

[3] Averaging of a generalized function by a kernel function was considered by e.g. Sobolev (1992) who provided the proof of a statement similar to our Theorem 1.

Proof. See Appendix A.

Next, consider the generalized derivative $f_{hK}(x)$ of $F_{hK}(x)$ that corresponds to the generalized density function locally to point x. Write for the univariate case for a given x

$$f_{hK}(x) = F'_{hK}(x). \tag{9}$$

Similarly

$$f_{hK}(x) = \frac{\partial^k F_{hK}(x)}{\partial x_1 ... \partial x_k}$$

in the multivariate case. Of course, if F were absolutely continuous in the neighborhood of x with ordinary density function $f(x)$ this would be

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{1}{h}\int F(\tilde{x})K(\frac{\tilde{x}-x}{h})d\tilde{x}\right) &= \frac{\partial}{\partial x}\left(\int F(x+hw)K(w)dw\right) \\ &= \int f(x+hw)K(w)dw \end{aligned}$$

and similarly in the multivariate case.

Assuming that K is a continuously differentiable function (9) is

$$\frac{\partial}{\partial x}\left(\frac{1}{h}\int F(\tilde{x})K(\frac{\tilde{x}-x}{h})d\tilde{x}\right) = -\frac{1}{h^2}\int F(\tilde{x})K'(\frac{\tilde{x}-x}{h})d\tilde{x},$$

thus

$$f_{hK}(x) = -\frac{1}{h}F_{hK'}(x) \tag{10}$$

is an ordinary function. Similarly

$$f_{hK}(x) = (-1)^k \frac{1}{(\Pi h_i)^2}\int F(\tilde{x})\frac{\partial^k K(\frac{\tilde{x}-x}{h})}{\partial x_1 ... \partial x_k}d\tilde{x} = (-1)^k \frac{1}{\Pi h_i}F_{hK'}(x) \tag{11}$$

in the multivariate case.

If K is not assumed to be differentiable (e.g. is a rectangular kernel) the definition (10) still holds but can be understood only as equality of generalized functions implying for $\psi \in D$:

$$(f_{hK}(x), \psi(x)) = (-\frac{1}{h}F_{hK'}(x), \psi(x)) = \frac{1}{h}(F_{hK}(x), \psi'(x)) \tag{12}$$

for all $\psi \in D$, $D \subseteq D_1$; a similar relation can be written for the multivariate case. Thus as long as either F or K is continuously differentiable $f_{hK}(x)$ is an ordinary function; in any case it is a generalized function.

If for a sequence of generalized functions, g_n, for any $\psi \in D$ and some number sequence $r(n)$ as $n \to \infty$ $(g_n, \psi) = O(r(n))$ holds, we use the notation $g_n \approx O(r(n))$.

Theorem 2 *As $\overline{h} \to 0$ and assuming that K satisfies Assumption A*
(a) convergence of generalized functions on R^k,

$$f_{hK}(x) \Rightarrow f(x),$$

holds for $D \subset D_k$; if the ordinary density function $f(x)$ exists and is continuous at x, this coincides with ordinary convergence: $f_{hK}(x) \to f(x)$;
(b) if $\psi \in D_{l+k}$ and K is a kernel of order l

$$f_{hK}(x) - f(x) \approx O(\overline{h}^l),$$

more specifically $(f_{hK}(x), \psi) - (f(x), \psi) =$

$$(-1)^l \sum_{m_1+\ldots+m_k=l} \int F(\tilde{x}) \left(\prod_{i=1}^{k} \frac{h_i^{m_i}}{m_i!} \right) \frac{\partial^{l+k}\psi}{\partial x_1^{m_{1i}+1} \ldots \partial x_k^{m_{ki}+1}}(\tilde{x}) d\tilde{x} \int K(w) w_1^{m_1} \ldots w_k^{m_{ki}} dw$$
$$+R(h),$$

where $R(h) = o(\overline{h}^l)$; if $\psi \in D_{l+k+1}$ then $R(h) = O(\overline{h}^{l+1})$.

Proof. See Appendix A.

Thus with the help of the kernel function K for $h \to 0$, we have constructed sequences of generalized functions (which may be ordinary functions e.g. if K is differentiable) that have support in a neighborhood of x and which converge as generalized functions of to the generalized derivative of the distribution function. The convergence rate of generalized functions can be controlled by appropriate selection of K and space D (as follows from (b) of Theorem 2). Of course if F is sufficiently differentiable $f_{hK}(x) - f(x) = O(\overline{h}^l)$ as ordinary functions.

The following example illustrates the case where F is not absolutely continuous and the sequence of ordinary functions f_{hK} diverges at rate $O(h^{-1})$ (even though convergence as generalized functions holds).

Example 2. Suppose that in some region Ω the distribution function can be defined as $F(x) = \alpha I(x - y)$ for some fixed y; K is differentiable. Then it is easy to compute that as $h \to 0$ convergence of ordinary functions

$$h f_{hK}(x) \to \begin{cases} 0, & \text{if } x \neq y; \\ \alpha K(0), & \text{if } x = y \end{cases}$$

holds. Thus in this case for the sequence of ordinary functions, $\sup f_{hK}(x) = O(h^{-1})$ in Ω. *It is easy to verify that as generalized functions $f_{hK}(x) \Rightarrow \alpha\delta(x - y)$, where the generalized function δ is Dirac's δ-function:*

$$(\delta(x - y), \psi(x)) = \psi(y).$$

2.3 The kernel estimator and its relation to generalized density

Consider a multivariate density; recall the notation in (8). From (11) integrating by parts we have

$$\begin{aligned} f_{hK}(x) &= (-1)^k \frac{1}{(\Pi h_i)^2} \int F(\tilde{x}) \frac{\partial^k K(\frac{\tilde{x}-x}{h})}{\partial x_1 ... \partial x_k} d\tilde{x} \\ &= \frac{1}{\Pi h_i} \int K(\frac{\tilde{x}-x}{h}) dF(\tilde{x}) = E_{\tilde{x}} K_h(\tilde{x}, x). \end{aligned}$$

If f_{hK} is not an ordinary function of x the equality may hold as equality of generalized functions only. A natural estimator for $f_{hK}(x)$ follows from the fact that it is an expectation; a sample average is used in estimation. The estimator based on a random sample of n observations $\{x_i\}$ from the distribution of x is

$$\widehat{f(x)} = \frac{1}{n \Pi h_i} \sum_{i=1}^{n} K(\frac{x_i - x}{h}) \tag{13}$$

and $E\widehat{f(x)} = f_{hK}(x)$.

We can thus interpret the kernel density estimator as an estimator of the local generalized density functional whether density exists as an ordinary function or not.

3 Limit process for the kernel estimator of generalized density

We now describe the limit process for the kernel estimator as a generalized random process. Note that since $E\widehat{f(x)} = f_{hK}(x)$ part (b) of Theorem 2 provides the convergence rate for the generalized bias function of the kernel estimator.

The following theorem describes the limit process of the kernel density estimator as a generalized Gaussian process (defined here in Appendix B).

Theorem 3 *For a kernel function K satisfying Assumption A, if $\overline{h} \to 0$ as $n \to \infty$ with $n\Pi h_i \to \infty$ and $\overline{h}^{2l+k} n \to 0$, the sequence of generalized random processes $(n\Pi h_i)^{\frac{1}{2}} \left(\widehat{f(x)} - f(x) \right)$ converges to a generalized Gaussian process with mean functional zero and covariance functional C which for any (linearly independent) $\psi_1, \psi_2 \in D_{l+k}$ provides*

$$\begin{aligned} (C, (\psi_1, \psi_2)) &= \int \psi_1(x) \psi_2(x) f(x) dx \int K(w)^2 dw \\ &= E(\psi_1(x) \psi_2(x)) \int K(w)^2 dw. \end{aligned} \tag{14}$$

Proof. See Appendix A.

If the density function $f(x)$ is continuous at x the covariance functional reduces to a functional given by an ordinary function, and the standard covariance result $f(x)\int K(w)^2 dw$ obtains. If $f(x) = 0$ the ordinary Gaussian process is degenerate and only the consistency rate holds. In the general case if for any $\psi \in D$ (not identically zero) the value $(C, (\psi, \psi))$ is positive the limit process in Theorem 3 is a proper (non-degenerate) generalized Gaussian process; here this holds as long as $E(\psi^2(x)) > 0$ for every non-trivial ψ. If the limit process is degenerate, consistency holds and as long as for some ψ the expectation $E(\psi^2(x))$ is positive there is a limit Gaussian process. Degeneracy of the generalized limit process process can result from the continuous density taking zero values on some open set and also can occur for distributions with singularities. The following examples present some cases where the distribution is not absolutely continuous.

Example 3. Discrete variables.

Case 1. When the discrete variables are represented correctly and the associated density is viewed as discrete, appropriate nonparametric estimators (e.g. those based on special kernels as in Ahmad and Cerrito, 1994) converge at parametric rates. In discrete-continuous models product kernels (e.g. Racine and Li, 2003) provide convergence at non-parametric rates that reflect the dimension of the continuous part only (alleviating "curse of dimensionality").

Case 2. When discrete variables are mislabelled as continuous, Example 2 demonstrates that if K rather than K_h were used as the kernel for the discrete variable the corresponding kernel estimator would converge as $n \to \infty$ to $\alpha K(0)$, where α is the discrete density, since

$$h\widehat{f(y)} \ \to \alpha K(0).$$

In other words, a rescaling of the kernel produces faster convergence (this can be viewed as a trivial case of reduced local dimension similar to the case in Example 5). The limit covariance functional is degenerate: $(C, (\psi, \psi)) = 0$ if $\psi(y) = 0$ at the mass point y.

Example 4. Suppose that the distribution is a mixture of an absolutely continuous strictly monotonic $F^c(x)$ and a discrete $F^d(x) = I(x - y)$, with weights α and $1 - \alpha$; the generalized density then is non-singular at every $x \neq y$ and is singular at y, and can be represented as a sum $\alpha f^c(y) + (1 - \alpha)\delta(x - y)$. Here $h\widehat{f(y)} \to (1-\alpha)K(0)$; unlike the previous example the limit of the rescaled kernel estimator will not provide a full description of the limit density since it misses the continuous component. The limit covariance functional is not degenerate.

Example 5. Reduced local dimension. Cantor-type distribution. Z-Q Lu (1999) discusses this famous example of reduced local dimension. Consider a random variable

$$\eta = 2\Sigma_{i=1}^{\infty} 3^{-i}\xi_i,$$

where ξ_i is Bernoulli (equals 1 with probability 1/2, 0 with probability 1/2). If one thinks of $\eta \in [0, 1]$ as a number in base 3 its digits are either 0 or 2 (no

ones) and any such number is equally probable. The distribution function on $[0,1]$ *is continuous but not absolutely continuous and for any point* $x \in [0,1]$ *and some* $r \to 0$

$$\Pr(\|u - x\| \le r) = c(x) r^{\log 2/\log 3}$$

which defines local dimension of the measure as $d = \log 2/\log 3 < 1$. *Consider a point* x *that does not have ones in its base three representation, then the sequence of kernel estimator functions* $\widehat{f(x)}$ *diverges, but at rate* $O(h^{d-1})$*: the rescaled function* $h^{1-d}\widehat{f(x)}$ *has a finite limit. The limit generalized Gaussian process is degenerate: for any neighborhood* U of x and $\psi \in D(U)$ *there exists an interval* $I \subset U$ *(of points with a one in some position of the base three representation) such that* $(C,(\psi,\psi)) = 0$ *for* ψ *with support in* I. *However if the distribution is a mixture of an absolutely continuous strictly monotonic and the Cantor distributions, the limit generalized process is no longer degenerate.*

These examples indicate that utilizing knowledge regarding the nature of the variables, in particular, local dimension (ranging from discrete variable cases to redundant variables to fractional dimension) and rescaling the kernel appropriately, e.g. by h for discrete variables and h^{1-d} for reduced local dimension (example 5) may result in a faster convergence rate. However, if it cannot be firmly established that reduced dimension is relevant (as e.g. in a mixture with an absolutely continuous measure) rescaling may lead to error and only the more general rates and results such as those provided in Theorems 2 and 3 apply.

The results on density estimation can be extended to estimation of derivatives of density. Under conditions of Theorem 3 as generalized functions on the space D_{l+2k} derivatives of the kernel estimator have as a limit process the generalized Gaussian process with mean functional zero and covariance functional given by generalized derivatives of the limit covariance functional for the kernel estimator itself, and so the covariance functional for the limit process for the derivatives of the kernel density estimator can be derived from C (see Appendix B).

4 Relaxing assumptions on the marginal distribution in the asymptotics of Nadaraya-Watson estimator

Nadaraya-Watson estimator of the conditional mean is widely used. Here general limit results that do not require making any assumptions on the marginal distribution are derived; additional assumptions provide consistency and the usual limit Gaussian distribution.

Consider the model

$$y_i = m(x_i) + u_i$$

where the conditional expectation function is

$$m(x) = E(y_i \,|x_i = x\,) = \int y dF(y\,|x\,);$$

$\sigma_u^2(x)$ is the conditional variance

$$\sigma_u^2(x) = var(u_i \,|x_i = x\,) = var(y_i \,|x_i = x\,).$$

Assume for simplicity that all bandwidths are equal: $h_i = h$. The Nadaraya-Watson kernel estimator of the conditional mean is

$$\widehat{m(x)} = \frac{\Sigma y_j K(\frac{x_j - x}{h})}{\Sigma K(\frac{x_j - x}{h})} = \frac{\frac{1}{nh^k}\Sigma y_j K(\frac{x_j - x}{h})}{\widehat{f(x)}}.$$

To accommodate various assumptions regarding smoothness of the conditional mean function recall the definition of a Hölder space $C_{v+\alpha}(E)$ for integer $v \geq 0$ and $0 < \alpha \leq 1$ (e.g. Mathematical Encyclopedia, 1977). It is a Banach space of bounded and continuous functions $g(x)$ defined on a set $E \subseteq R^k$ which are v times continuously differentiable with all the $v-th$ order derivatives $g_\iota^{(v)}(x)$ satisfying Hölder's condition for α :

$$\left| g_\iota^{(v)}(x + \Delta x) - g_\iota^{(v)}(x) \right| \leq A(x) \left\| \Delta x \right\|^\alpha$$

for every $x, x + \Delta x \in E$.

If $a = 1$ the $v-$th derivatives satisfy the Lipschitz condition; if $v = 0$ the function itself satisfies the Hölder (fractional Lipschitz) condition.

Denote by U some open neighborhood of x in R^k.

Assumption B (conditional moments).

(a). The conditional mean function $m(x) \in C_{v+\alpha}(U)$.

(b). The conditional variance $\sigma_u^2(x)$ is continuous on U and for some $\delta > 0$ $\sigma_u^{2+\delta}(x)$ is uniformly bounded on U.

Assumption B(a) for $v \geq 2$ implies assumptions often made in the literature on asymptotics of the Nadaraya-Watson estimator which discusses asymptotic bias reduction (Bierens, 1987), while for example Lu (1999) argues that if the conditional mean is differentiable the local linear estimator (see, e.g. Fan and Gijbels, 1995) is preferable[4] and thus for the kernel estimator it makes more sense to assume $v = 0$, i.e. a fractional Lipschitz (Hölder) condition. The following Theorems examine the behavior of the Nadaraya-Watson estimator under assumptions that include both of these cases. The space of test functions, D, here is restricted to functions with support contained in U, $D(U)$.

Theorem 4 *(a) If as $n \to \infty$ the bandwidth $h \to 0$ and $nh^k \to \infty$; Assumption A holds for $l \geq 1$, assumption B(a) holds with $v \geq l$, B(b) holds and $\lim_{h \to 0} h^l \left(nh^k\right)^{1/2} = 0$, then*

$$(nh^k)^{\frac{1}{2}} \widehat{f(x)} \left(\widehat{m(x)} - m(x) \right)$$

[4]Local linear and local polynomial estimators perform better than Nadaraya-Watson estimator when density is not smooth but the properties of those estimators, such as mean square error, have only been derived for continuous density and are not yet available for distributions with singularities.

converges as a generalized random process on $D_{l+k}(U)$ to a generalized Gaussian process with expectation functional zero and limit covariance functional C_x given by

$$(C_x, (\psi_1, \psi_2)) = E\left(\psi_1(x)\psi_2(x)\sigma^2(x)\right) \int K(w)^2 dw. \tag{15}$$

(b) If as $n \to \infty$ the bandwidth $h \to 0$ and $nh^k \to \infty$; Assumption A(a-b) holds, assumption B(a) holds with $v = 0, 0 < \alpha \leq 1$, B(b) holds and $\lim_{h\to 0} h^\alpha \left(nh^k\right)^{1/2} = 0$, then

$$(nh^k)^{\frac{1}{2}} \widehat{f(x)} \left(\widehat{m(x)} - m(x)\right)$$

converges in distribution to a generalized Gaussian process on $D_k(U)$ with expectation functional zero and the covariance functional C_x given by (15).

Proof. See Appendix.

The Theorem in parts (a) and (b) provides an asymptotic (generalized) Gaussian process for he conditional mean estimator weighted by $\widehat{f(x)}$; as in Theorem 3 the limit generalized Gaussian process may be degenerate. When continuous density exists results such as Theorem 4 are viewed as intermediate (e.g. Bierens, 1987) to obtaining the limit distribution for $\widehat{m(x)}$ itself; here this limit distribution is given in Corollary 2. Without further assumptions on the behavior of $\widehat{f(x)}$ only the general result of Theorem 4 holds. It could be used to construct confidence intervals.

The following assumption allows to examine the limiting behavior for $\widehat{m(x)}$.

Assumption C (condition on $\widehat{f(x)}$) *As $n \to \infty, h \to 0$ for some η there exists $b > 0$ such that for $x \in U$*

$$\Pr(\widehat{f(x)} > h^\eta b) \to 1.$$

This assumption is a condition on the density estimator; it permits convergence to zero for $\eta > 0$. When an ordinary density function that is continuous and positive at x exists, the assumption obviously holds in some U with $\eta = 0$. The following corollary establishes the consistency rate for $\widehat{m(x)}$.

Corollary 1. *Under the conditions of Theorem 4 and Assumption C, for bandwidths that additionally satisfy $nh^{k+2\eta} \to \infty$ when $\eta > 0$, the generalized random process satisfies*

$$(nh^{k+2\eta})^{\frac{1}{2}} \left(\widehat{m(x)} - m(x)\right) \approx O_p(1).$$

Proof. See Appendix A.

The next corollary provides the limit process in the case of a continuous density function.

Corollary 2. *Under the conditions of Theorem 4 and assuming that an ordinary density function $f(\cdot)$ exists, is continuous at x* and *$f(x) > 0$, the*

sequence of ordinary random functions $(nh^k)^{\frac{1}{2}}\left(\widehat{m(x)}-m(x)\right)$ *converges as a generalized sequence to an (ordinary) Gaussian distribution*

$$(nh^k)^{\frac{1}{2}}\left(\widehat{m(x)}-m(x)\right)\Rightarrow_d N\left(0,\frac{\sigma^2(x)}{f(x)}\int K(w)^2dw\right); \tag{16}$$

if ordinary convergence to a limit distribution for $(nh^k)^{\frac{1}{2}}\left(\widehat{m(x)}-m(x)\right)$ *obtains, convergence to the Gaussian distribution* (16) *holds in an ordinary sense.*

Proof. See Appendix A.

A standard sufficient condition for ordinary convergence for $(nh^k)^{\frac{1}{2}}\left(\widehat{m(x)}-m(x)\right)$ in Corollary 2 is that the density be l times continuously differentiable (Bierens, 1987). Assumptions on h in Theorem 4 and Corollary 2 imply zero asymptotic bias. More generally, denote $\lim_{h\to 0}h^l\left(nh^k\right)^{1/2}$ by β; then from (20) in Appendix A the asymptotic bias functional can be written as

$$\beta\sum_{\substack{s_1+...+s_k=r=l-t;\\ l-1\geq t=m_1+...m_k\geq 0}}\frac{1}{m_1!...m_k!}\frac{1}{s_1!...s_k!}\frac{\partial^t}{\partial\tilde{x}_1^{m_1}...d\tilde{x}_k^{m_k}}\left(\frac{\partial^r m(x)}{\partial\tilde{x}_1^{s_1}...d\tilde{x}_k^{s_k}}f(x)\right)$$
$$\times\int w_1^{m_1+s_1}...w_k^{m_k+s_k}K(w)dw.$$

If $\beta=0$ the asymptotic bias is a generalized bias function that is given by an ordinary function ($\equiv 0$) and Corollary 2 applies. If $0<\beta<\infty$, the asymptotic bias may not be given by an ordinary function unless the density is appropriately smooth.

5 Conclusions

For kernel estimator of density $\widehat{f(x)}$ and for the Nadaraya-Watson estimator $\widehat{m(x)}$ of the conditional mean in the i.i.d. case this paper provides general asymptotic results that can be obtained without making any assumptions regarding the underlying distribution of x. Density can always be defined as a generalized derivative of the distribution function and the kernel estimator $\widehat{f(x)}$ represents an estimator of the generalized density, f. Limit processes for the estimators in Theorems 3 and 4 for appropriate rates of bandwidths are described as generalized Gaussian processes with mean functional zero and a non-zero covariance functional (possibly degenerate).

The generalized functions approach can be extended to deriving limit moments and limit processes for other estimators in situations where it may be of benefit to relax assumptions regarding existence and smoothness of the density function.

6 Appendix A.

Proof of Theorem 1.

We have

$$\begin{aligned} & \lim_{h\to 0}\int\int F(x+hw)K(w)dw\psi(x)dx \\ = & \lim_{h\to 0}\int\int F(y)\psi(y-hw)K(w)dwdy \\ = & \int F(y)\psi(y)dy\int K(w)dw=(F,\psi) \end{aligned}$$

where the second equality follows from ψ being a continuous function. If F is continuous at x then by interchanging the integral and limit in the first line we get

$$\lim_{h\to 0}\int F(x+hw)K(w)dw=F(x).$$

■

Proof of Theorem 2.

(a) Since ψ is a continuously differentiable function ($\psi\in D_k$)

$$\begin{aligned} (f_{hK},\psi) &= (-1)^k\int\frac{\partial^k}{\partial x_1...\partial x_k}\left(\frac{1}{\Pi h_i}\int F(\tilde{x})K(\frac{\tilde{x}-x}{h})d\tilde{x}\right)\psi(x)dx \\ &= \int\left(\frac{1}{\Pi h_i}\int F(\tilde{x})K(\frac{\tilde{x}-x}{h})d\tilde{x}\right)\frac{\partial^k}{\partial x_1...\partial x_k}\psi(x)dx \\ &= (-1)^k\int\int F(\tilde{x})\frac{\partial^k}{\partial x_1...\partial x_k}\psi(\tilde{x}-hw)K(w)dwdy \qquad (17) \\ &\to (f,\psi)\int K(w)dw=(f,\psi) \end{aligned}$$

where the second equality is obtained by integration by parts and using finite support property of ψ; the third equality follows from change of variable; the last result uses continuity of $\frac{\partial^k}{\partial x_1...\partial x_k}\psi$ and $\int K=1$.

Consider f_{hK} defined in (9). For F univariate absolutely continuous in the neighborhood of x with continuous density function $f(x)$

$$\begin{aligned} f_{hK}(x) &= \frac{\partial}{\partial x}\left(\frac{1}{h}\int F(\tilde{x})K(\frac{\tilde{x}-x}{h})d\tilde{x}\right)=\int f(x+hw)K(w)dw \\ &\to f(x)\int K(w)dw=f(x) \end{aligned}$$

by change of variable, continuity of f and the assumption $\int K=1$, similarly in the multivariate case.

(b) For the univariate case using (12) and applying change of variable

$$(f_{hK},\psi)-(f,\psi)=-\int\int F(\tilde{x})[\psi'(\tilde{x}-hw)-\psi'(\tilde{x})]K(w)dwd\tilde{x}.$$

From expanding $\psi'(\tilde{x}-hw)$ and making use of kernel order this becomes

$$(-1)^l \frac{h^l}{l!} \int F(\tilde{x})\psi^{l+1}(\tilde{x})d\tilde{x} \int K(w)w^l dw + R(\overline{h})$$

where the remainder term $R(\overline{h})$ is $o(\overline{h}^l)$ and if $\psi \in D_{l+1}$ is $O(\overline{h}^{l+1})$. Consider now the multivariate case: expansion of $\frac{\partial^k}{\partial x_1 ... \partial x_k}\psi(\tilde{x}-hw) =$

$$\sum_{m_1+...+m_k=1}^{l} \left(\frac{1}{(m_1+1)!...(m_k+1)!} \frac{\partial^{m_1+...+m_k+k}\psi(\tilde{x})}{\partial \tilde{x}_1^{m_1+1}...d\tilde{x}_k^{m_k+1}} \right) (-h)^{m_1+...+m_k} w_1^{m_1}...w_k^{m_k} dw +$$

$$+R(\overline{h}) \text{ with } R(\overline{h}) = o(\overline{h}^l);\ \text{if } \psi \in D_{l+k+1},\ R(\overline{h}) = O(\overline{h}^{l+1}).$$

Substituting into (17), subtracting the limit and using the order of kernel we get that

$$\begin{aligned}
& (f_{hK},\psi) - (f,\psi) \\
= \ & (-1)^l \int F(\tilde{x}) \left[\sum_{m_1+...+m_k=l} \left(\frac{1}{(m_1+1)!...(m_k+1)!} \frac{\partial^{l+k}\psi(\tilde{x})}{\partial \tilde{x}_1^{m_1+1}...d\tilde{x}_k^{m_k+1}} \right) (-h)^l + R(\overline{h}) \right] d\tilde{x} \\
& \int K(w) w_1^{m_1}...w_k^{m_k} dw_1...dw_k + R(\overline{h}) \\
= \ & O(h^l).
\end{aligned}$$

■

Proof of Theorem 3.

Define a (generalized) function

$$e_{nhj}(x) = \frac{1}{\Pi h_i} K(\frac{x-x_j}{h}) - f(x)$$

and consider $e_{hn}(x) = \frac{1}{n}\sum_{j=1}^{n} e_{nhj}(x)$; it equals $\widehat{f(x)} - f(x)$.

By Theorem 2 (b) $Ee_{hni}(x) \approx O(\overline{h}^l)$.

Next consider $T_{ij} = E(e_{hni}(x),\psi_1)(e_{hnj}(x),\psi_2))$.

For $i \neq j$ by independence

$$\begin{aligned}
E(T_{ij}) &= E(e_{hni}(x),\psi_1)(e_{hnj}(x),\psi_2) = E(e_{hni}(x),\psi_1)E(e_{hnj}(x),\psi_2) \\
&= O(\overline{h}^{2l}).
\end{aligned}$$

For $i = j$

$$\begin{aligned}
E(T_{ii}) &= E(e_{hni}(x),\psi_1)(e_{hni}(x),\psi_2) \\
&= \int \left[\int \frac{1}{\Pi h_i} K(\frac{x_i - x}{h})\psi_1(x)dx - \int f(x)\psi_1(x)dx \right] \times \\
&\quad \left[\int \frac{1}{\Pi h_i} K(\frac{x_i - x}{h})\psi_2(x)dx - \int f(x)\psi_2(x)dx \right] f(x_i)dx_i \\
&= E(T_{ii}^1 + T_{ii}^2),
\end{aligned}$$

where

$$T_{ii}^1 = \left(\int \frac{1}{\Pi h_i} K(\frac{x_i - x}{h}) \psi_1(x) dx \right) \left(\int \frac{1}{\Pi h_i} K(\frac{x_i - x}{h}) \psi_1(x) dx \right).$$

It is easy to see that $ET_{ii}^2 = -E\psi_1 E\psi_2$. To express ET_{ii}^1 as a bilinear functional applied to (ψ_1, ψ_2) the order of integration has to be changed. Consider now for any fixed (x, y)

$$\int \frac{1}{\Pi h_i^2} K(\frac{x_i - x}{h}) K(\frac{x_i - y}{h}) f(x_i) dx_i.$$

Substituting $\frac{x_i - x}{h} = w$ this becomes

$$\int \frac{1}{\Pi h_i} K(w) K(w + \frac{x - y}{h}) f(x + hw) dw;$$

if $x - y \neq 0$ for small enough h we have $\left| \frac{x-y}{h} \right| > 2 \left| support(K) \right|$ and then $K(w + \frac{x-y}{h}) = 0$. If $x = y$ this expression multiplied by Πh_i becomes

$$\int K(w)^2 f(x + hw) dw.$$

By Theorem 2 as $\overline{h} \to 0$ it converges as a generalized function to $f(x) \int K(w)^2 dw$. Thus $\Pi h_i ET_{ii}^1 \to \int K(w)^2 dw E(\psi_1 \psi_2)$; it is easy to see that $\Pi h_i ET_{ii}^2 \to 0$ and then $\Pi h_i ET_{ii} \to \int K(w)^2 dw E(\psi_1 \psi_2)$. This provides the limit covariance.

Consider now

$$\begin{aligned} \eta_{hni}(x) &= n^{\frac{1}{2}} \Pi h_i^{\frac{1}{2}} e_{hni}(x) - E(n^{\frac{1}{2}} \Pi h_i^{\frac{1}{2}} e_{hni}(x)); \\ \eta_{hn}(x) &= \frac{1}{n} \sum \eta_{hni}(x). \end{aligned} \tag{18}$$

This generalized random function has expectation zero. In the covariance the terms where $i \neq j$ are zero and

$$E(\eta_{hni}(x), \psi_1)(\eta_{hni}(x), \psi_2)$$

is $O(1)$ and converges to $\int K(w)^2 dw E(\psi_1 \psi_2)$.

Next we show that for any set of linearly independent functions $\psi_1, ..., \psi_m \in D$ with $E(\psi_i^2) > 0$ the joint distribution of the vector $\overrightarrow{\eta}_{hn} = ((\eta_{hn}, \psi_1) ..., (\eta_{hn}, \psi_m))'$ converges to a multivariate Gaussian. Define similarly the vector $\overrightarrow{\eta}_{hni}$ with components (η_{hni}, ψ_l). Denote by Σ the $m \times m$ matrix with ts component $\{\Sigma\}_{ts} = (C, (\psi_t, \psi_s))$ where the functional C is given by (14). Denote by $\hat{\Sigma}_n$ the covariance matrix of $\overrightarrow{\eta}_{hni}$. By the convergence results for T_{ij}, $\hat{\Sigma}_n \to_p \Sigma$. Since the functions $\psi_1, ..., \psi_m$ are linearly independent and $E(\psi_i^2) > 0$ the matrix Σ and thus $\hat{\Sigma}$ (in probability for large enough n) is invertible. Define ξ_{hni} to equal $\hat{\Sigma}^{-1/2} \overrightarrow{\eta}_{hni}$, then $\hat{\Sigma}^{-1/2} \overrightarrow{\eta}_{hni} - \Sigma^{-1/2} \overrightarrow{\eta}_{hni} \to_p 0$.

Next, consider an $m \times 1$ vector λ with $\lambda' \lambda = 1$. The random variables $\lambda' \xi_{hni}$ are independent with expectation 0, $var \sum \lambda' \xi_{hni} = 1$; they satisfy the Liapunov

condition: $\sum E\left|\lambda'\xi_{hni}\right|^{2+\delta} \to 0$ for $\delta > 0$ since the kernel function is bounded with finite support. Thus

$$\sum \lambda'\xi_{hni} \to_d N(0,1)$$

and by the Cramer-Wold theorem convergence to a limit Gaussian process for $\hat{\Sigma}_n^{-1/2}\overrightarrow{\eta}_{hn}$ and thus for $\Sigma^{-1/2}\overrightarrow{\eta}_{hn}$ follows. If there exists a non-trivial ψ (linearly independent of $(\psi_1...,\psi_m)$) such that $E\psi^2 = 0$ then for any any ψ_i the limit covariance functional provides $(C,(\psi,\psi_i)) = 0$ in which case the limit process for $\overrightarrow{\eta}_{hn} = ((\eta_{hn},\psi_1)...,(\eta_{hn},\psi_m),(\eta_{hn},\psi))'$ is a degenerate Gaussian process.■

Proof of Theorem 4.

First examine the mean functional. For the generalized expectation by using iterated expectation and then the usual change of variable $\frac{x_i-x}{h} = w$

$$\begin{aligned}
E(\widehat{m(x)}\widehat{f(x)} - m(x)\widehat{f(x)}) &= \frac{1}{h^k}E[E_{|x_i}(y_iK(\frac{x_i-x}{h}) - m(x)K(\frac{x_i-x}{h})] \\
&= \frac{1}{h^k}E[m(x_i)K(\frac{x_i-x}{h}) - m(x)K(\frac{x_i-x}{h})] \\
&= \int \frac{1}{h^k}[m(x_i)K(\frac{x_i-x}{h}) - m(x)K(\frac{x_i-x}{h})]f(x_i)dx_i \\
&= \int [m(x+hw)-m(x)]K(w)f(x+hw)dw.
\end{aligned}$$

For any test function $\psi(x) \in D$ write

$$\begin{aligned}
&E\left((\widehat{m(x)}\widehat{f(x)} - m(x)\widehat{f(x)}),\psi(x)\right) = \\
&= \int [m(x+hw)-m(x)]K(w)f(x+hw)dw\psi(x)dx \\
&= \int [m(\tilde{x})-m(\tilde{x}-hw)]K(w)f(\tilde{x})\psi(\tilde{x}-hw)dwd\tilde{x} \qquad (19)
\end{aligned}$$

where the last line is obtained by the change of variable $\tilde{x} = x + hw$.

In case (a) the expansion (in ordinary continuous functions)

$$\begin{aligned}
m(\tilde{x}) - m(\tilde{x}-hw) &= \sum_{s_1+...s_k=1}^{l} h^t \frac{1}{s_1!...s_k!}\frac{\partial^t m(\tilde{x})}{\partial \tilde{x}_1^{s_1}...d\tilde{x}_k^{s_k}} w_1^{s_1}...w_k^{s_k} + R, \\
\text{with } R &= O(h^{l+\alpha})
\end{aligned}$$

holds for $l \geq 1$. Since $\psi \in D_{l+k}$ we can expand

$$\psi(\tilde{x}-hw) = \sum_{m_1+...+m_k=0}^{l} \left(\frac{1}{m_1!...m_k!}\frac{\partial^{m_1+...+m_k}\psi(\tilde{x})}{\partial y_1^{m_1}...dy_k^{m_k}}\right)(-h)^l w_1^{m_1}...dw_k^{m_k} + o(h^l).$$

Substituting the expansions into (19) and combining with kernel order we get that the limit mean functional in part (a) is $O(h^l)$ and can be expressed as

$$h^l \sum_{\substack{s_1+...+s_k=r=l-t;\\ l-1\geq t=m_1+...m_k\geq 0}} \frac{1}{m_1!...m_k!}\frac{1}{s_1!...s_k!}\frac{\partial^t}{\partial\tilde{x}_1^{m_1}...d\tilde{x}_k^{m_k}}\left(\frac{\partial^r m(x)}{\partial\tilde{x}_1^{s_1}...d\tilde{x}_k^{s_k}}f(x)\right)$$
$$\times\int w_1^{m_1+s_1}...w_k^{m_k+s_k}K(w)dw \tag{20}$$

where the appropriate generalized derivatives of the density exist for $\psi\in D_{l+k}$. Under the conditions on the bandwidth the mean functional for $(nh^k)^{1/2}(\widehat{m(x)}\widehat{f(x)}-m(x)\widehat{f(x)})$ converges to zero.

For case (b) using in (19) Hölder continuity of $m(x)$ (Assumption B(a) for $v=0, 0<\alpha\leq 1$) together with boundedness of the kernel function, smoothness of ψ and boundedness of $\int f(\tilde{x})\psi(\tilde{x})d\tilde{x}$ we get that $\left|E\left((\widehat{m(x)}\widehat{f(x)}-m(x)\widehat{f(x)}),\psi(x)\right)\right| = O(h^\alpha)$ and thus the mean functional for $(nh^k)^{1/2}(\widehat{m(x)}\widehat{f(x)}-m(x)\widehat{f(x)})$ converges to zero under the condition on the bandwidth.

Next we derive the covariance functional for both (a) and (b). Define a (generalized) function

$$e_{nhj}(x)=\frac{1}{h^k}\left(y_iK(\frac{x-x_j}{h})-m(x)K(\frac{x-x_j}{h})\right)$$

and consider $e_{hn}(x)=\frac{1}{n}\sum_{j=1}^n e_{nhj}(x)$; it equals $\widehat{m(x)}\widehat{f(x)}-m(x)\widehat{f(x)}$.

Next consider $T_{ij}=E(e_{hni}(x),\psi_1)(e_{hnj}(x),\psi_2))$.

For $i\neq j$ by independence

$$\begin{aligned} E(T_{ij}) &= E(e_{hni}(x),\psi_1)(e_{hnj}(x),\psi_2)=E(e_{hni}(x),\psi_1)E(e_{hnj}(x),\psi_2)\\ &= \begin{cases} O(h^{2l}) & \text{for case (a);}\\ O(h^{2\alpha}) & \text{for case (b).}\end{cases}\end{aligned}$$

For $i=j$ we have $E(T_{ii})=$

$$\begin{aligned} & E(e_{hni}(x),\psi_1)(e_{hni}(x),\psi_2)\\ = & \int E_{|x_i}\left[\int\frac{1}{h^k}K(\frac{x_i-x}{h})\left(y_i-m(x)\right)\psi_1(x)dx\right]\\ & \times\left[\int\frac{1}{h^k}K(\frac{x_i-x}{h})\left(y_i-m(x)\right)\psi_1(x)dx\right]f(x_i)dx_i.\end{aligned}$$

To express ET_{ii} as a bilinear functional applied to (ψ_1,ψ_2) the order of integration has to be changed. For any fixed (x,y)

$$\int\frac{1}{h^{2k}}E_{|x_i}\left(y_i-m(x)\right)\left(y_i-m(y)\right)K(\frac{x_i-x}{h})K(\frac{x_i-y}{h})f(x_i)dx_i.$$

Substituting $\frac{x_i-x}{h} = w$ this becomes

$$\int \frac{1}{h^k} E_{|x_i} \left(y_i - m(x+hw)\right)\left(y_i - m(y)\right) K(w) K(w + \frac{x-y}{h}) f(x+hw) dw;$$

if $x - y \neq 0$ for small enough h we have $\left|\frac{x-y}{h}\right| > 2\left|support(K)\right|$ and then $K(w + \frac{x-y}{h}) = 0$. If $x = y$ this expression multiplied by h^k becomes

$$\int E_{|x+hw} \left(y_i - m(x)\right)^2 K(w)^2 f(x+hw) dw. \tag{21}$$

We have $E_{|x+hw} \left(y_i - m(x)\right)^2 =$

$$\begin{aligned} & E_{|x+hw} \left(y_i - m(x+hw)\right)^2 - 2(m(x+hw) - m(x)) E_{|x+hw} \left(y_i - m(x+hw)\right) \\ & +(m(x+hw) - m(x))^2 \\ = \quad & \sigma^2(x_i) + r \end{aligned}$$

with (since $E_{|x+hw} \left(y_i - m(x+hw) = 0\right)$

$$r = (m(x+hw) - m(x))^2 = \begin{cases} O(h^2) & \text{in case (a);} \\ O(h^{2\alpha}) & \text{in case (b)} \end{cases}$$

uniformly over U. As $h \to 0$ (21) converges as a generalized function to $f(x)\sigma^2(x) \int K(w)^2 dw$ by Assumption B(b) and Theorem 2. Thus

$$h^k E T_{ii} \to E(\psi_1(x)\psi_2(x)\sigma^2(x)) \int K(w)^2 dw.$$

This provides the limit covariance functional C_x in (15).

Consider now

$$\begin{aligned} \eta_{hni}(x) &= n^{\frac{1}{2}} h^{\frac{k}{2}} e_{hni}(x) - E(n^{\frac{1}{2}} h^k e_{hni}(x)); \\ \eta_{hn}(x) &= \frac{1}{n} \sum \eta_{hni}(x). \end{aligned} \tag{22}$$

This generalized random function has expectation zero. In the covariance the terms where $i \neq j$ are zero and

$$E(\eta_{hni}(x), \psi_1)(\eta_{hni}(x), \psi_2)$$

is $O(1)$ and converges to $E(\psi_1(x)\psi_2(x)\sigma^2(x)) \int K(w)^2 dw$.

The proof that the limit is a Guassian distribution is identical to the concluding part of the proof of Theorem 3. Here for any set of linearly independent functions $\psi_1, ..., \psi_m \in D$ the joint distribution of the vector $\overrightarrow{\eta}_{hn} = ((\eta_{hn}, \psi_1)..., (\eta_{hn}, \psi_m))'$ converges to a multivariate Gaussian. If $E(\psi_i^2 \sigma^2) > 0$ the limit distribution is non-degenerate; if for some ψ the value of the covariance functional $E(\psi^2\sigma^2) = 0$ the limit process is degenerate.

■

Proof of Corollary 1.
Write

$$\begin{aligned} & (nh^{k+2\eta})^{1/2}(\widehat{m(x)} - m(x)) \\ = \quad & \left(b^{-1}(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)} - m(x))\right) b(\widehat{f(x)}h^{-\eta})^{-1}. \end{aligned} \tag{23}$$

From Theorem 4 the first factor in the product in (23) as a generalized random process is bounded in probability

$$b^{-1}(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)} - m(x)) \approx O_p(1);$$

from Assumption C

$$\Pr\left(b\left|\widehat{f(x)}\right| h^{-\eta-1} < 1\right) \to 1.$$

thus the ordinary function $b(\widehat{f(x)}h^{-\eta})^{-1} = O_p(1)$ and

$$(nh^{k+2\eta})^{1/2}(\widehat{m(x)} - m(x)) \approx O_p(1).$$

■

Proof of Corollary 2.

Under the conditions of the corollary $(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)} - m(x))$ converges by Theorem 4 to a generalized Gaussian process with the limit mean functional zero and the limit covariance functional given by an ordinary function $\sigma^2(x)f(x)\int K(w)^2dw$. The estimator is an ordinary integrable random variable in U that as a generalized random variable converges to a Gaussian generalized limit process that is given by an ordinary Gaussian zero-mean process. Thus the generalized limit process for the estimator is an ordinary Gaussian process

$$N(0, \sigma^2(x)f(x)\int K(w)^2dw).$$

The fact that it has zero mean follows from the condition on h and does not require differentiability of $f(x)$. Consider

$$\begin{aligned} & (nh^k)^{1/2}(\widehat{m(x)} - m(x)) \\ = \quad & \left(f(x)^{-1}(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)} - m(x))\right) f(x)(\widehat{f(x)})^{-1} \\ = \quad & \left(f(x)^{-1}(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)} - m(x))\right) \sum_{n=0}^{\infty}\left(-\frac{\widehat{f(x)} - f(x)}{f(x)}\right)^n \\ = \quad & f(x)^{-1}(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)} - m(x)) + o_p((nh^k)^{1/2}, \end{aligned}$$

where $f(x)(\widehat{f(x)})^{-1}$ is expanded as $\sum_{n=0}^{\infty}\left(-\frac{\widehat{f(x)}-f(x)}{f(x)}\right)^n$ and the last line uses the ordinary consistency of the kernel density estimator. It follows that the limit process for $(nh^k)^{1/2}(\widehat{m(x)} - m(x))$ coincides with the limit process for

$f(x)^{-1}(nh^k)^{1/2}\widehat{f(x)}(\widehat{m(x)}-m(x))$ and that is $N(0,\sigma^2(x)\,(f(x))^{-1}\int K(w)^2dw)$. If a sequence of ordinary functions converges in an ordinary sense to a limit distrbution and as a generalized sequence converges to an ordinary limit, then convergence to this limit holds in an ordinary sense, thus ordinary convergence in(16) obtains.

■

7 Appendix B.

In this Appendix useful results from different sources on generalized functions and generalized random processes are collected and presented in the form and notation that suits this paper.

7.1 Generalized functions: Definitions, properties and examples.

Here we summarize some of the definitions and results from Gel'fand and Shilov (1964, v.1 and v.2) and Ch. 1 of Sobolev (1992).

Spaces of test functions. The space K (v.1,1.2). Consider all infinitely differentiable real functions on R^k with finite support; this is a linear space. Convergence is defined for a sequence $\psi_1,...\psi_n,...$ if all ψ_n are zero outside a bounded interval and on it converge (uniformly) as well as each of the derivatives. K is a non-metrizable topological space.

The space $K(a)\subset K$ consists of $\psi\in K$ such that $\psi(x)=0$ for $\|x\|>a$.

The space S (v.1,1.10) is defined as that of all infinitely differentiable real functions on R^k that go to zero at infinity faster than any power; this is a linear space and topology can be defined similarly.

Spaces D_m, $m=0,1,...$ consist of functions with finite support that have m continuous derivatives. On R^k the space D_m contains all ψ with continuous derivatives $\frac{\partial^l\psi}{\partial^{l_1}x_1...\partial^{l_k}x_k}$ with $l_1+...+l_k\le m$. Any function in D can be approximated by a product of univariate functions.

Properties (from embedding diagrams, Sobolev, p. 56; notation: our D_m is $\overset{o}{C}{}^{(m)}$ there):

(i) $K\subset D_m$ and $D_m\subset D_{m'}$ for any m, where $m'<m$; also $K\subset S$;

(ii) each of the subspaces is dense in the larger space in the topology of that space.

We denote a generic space of test functions by D.

Generalized functions.

Ordinary functions (locally summable). A real function defined on R^k and Lebesgue-integrable on any bounded set (locally summable) is an ordinary function (v.1, 1.3). Each ordinary function f defines a functional on D: for $\psi \in D$ the value of the functional (f, ψ) can be defined by

$$(f, \psi) = \int_{-\infty}^{+\infty} f(x)\psi(x)dx. \tag{24}$$

It is easy to see that this is a linear continuous functional. Note that if two ordinary functions differ they define different functionals (v.2, 1.5).

Generalized functions. Denote the space of linear continuous functionals on D by D'. For any D the linear continuous functionals form a linear space D' with the weak topology: a sequence of functionals in D', g_n, converges to g if for any $\psi \in D$ $(g_n, \psi) \to (g, \psi)$.

Define a generalized function as a functional from the space D'. If it is given by (24) it is a regular functional (function); if it cannot be represented in the form (24) it is a singular functional (function) (v.1, 1.3). Usually the same notation (24) is used for any functional even though for singular functionals it does not have the ordinary interpretation. An example of a singular function is the $\delta-$function defined by $(\delta, \psi) = \psi(0)$. If a generalized function can be represented as

$$(f, \psi) = \sum_{k=0}^{p} \int_{-\infty}^{\infty} f_k(x)\psi^{(k)}(x)dx \tag{25}$$

where $f_0, ...f_k$ are ordinary functions, it is said that f has an order of singularity $\leq p$. E.g. an ordinary function has order of singularity zero, the δ-function has order of singularity ≤ 1, and since it is not an ordinary function (with order of singularity zero), its order of singularity is exactly 1.

Support of a generalized function f is defined (v.1,1.4) as the set of its essential points, where an essential point x is such that for any neighborhood $U(x)$ there is a test function $\psi \in D$ with support contained in $U(x)$ for which $(f, \psi) \neq 0$.

Convergence of generalized functions is defined as weak convergence of functionals: $f_n \Rightarrow f$ iff for any $\psi \in D$ the sequence of values of the functionals converges: $(f_n, \psi) \to (f, \psi)$. Generalized functions form a complete linear space D' (v.1, 1.8). The subspace of functionals given by (24) is dense in D' (v.1, 1.5 for the space K).

Properties (Sobolev, 1992, p.59; notation: our D'_m is $\overset{o}{C}{}^{(m)\#}$):
(i) $D'_m \subset K'$; $D'_{m'} \subset D'_m$ for any m where $m' < m$; also $S' \subset K'$;
(ii) for any of the spaces $D \subset D'$ and is dense there in the weak topology.

Derivatives of generalized functions. For any generalized function f on R its derivative, f', is defined as long as $D \subseteq D_1$ (v.1, 2.1)

$$(f', \psi) = (f, -\psi') \tag{26}$$

It is easy to check that this definition provides an ordinary derivative if the functional was differentiable as an ordinary function. Differentiation is a continuous operation (v.1, 2.4).

Any continuous functional on D of degree of singularity less than m can be extended to a continuous functional on D_m.

One can similarly consider multivariate generalized functions (examples in v.1, 2.3). For $x = (x_1, ..., x_k) \in R^k$ and $D(R^k)$ a continuous linear functional $F \in D'$ and its value on $\psi \in D(R^k)$, (F, ψ), is similarly defined. If $F(x_1, ..., x_k)$ is an ordinary locally summable function then

$$(F, \psi) = \int .. \int F(x_1, ..., x_k)\psi(x_1, ..., x_k)dx_1...dx_k,$$

which we write as $\int F(x)\psi(x)dx$. For any $\in D(R^k)$ the derivative

$$(\frac{\partial^k F(x)}{\partial x_1...\partial x_k}, K) = (-1)^k(F, \frac{\partial^k \psi(x)}{\partial x_1...\partial x_k})$$

can be defined as long as $D \subseteq D_k$.

7.2 Generalized random processes: Definitions, properties and examples.

Here we summarize some definitions and results from Ge'fand and Vilenkin (1964, v.4).

If f is a continuous linear functional on the space D (Gelfand and Vilenkin consider only D coinciding with the space K) and additionally (f, ψ) is a random variable for any $\psi \in D$ which implies that for any number l of $\psi_1, ...\psi_l \in D$ the set $(f, \psi_1), ..., (f, \psi_l)$ has a joint probability distribution, then f defines a generalized random function on D. (pp. 241-243; the notation is different there from ours). Gel'fand and Vilenkin distinguish between a generalized random process defined by f when the functions in D are univariate and a generalized random field for the case of multivariate functions. We shall not make that distinction and refer to a generalized random process (univariate or multivariate) in all cases.

An expectation or mean functional is defined by (again changing notation, p.246)

$$m(\psi) = (E(f), \psi) = E(f, \psi)$$

if $E(f, \psi)$ defines a continuous linear functional on D.

If for any ψ_1, ψ_2 the expectation $E((f, \psi_1), (f, \psi_2))$ exists then the "correlation" functional of the process is given by

$$B(\psi_1, \psi_2) = E((f, \psi_1), (f, \psi_2)),$$

and the covariance functional by

$$C(\psi_1, \psi_2) = B(\psi_1, \psi_2) - m(\psi_1)m(\psi_2)$$

provided these functionals exist. These functionals are bilinear in ψ_1, ψ_2. (p. 247). Higher order moments are similarly defined.

A generalized random process is a generalized Gaussian process if for any linearly independent $\psi_1, ..., \psi_l$ from D the joint distribution of $(f, \psi_1), ..., (f, \psi_l)$ is Gaussian. (p.248). If for any $\psi \in D$ (not identically zero) the covariance functional $(C, (\psi, \psi)) > 0$ the limit process is a proper (non-degenerate) generalized Gaussian process[5]. A generalized Gaussian process is uniquely determined by its mean functional and "correlation" (or covariance) bilinear functional (Theorem 1, p.250-251).

Generalized Gaussian processes are differentiable; for example, the derivative of a generalized Gaussian process with mean functional zero and correlation functional $B(\psi_1, \psi_2)$ is a generalized Gaussian process with correlation functional B' given by $B'(\psi_1, \psi_2) \equiv B(\psi_1', \psi_2')$ (p. 257).

References

[1] Ahmad, I.A. and P.B. Cerrito (1994) Nonparametric Estimation of Joint Discrete-continuous Probability Densities with Applications, Journal of Statistical Planning and Inference 41, 349-364.

[2] Bierens, H.J. (1987) Kernel Estimation of Regression Functions, in T.F.Bewley (ed.) Advances in Econometrics , Cambridge University Press, v.1, 99-144.

[3] Devroye, L.P. and L. Györfi (1985) Nonparametric Density Estimation, Wiley, New York.

[4] Fan, J and Z. Gijbels (1995) Local Polynomial Modelling and its Applications, Chapman and Hall, London.

[5] Frigyesi, A. and O. Hössjer (1998) A Test for Singularity, Statistics and Probability Letters, v.40, 215-226.

[6] Frigyesi, A. (2004) Topics in Multifractal Measures, Nonparametrics and Biostatistics, Centre for Mathematical Sciences, Lund University, Doctoral thesis.

[7] Gel'fand, I.M. and G.E.Shilov (1964) Generalized Functions, Vol.1, Properties and Operations, Academic Press, San Diego.

[8] Gel'fand, I.M. and G.E.Shilov (1964) Generalized Functions, Vol.2, Spaces of Test functions and Generalized Functions, Academic Press, San Diego.

[5] Gel'fand and Vilenkin discuss the results for proper Gaussian processes; some of those generalize easily to degenerate processes.

[9] Gel'fand, I.M. and N.Ya.Vilenkin (1964) Generalized Functions, Vol.4, Applications of Harmonic Analysis, Academic Press, San Diego.

[10] Green, D.A. and W.C. Riddell (1997), Qualifying for unemployment insurance: an Empirical Analysis, Economic Journal, v. 107, 67-84.

[11] Halperin, I. (1952) Introduction to the Theory of Distributions (based on Lectures by L. Schwartz), University of Toronto Press.

[12] Härdle, W. , G. Kerkyacharian, D. Picard and A. Tsybakov (1998) Wavelets, Approximations and Statistical Applications, Springer-Verlag.

[13] Li, Q. and J. Racine (2003) Nonparametric Estimation of Distributions with Categorical and Continuous Data, Journal of Multivariate Analysis, v. 86, 266-292.

[14] Lu, Z-Q. (1999) Nonparametric Regression with Singular Design, Journal of Multivariate Analysis, v. 70, pp. 177-201.

[15] Mathematical Encyclopedia - Matematicheskaya Encyclopedia (Russian), ed. I.M.Vinogradov, Moscow, 1977.

[16] Müller, H-G. (1992) Change-points in Nonparametric Regression Analysis, Annals of Statistics, v. 20, 737-761.

[17] Pagan, A. and A. Ullah (1999) Nonparametric Econometrics, Cambridge University Press.

[18] Phillips, P.C.B. (1991) A Shortcut to LAD Estimator Asymptotics, Econometric Theory, 7,450-463

[19] Phillips, P.C.B. (1995) Robust Non-Stationary Regression, Econometric Theory, 11, 912-951.

[20] Schwartz, L. (1950) Théorie des Distributions, v.1,2.. Hermann, Paris.

[21] Schennach, S. (2004) Estimation of Nonlinear Models with Measurement Error, Econometrica, v.72, pp. 33-75.

[22] Sobolev, S. (1992) Cubature formulas and Modern Analysis, Gordon and Breach Science Publishers, S.A..

[23] Zinde-Walsh, V. (2002) Asymptotic Theory for some High Breakdown Point Estimators. *Econometric Theory* 18, 1172-1196.

[24] Zinde-Walsh, V. and P. C. B. Phillips (2003) Fractional Brownian motion as a differentiable generalized Gaussian process, in K. Athreya, M. Majumdar, M. Puri and W. Waymire (eds.) Probability, Statistics and their Applications: Papers in Honor of Rabi Bhattacharya. Institute of Mathematical Statistics Lecture Notes-Monograph Series, Beachwood, Ohio, v. 41, 285-292. .