Neural Network Interest Group

Título/*Title*:
Nonparametric Density and Entropy Estimation (with a Focus on the Parzen Window Method)

Autor(es)/*Author(s)*:
J. P. Marques de Sá

Relatório Técnico/*Technical Report* No. 3 /2006

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:

Nonparametric Density and Entropy Estimation
(with a Focus on the Parzen Window Method)

Autor(es)/*Author(s)*:

J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 3 /2006

Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Julho 2006

Contents

1 Nonparametric Density Estimators

1.1 Estimability of Functionals

The first question to be addressed is whether or not a given functional $q(F)$, where F belongs to a family of distributions $\mathcal{F}$, is estimable based on a sequence of i.i.d. random variables $X_1,\ldots, X_n$.
Reference [15] defines estimability in the following way: $q(F)$ is *estimable with n observations* if there exists a statistic $\delta(X_1,\ldots, X_n)$ such that

$$E_F\left[\delta\left(X_1,...,X_n\right)\right]=q(F)$$

Therefore, estimability means the existence of unbiased estimators.
Reference [15] explains the necessary and sufficient conditions of estimability for a convex family[1] of distribution functions and presents examples of estimable and non-estimable functionals. Here are some of them:

Examples of estimable functionals:

- The variance: $q(F)=\sigma^2(F)$.
- $q(F) = F(x_0)$ for some fixed $x_0 \in \Re$.
- $q(F)=\int_{\Re}\exp(it_0x)F(dx)$

Examples of non-estimable functionals:

- $q(F) = f(x_0)$ for some fixed $x_0 \in \Re$.
- The regression function of Y on X: $q(F)=\int_R yf(x,y)dy/\int_R f(x,y)dy$
- The conditional density of Y given x: $q(F)=f(x,y)/\int_R f(x,y)dy$

Although unbiased estimators do not exist in general for f, it is possible to define sequences of density estimators, $\hat{f}_n$, *asymptotically unbiased*:

$$\lim_{n\to\infty} E_F[\hat{f}_n(x)]=f(x)$$

1.2 Histogram-Based Density Estimator

We are given a random sample $\{x_1,\ldots, x_k, \ldots, x_n\}$observations of i.i.d. r.v. from an unknown absolutely continuous pdf.

We restrict ourselves to the univariate case.

[1] $\mathcal{F}$ is a convex family if for every $F,G\in\mathcal{F}$ and $0 \le \alpha \le 1$, $\alpha F + (1-\alpha)G\in\mathcal{F}$.

If the unknown pdf, $g(x)$, has an infinite support we content ourselves with estimating the truncated density

$$f(x) = \begin{cases} g(x) / \int_a^b g(t)dt & x \in [a,b] \\ 0 & otherwise \end{cases}$$

Let us partition the interval by $a = t_0 < t_1 < \ldots < t_i < \ldots < t_m = b$. (We use "$t_i$" for no confusion with the x_k.)

Let us denote:

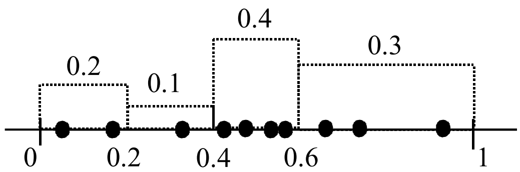
$T_i = [t_i, t_{i+1}[$;

$q_i = \sum_{k=1}^{n} I_{x_k \in T_i}$, $\quad t \in T_i$ (# cases falling in T_i);

$l(T_i) = t_{i+1} - t_i$.

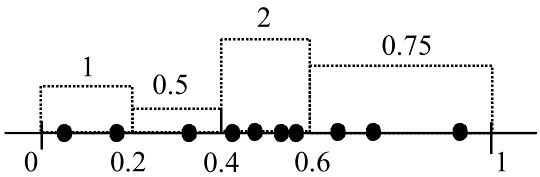
<u>Histogram</u>:

$$p(t) = \begin{cases} q_i / n & t \in T_i \\ q_{m-1} / n & t = b \\ 0 & t \notin [a,b] \end{cases}$$



<u>Histogram-based density estimator</u>:

$$\hat{f}_H(t) = \begin{cases} p(t) / l(T_i) & t \in T_i \\ p(t) / l(T_{m-1}) & t = b \\ 0 & t \notin [a,b] \end{cases}$$



1×0.2 + 0.5×0.2 + 2×0.2 + 0.75×0.4 =1

Rationale: The variable q_i is a multinomial r.v. Thus, q_i / n estimates $\int_{T_i} f(t)dt$. If f is absolutely continuous and T_i is small, then $f(t) \approx f(t_i)$ for $t \in T_i$. Hence, $q_i / (n \times l(T_i))$ estimates $f(t)$.

Properties (for details, see [9]):

- Let us assume an estimator based on assigning quantities c_i to the T_i intervals. Among all such estimators $\hat{f}_H$ uniquely maximizes the likelihood $L(c_0,\ldots, c_{m-1})$.

- <u>Theorem</u>: Suppose that f is bounded and has continuous derivatives up to order three except at the endpoints of $[a,b]$. Suppose equal spacing, $t_{i+1} - t_i = 2h(n) \equiv 2h_n$. Then, if $n \to \infty$ and $h_n \to 0$ such that $nh_n \to \infty$, for $x \in [a,b]$

$$MSE\left(\hat{f}_H(x)\right) = E\left[\left(\hat{f}_H(x) - f(x)\right)^2\right] \to 0$$

i.e., $\hat{f}_H$ is a consistent estimator for $f(x)$.

- The proof of the above Theorem leads to the results[2]

$$MSE\left(\hat{f}_H(x')\right) = \frac{f(x')}{2nh_n} + \frac{h_n^4}{36}\left|f''(x')\right|^2 + O(1/n) + O(h_n^5)$$

and

$$MSE\left(\hat{f}_H(x)\right) \leq \frac{f(x')}{nh_n} + 2\left|f'(x')\right|^2 h_n^2 + O(1/n) + O(h_n^3),$$

is based on a Taylor series development around the midpoint x' of the interval containing x and uses the well-known result[3]:

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

- From the formula of $MSE\left(\hat{f}_H(x)\right)$ one may select $h_n = \left[\frac{f(x')}{4\left(f'(x')\right)^2}\right]^{1/3} n^{-1/3}$ to obtain convergence throughout the kth interval of order $n^{-2/3}$.

- The integrated mean square error is minimized by selecting

$$h_n = \left[\frac{1}{4\int\left(f'(x)\right)^2 dx}\right]^{1/3} n^{-1/3}$$

to obtain

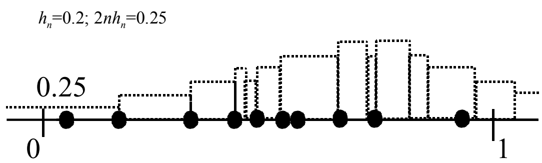
$$\int MSE\left(\hat{f}_H(x)\right) = IMSE \leq 3\left[\frac{1}{2}\int\left(f'(x)\right)^2 dx\right]^{1/3} n^{-2/3} + O\left(\frac{1}{n} + h_n^3\right)$$

1.3 Rosenblatt's Kernel Estimator

Rosenblatt's estimator (introduced in 1956) is an extension of the histogram-based estimator:

$$\hat{f}_n(x) = \frac{\#\text{ sample points in } \left]x - h_n, x + h_n\right]}{2nh_n},$$

i.e., we shift the interval such as to center it at x.



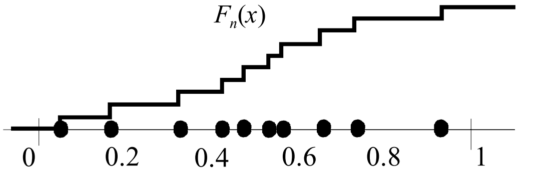
[2] Note that $\hat{f}_H(x)$ is a r.v. (dependent on $\{X_1, \ldots, X_k, \ldots, X_n\}$); $f(x)$ is a constant.

[3] Therefore a convergence in the MSE sense is equivalent to a convergence of the mean ($E[\hat{f}_H] \underset{n\to\infty}{\rightarrow} f$) together with a convergence of the variance towards zero ($Var[\hat{f}_H] \underset{n\to\infty}{\rightarrow} 0$).

The estimate can also be written as:

$$\hat{f}_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n}$$

where $F_n(x)$ is the <u>empirical distribution</u> (also called <u>empirical measure</u> in the previous tutorial).

$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \in A}\ , A \subset \Re$ For $A =]-\infty, x]$, $\mu_n(A) = F_n(x)$	

The <u>shifted histogram estimator</u> of Rosenblatt can be represented as:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} w\left(\frac{x - x_i}{h_n}\right)$$

where $w(u) = \begin{cases} 1/2 & |u| < 1 \\ 0 & otherwise \end{cases}$ is the kernel (rectangular).

Properties (for details, see [9]):

- In the same conditions as above:

$$MSE\left(\hat{f}_H(x)\right) = \frac{f(x)}{2nh_n} + \frac{h_n^4}{36}\left|f''(x)\right|^2 + o\left(\frac{1}{nh_n} + h_n^4\right)$$

- One may minimize the first two terms in the above formula, selecting $h_n = \left[\frac{9f(x)}{2\left(f''(x)\right)^2}\right]^{1/5} n^{-1/5}$ to obtain an MSE of order $n^{-4/5}$. Therefore the MSE of Rosenblatt's estimator decreases faster than the fixed grid histogram estimator (order of $n^{-2/3}$).

- The integrated mean square error is minimized by selecting $h_n = \left[\frac{9}{2\int\left(f''(x)\right)^2 dx}\right]^{1/5} n^{-1/5}$, yielding $IMSE \sim n^{-4/5}$.

1.4 Parzen Window Estimator

The Parzen window estimator is a generalization of the shifted-histogram estimator, introduced by Parzen in 1962 [1]:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x - x_i}{h_n}\right),$$

where $K(x)$, the kernel function, is any Borel function[4] satisfying:

i. Boundedness: $\sup_{\Re}|K| < \infty$

ii. $K \in L_1$: $\int |K| < \infty$

iii. Decreasing faster than $1/x$: $\lim_{x\to\infty}|xK(x)| = 0$

iv. $\int K = 1$.

The Parzen window estimator can also be written as a convolution of the window with the (derivative of the) empirical distribution:

$$\hat{f}_n(x) = \int \frac{1}{h_n}K\left(\frac{x - y}{h_n}\right)dF_n(y) = \int K_{h_n}(x - y)dF_n(y),$$

where $K_{h_n}(x) = \frac{1}{h_n}K\left(\frac{x}{h_n}\right)$. The positive constants h_n are the bandwidths. Note that $\int \left|K_{h_n}\right| = \int |K|$.

Convolutions enjoy a series of properties given in Appendix. Particularly note the smoothing imposed by convolutions with a large class of kernels (Fourier Transform property). For a large class of kernels $\hat{f}_n(x)$ is a blurred, smoothed, version of $f(x)$.

In the following we often use, for simplicity reasons, the notation h, K_h and f_n instead of h_n, K_{h_n} and $\hat{f}_n$, respectively.

A central role in the consistency of this estimator is played by the following:

Lemma (Bochner, 1960): Let K be a Borel function satisfying i, ii and iii. Let $g \in L_1$ and

$$g_n(x) = \int K_h(x - y)g(y)dy = K_h \otimes g$$

If h_n is a sequence of positive constants having $\lim_{n\to\infty} h = 0$ the following holds (at every continuity point of g):

[4] A Borel function is a measurable function. A continuous function is a Borel function.

$$\lim_{n\to\infty} g_n(x) = g(x)\int K(y)dy$$

•

In [2] (Devroye, 2001) this Lemma is stated as an equivalent Theorem, stating:

$$\lim_{h\to 0}\int\left|g\otimes K_h - g\int K\right| = 0$$

Sometimes the Parzen window estimator is written as

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$$

to stress the fact that $f_n(x)$ is a r.v.
The r derivative of $f(x)$ is estimated by [6]

$$f_n^{(r)}(x) = \frac{1}{nh^{r+1}}\sum_{i=1}^{n} K^{(r)}\left(\frac{x-X_i}{h}\right).$$

<u>Properties (for details see [1], [2], [9], [11], [14-16]):</u>

- If K is an even function we have:

$$\mu_n = \bar{x}\,; \qquad \sigma_n^2 = s^2 + h^2\int x^2 K(x)dx$$

  The proofs are in [9].
- The estimate is unbiased: $\lim_{n\to\infty} E[f_n(x)] = f(x)$. A direct corollary of the above Lemma. The proof is in Appendix.
- If in addition to $\lim_{n\to\infty} h = 0$ the bandwidths satisfy $nh_n \to \infty$ (they decrease less than $1/n$) the estimate verifies:

$$\lim_{n\to\infty} nh\mathrm{V}[f_n(x)] = f(x)\int K^2(y)dy$$

  For a Gaussian kernel: $\lim_{n\to\infty} \mathrm{V}[f_n(x)] = \dfrac{f(x)}{2nh\sqrt{\pi}}$
- From the two preceding results follows that the estimate is consistent:

$$MSE(f_n(x)) \to 0\,.$$

- The consistent estimate, for a density having r derivatives, verifies:

$$MSE(f_n(x)) \sim \frac{f(x)}{nh_n}\int_{-\infty}^{\infty} K^2(y)dy + h_n^{2r}k_r^2\left|f^{(r)}(x)\right|^2,$$

  where k_r is the *characteristic exponent* of the Fourier transform of $K(x)$, that we denote $k(u)$, defined as:

$$k_r = \lim_{u\to 0}\left[\frac{1-k(u)}{|u|^r}\right]$$

For the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad \Leftrightarrow \qquad k(u) = e^{-u^2/2}$$

$$k(u) = 1 + \sum_{i=1}^{\infty} \frac{(-u^2/2)^i}{i!} = 1 - \frac{u^2}{2} + O(u^4)$$

Thus: $k_r = ½$, for $r = 2$.

Any even kernel having $x^2K(x) \in L_1$ has a nonzero finite k_r for $r = 2$.

- The optimal MSE is given by:

$$MSE_{opt}\left(f_n(x)\right) \sim (2r+1)\left\{\frac{f(x)}{2nr}\int_{-\infty}^{\infty} K^2(y)dy\right\}^{2r/(2r+1)} \left|k_r f^{(r)}(x)\right|^{2r/(2r+1)}$$

Thus, the decrease of the MSE is of order $n^{-2r/(2r+1)}$. Therefore, for symmetric $x^2K(x) \in L_1$ kernels the decrease obtainable is of order $n^{-4/5}$ as good as for the shifted histogram.

- The optimal integrated mean square error of the consistent estimate in the above conditions is obtained for:

$$h_n = n^{-1/(2r+1)}\alpha(K)\beta(f)$$

with

$$\alpha(K) = \left[\frac{\int K^2(y)dy}{2r\left(\int y^r K(y)dy/r!\right)^2}\right]^{1/(2r+1)}$$

$$\beta(f) = \left[\int \left|f^{(r)}(y)\right|^2 dy\right]^{-1/(2r+1)}$$

For symmetric $x^2K(x) \in L_1$ kernels we have:

$$h_n = n^{-1/5}\alpha(K)\beta(f)$$

with

$$\alpha(K) = \left[\frac{\int K^2(y)dy}{\left(\int y^2 K(y)dy\right)^2}\right]^{1/5}$$

$$\beta(f) = \left[\int \left|f''(y)\right|^2 dy\right]^{-1/5}$$

Some values for $\alpha(K)$:

K		$\alpha(K)$
$K(y) = 1/2$	$\lvert y\rvert \le 1$	1.3510
$K(y) = 1 - \lvert y\rvert$	$\lvert y\rvert \le 1$	1.8882
$K(y) = \frac{15}{16}\left(1-y^2\right)^2$	$\lvert y\rvert \le 1$	2.0362
$K(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$	$\lvert y\rvert < \infty$	0.7764

The difficulty lies in the fact that $\beta(f)$ is generally unknown. One could consider iteratively improving an estimate of $\beta(f)$.

For the Gaussian density with standard deviation σ, we have:

$$\int \left| f''(y) \right|^2 dy \approx 0.212\sigma^{-5} \quad \Rightarrow \quad \beta(f) \approx 1.3637\sigma \quad \Rightarrow \quad h_n \approx 1.06\sigma n^{-1/5}$$

Some values of optimal h_{100} using a Gaussian kernel:

Density	$\beta(f)$	h_{100}
$N(0,1)$	1.3637	0.42
$.5N(-1.5,1)+.5N(1.5,1)$	1.6177	0.50
t_5	1.0029	0.31
$F_{10,10}$	0.4853	0.15

Quoting reference [9]: "kernel estimators are not in general robust against poor choices of h_n".

Reference [11] mentions the use of $h_n = 0.79Rn^{-1/5}$, where R is the interquartile range, for skew distributions.

- The optimal IMSE for symmetric $x^2K(x) \in L_1$ kernels is given by:

$$IMSE = \frac{5}{4} C(K) \beta^{-1}(f) n^{-4/5}$$

The quantity $C(K)$ is minimized for the Epanechnikov kernel:

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & -\sqrt{5} \le t \le \sqrt{5} \\ 0 & otherwise \end{cases}$$

Reference [11] indicates the efficiencies (IMSE or $C(k)$ ratios) of other kernels compared to K_e. The efficiency of the Gaussian kernel is ≈ 0.9512.

- The errors of the consistent estimate are asymptotically normal:

$$\lim_{n\to\infty} P\left\{\frac{f_n(x) - E[f_n(x)]}{\sigma[f_n(x)]} \le c\right\} = N_{0,1}(c)$$

- By the bounded difference inequality:

$$P\left(\left|\int |f_n - f| - E\left[\int |f_n - f|\right]\right| \ge t\right) \le 2e^{-t^2/2n\left(\int |K|\right)^2}$$

- Schuster's Lemma [6]: If f and its r+1 derivatives are bounded and if $\{\varepsilon_n\}$ is a sequence of positive numbers such that $h_n = o(\varepsilon_n)$, then there exist positive constants C_1 and C_2 such that

$$P\left\{\sup\left|f_n^{(r)} - f^{(r)}\right| > \varepsilon_n\right\} \le C_1 \exp(-C_2 n \varepsilon_n^2 h_n^{2r+2})$$

  for sufficiently large n.

- Suppose that one of the X_i changes value while the other n-1 data points remain fixed. Let f_n^* denote the new perturbed estimate. Then: $\int\left|f_n - f_n^*\right| \le \frac{2}{n}\int |K|$ (Parzen window estimates are stable). See proof in Appendix.

- The Parzen window estimator is a regularized estimate of the density ([3]).

2 Entropy Estimation Based on Parzen Windows

2.1 Plug-in Estimates

We only consider the Shannon functional: $H(f) = -\int f(x) \ln f(x) dx$.

Plug-in estimates [5] are based on using a density estimate f_n obtained from the data. There are four types of plug-in estimators:

- Integral estimator
- Resubstitution estimator
- Splitting data estimator
- Cross-validation estimator

We'll only consider the first two:

1. <u>Integral estimator</u>: $H_n(f) = -\int_{A_n} f_n(x) \ln f_n(x) dx$

This estimator requires numerical integration. A_n typically excludes tail values of the distribution.

<u>Theorem</u> (strong consistency; Dmitriev and Tarasenko, 1973: [6]): Assume that a function M exists such that

$$\sup_{|y| \le x} \frac{1}{f(y)} \le M(x) \quad \forall x$$

If $h(n) = n^{-1/4}$ and $A_n = [-k_n, k_n]$ with $k_n = M^{-1}(n^{1/10})$, then $H_n(f)$ converges to $H(f)$ a.s.

2. <u>Resubstitution estimator</u>: $H_n(f) = -\frac{1}{n}\sum_{i=1}^{n} \ln f_n(X_i)$

(This estimator seems to have been first proposed by I.A. Ahmad and P-E Lin in 1976; [8].)

<u>Properties for discrete distributions</u> [7]:

- The resubstitution estimate is strongly universally consistent (also consistent in L_2).
- $\mathrm{E}[H_n] \le H$; $\mathrm{V}[H_n] \le \ln^2 n / n$
- $P\{|H_n - \mathrm{E}[H_n]| > \varepsilon\} \le 2e^{-n\varepsilon^2 / 2\ln^2 n}$
- There is no universal convergence rate of $|H_n - H|$. In other words, the convergence of H_n to H can be arbitrarily slow.

<u>For continuous distributions</u> [8], [12]:

- L_1 consistency [8]: If $nh_n \to \infty$ as $n \to \infty$, $\int [\ln f]^2 f < \infty$, f' is continuous and $\sup|f'| < \infty$, $\int |u| K(u) du < \infty$ then $\mathrm{E}[|H_n - H|] \underset{n\to\infty}{\to} 0$.

- L_2 consistency [8]: If, in addition, $\int (f'(x)/f(x))^2 f(x) < \infty$ (finite Fisher information number) then $\mathrm{E}[|H_n - H|^2] \underset{n\to\infty}{\to} 0$

Refernce [8] states that the above conditions are mild and are satisfied by the following distributions: Gamma distribution with $\alpha = 1$ and $\beta = 0$ or $\alpha > 2$ and $\beta > 0$; Weibull distribution with parameters $\alpha > 0$ and $\beta > 2$; normal distribution.

- Almost sure consistency: $H_n \underset{n\to\infty}{\to} H$ a.s., under certain mild conditions stated in [13] (the multivariate case is studied).
- Reference [13] presents upper bounds for the moments of $|H_n - H|$. The formulas are complex, dependent on the support limits (stated above as [a, b]; denoted [-K_n, K_n] in [13], since they vary with n) and applicable only when $\varphi(u) = \inf\{f(x);\, |x| \le u\} > 0$ (analogous for the multivariate case). Therefore, their formulas do not apply to densities with restricted support.
  Here is the formula for the first order moment (univariate case):

$$E[H_n - H] \le \left(\varepsilon_n + \frac{K(0)}{nh_n}\right)\varphi^{-1}(K_n) + c_1(1)\exp(-\frac{1}{2}c_2 n\varepsilon_n^2 h_n^2) + c_2(1)n^{-1/2}|\ln\varphi(K_n)| + c_3(1)\ln K_n / K_n^{\rho_1 - 1}$$

3 Appendix

3.1 Asymptotic Notation

- $f(x) = O(g(x))$ if there are constants c, x_0 > 0 such that $|f(x)| \le c|g(x)|,\quad \forall x \ge x_0$. In other words, an $O(g(x))$ term (an asymptotic upper bound; "order of $g(x)$") deviates in absolute value less than $c|g(x)|$ after a given x_0.
- $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) = 0$. In other words, an $o(g(x))$ term converges to zero faster than $g(x)$.
- $f(x) = o(g(x))$ implies $f(x) = O(g(x))$, but not vice-versa.
- $f(x) \sim g(x)$ if $\lim_{x\to\infty} f(x)/g(x) = 1$ ("f goes asymptotically to g").
- $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $g(x) = O(f(x))$ ("asymptotic tight bound").

Examples:

- $\sin x = O(1)$
- $x\sin x = O(x)$. However, $x\sin x \ne o(x)$.
- $e^x = 1 + x + \frac{x^2}{2} + O(x^3)$
- $\int kx^3 dx = O(x^4)$
- $x = o(x^2)$. Also $x = O(x^2)$.
- In the calculation of $MSE\left(\hat{f}_H(x')\right)$ for the histogram-based density estimator, the expression of the variance is:

$$Var(\hat{f}_H(x')) = \frac{1}{2nh_n}\left[f(x') - 2h_n f^2(x') + \frac{h_n^2}{6} f''(x') + O(h_n^3)\right]$$

Hence:

$$Var(\hat{f}_H(x')) = \frac{f(x')}{2nh_n} + \frac{1}{2n}\left[-2f^2(x') + \frac{h_n}{6} f''(x') + O(h_n^2)\right],$$

since $O(h_n^3)/h_n = O(h_n^2)$. Moreover, since $h_n \to 0$, we have:

$$\frac{1}{2n}\left[-2f^2(x') + \frac{h_n}{6} f''(x') + O(h_n^2)\right] = O\left(\frac{1}{n}\right)$$

Finally:

$$Var(\hat{f}_H(x')) = \frac{f(x')}{2nh_n} + O\left(\frac{1}{n}\right)$$

Link: http://en.wikipedia.org/wiki/Big_O_notation

3.2 Convolution properties

- $f \otimes g = g \otimes f$
- $f \otimes (g \otimes h) = (f \otimes g) \otimes h$
- $(f + g) \otimes K = f \otimes K + g \otimes K$
- $(af) \otimes K = a(f \otimes K), \quad a \in \Re$
- $\frac{d}{dx}(f \otimes g) = \frac{df}{dx} \otimes g = f \otimes \frac{dg}{dx}$
- $F(g \otimes K) = F(g) \times F(K), \quad F \equiv$ Fourier transform
- $\int |f \otimes K| \le \int |f| \times \int |K|$ (Young's inequality)
- $\int |f \otimes K - g \otimes K| \le \int |K| \int |f - g|$ (convolution lowers total variation; the proof is based on Young's inequality)

3.3 Proof of the convergence of $E[f_n]$ to f

We have:

$$E[f_n(x)] = E[K_h(x - X)] = \int_{-\infty}^{\infty} K_h(x - y) f(y) dy$$

Applying Bochner's Lemma (if the respective conditions are satisfied):

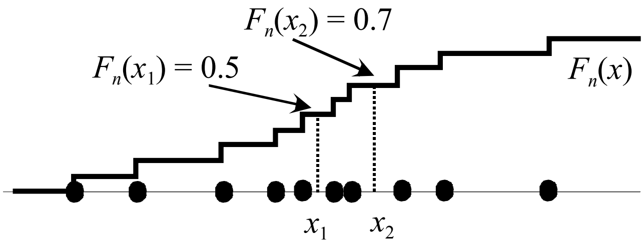
$$\lim_{n \to \infty} E[f_n(x)] = f(x) \int_{-\infty}^{\infty} K(y) dy = f(x)$$

The original justification ([10]) for this result ran like this: Consider:

$$\hat{f}_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n} \text{ where } F_n(x) = \frac{\#\text{ sample points} \le x}{n}$$

Partition the real line into three intervals: $]-\infty, x_1]$, $]\, x_1, x_2]$, $]x_2, +\infty[$. Denote:

$$Y_1 = F_n(x_1);\quad Y_2 = F_n(x_2) - F_n(x_1);\quad Y_3 = 1 - F_n(x_2)$$



$y_1 = 0.5$
$y_2 = 0.2$
$y_3 = 0.3$

Then, (nY_1, nY_2, nY_3) is a trinomial r.v. with probabilities $(F(x_1), F(x_2) - F(x_1), 1 - F(x_2))$. Thus, we have $E[F_n(x)] = F(x)$.

3.4 Proof of the Result on the Perturbed Density Estimate

Assume w.l.o.g. that is the value of X_1 that changes. We have:

$$\left|f_n - f_n^*\right| = \frac{1}{n}\left|\left(K_h(x-x_1) - K_h\left(x-x_1'\right)\right)\right| \le \frac{1}{n}\left(K_h(x-x_1) + K_h\left(x-x_1'\right)\right)$$

Therefore:

$$\int\left|f_n - f_n^*\right| \le \frac{1}{n}\int\left(K_h(x-x_1) + K_h\left(x-x_1'\right)\right)dx = \frac{2}{n}\int|K|$$

If $\int|K| = 1$: $\int\left|f_n - f_n^*\right| \le \frac{2}{n}$

References

[1] - Emanuel Parzen (1962) On Estimation of a Probability Density Function and Mode. Annals Math. Stat., 33:1065-1076.

[2] - Luc Devroye, Gábor Lugosi (2001). Combinatorial Methods in Density Estimation. Springer-Verlag

[3] - Vladimir Vapnik (1998) Statistical Learning Theory. John Wiley & Sons, Inc.

[4] - A. Kolmogorov, S. Fomin (1999) Elements of the Theory of Functions and Functional Analysis. Dover Pub. Inc.

[5] - J. Beirlant, EJ Dudewicz, L Györfi, EC van der Meulen (1997) Nonparametric Entropy Estimation: An Overview. Int. J. Math. Stat. Sci., 6(1):17-39.
[6] - Yu. G. Dmitriev, F.P. Tarasenko (1973) On the Estimation of Functionals of the Probability Density and its Derivatives. Theory of Probability and its Applications, 18:628-633.
[7] - András Antos, Iannis Kontoyiannis (2001) Convergence Properties of Functional Estimates for Discrete Distributions. Random Structures & Algorithms, 19: 163-193.
[8] - I.A. Ahmad, P-E Lin (1976) A Nonparametric Estimation of the Entropy for Absolutely Continuous Distributions. IEEE Tr. IT, pp. 372-375.
[9] - Richard A. Tapia, James R. Thompson (1978) Nonparametric Probability Density Estimation. The John Hopkins University Press.
[10] – Murray Rosenblatt (1956) Remarks on some nonparametric estimates of a density function. Annals of Math. Statistics, 27:832-835. (Not available.)
[11] – B.W. Silveman (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall Ltd.
[12] – Abdelkader Mokkadem (1989) Estimation of the Entropy and Information of Absolutely Continuous Random Variables. IEEE Tr. IT, 35 (1): 193-196.
[13] – A.V. Ivanov, M.N. Rozhkova (1982) Properties of the Statistical Estimate of the Entropy of a Random Vector with a Probability Density. Problems Inform. Transmission, 17: 171-178.
[14] – R.O. Duda, P.E. Hart, D.G. Stork (2001) Pattern Classification. John Wiley & Sons, Inc.
[15] – B.L.S. Prakasa Rao (1983) Nonparametric Functional Estimation. Academic Press Inc.
[16] – M. P. Wand, M. C. Jones (1995) Kernel Smoothing. Chapman & Hall.