ON APPROXIMATION OF ERROR MEASURE IN KERNEL DENSITY ESTIMATION UNDER DEPENDENCE

E. Krasnogir
Belarusian State University, Minsk, BELARUS
e-mail: krasno@tut.by

Abstract

The problem of the probability density estimation by using n size sample of stationary process is considered. We investigate conditions which the coefficients in approximation of Mean Integrated Squared Error should satisfy to receive the consistent estimator. The clear approximation formula of optimal bandwidth for dependent data is given.

1 Introduction

Let X be a stationary process with marginal density $f(x)$ and let $\{X_i\}_{i=1}^n$ be a sample of size n from this process at the discrete times $t,\ t+\Delta,\ \ldots,\ t+(n-2)\,\Delta,\ t+(n-1)\,\Delta$, where Δ is the time between observations. The nonparametric estimator of $f(x)$ is defined by $f_h(x) = \frac{1}{nh}\sum\limits_{i=1}^{n} K(\frac{x-X_i}{h})$ [1], where h is a smoothing parameter or bandwidth and $K(x)$ is a kernel function which is a symmetric probability density [1]. In nonparametric kernel estimation it is supposed [1] that

$$h \to 0, hn \to \infty \text{ as } n \to \infty. \tag{1}$$

The statistical properties of $f_h(x)$ depend closely on the bandwidth h [1]. To evaluate the optimal h it is necessary to choose a measure of distance between the true density $f(x)$ and the estimator $f_h(x)$. Especially common choice is the Mean Integrated Squared Error (MISE) [2]

$$MISE(h) \equiv \int\limits_{-\infty}^{+\infty} E\left[\left(f_h(x) - f(x)\right)^2\right] dx. \tag{2}$$

2 Approximation and Order of Coefficients

Since it is impossible to find optimal bandwidth in explicit form from expression (2), we determine it approximately. Using of Taylor's expansion for (2) in a neighborhood of point $h = 0$, choosing from this representation only those terms which depend on h (up to the order h^4) gives following approximation for $MISE(h)$:

$$g(h) = \frac{\nu_2}{hn} - \alpha h^2 + \beta h^4. \tag{3}$$

Here it is denoted

$$\alpha = \mu_2 \left(\int\limits_{-\infty}^{\infty} f(x) f^{(2)}(x) dx - \frac{2}{n^2} \sum_{j=1}^{n-1} (n-j) \int\limits_{-\infty}^{\infty} {}_j f_{0,\,2}^{(2)}(x,\, x)\, dx \right),$$

$$\beta = \frac{1}{6n^2} \sum_{j=1}^{n-1} (n-j) \int\limits_{-\infty}^{\infty} \left({}_j f_{0,\,4}^{(4)}(x,\, x)\, \mu_4 + 3\, {}_j f_{2,\,2}^{(4)}(x,\, x)\, \mu_2^2 \right) dx - \frac{\mu_4}{12} \int\limits_{-\infty}^{\infty} f(x) f^{(4)}(x) dx,$$

$$\nu_2 = \int\limits_{-\infty}^{\infty} K^2(u) du, \quad \mu_r = \int\limits_{-\infty}^{\infty} u^r K(u) du, \quad {}_j f_{k,\,m-k}^{(m)}(x,\, x) = \left. \frac{\partial^m f_j(t,\, s)}{\partial t^k \partial s^{m-k}} \right|_{t=x,\, s=x},$$

where $f_j(t,\, s)$ is a joint density of $(X_k,\, X_{k+j})$, $k = \overline{1,\, n-j}, j = \overline{1,\, n-k}$.

The necessary condition of extremum may be written as

$$h^4 - \frac{\alpha h^2}{2\beta} - \frac{\nu_2}{4\beta h n} = 0. \tag{4}$$

Thus, the optimal bandwidth for a sufficiently great size of sample is determined from the equation (4).

Provided that $\alpha > 0$, $\beta > 0$, the solution of this biquadratic equation is

$$h^2 = \frac{\alpha}{4\beta} + \frac{\alpha}{4\beta} \sqrt{1 + \frac{4\beta\nu_2}{\alpha^2 n h}}. \tag{5}$$

Assuming (1), let's find conditions which α and β should satisfy. There are two situations:

Situation 1.

The data arrive online and interval Δ remains constant. By using (3), we identify the order of α and β. It is possible to show that at this situation $\alpha \cong An^k$, $k \in [-1,\, 0]$, $\beta \cong Bn^m$, $m \leq 0$, where $\phi(n) \cong \varphi(n)$ means $\lim\limits_{n\to\infty} \frac{\phi(n)}{\varphi(n)} = 1$. Let's find conditions which α and β should satisfy to provide (1) as $n \to \infty$. The result depends on the behavior of the radicand in expression (5) as $n \to \infty$:

Case 1. The second term in radicand tends to zero as $n \to \infty$.

Case 2. The second term in radicand tends to infinity as $n \to \infty$.

Case 3. The second term in radicand tends to constant c as $n \to \infty$.

Then using (1) and asymptotic expression of radicand we have $h^2 \cong \frac{\alpha}{2\beta} \cong \frac{An^k}{2Bn^m}$ for case 1, $h^2 \cong \frac{\alpha}{4\beta}\sqrt{\frac{4\beta\nu_2}{\alpha^2 n h}}$ or $h \cong \left(\frac{\nu_2}{4\beta n}\right)^{1/5} \cong \left(\frac{\nu_2}{4Bn^{m+1}}\right)^{1/5}$ for case 2, $h \cong \frac{4B\nu_2}{A^2 c} n^{m-2k-1}$ or $h^2 \cong \frac{\alpha}{4\beta} + \frac{\alpha}{4\beta}\sqrt{1+c} \cong \frac{An^k}{4Bn^m}\left(1 + \sqrt{1+c}\right)$ for case 3. Respectively we get the following set of inequalities:

$$\begin{cases} k - m < 0, \\ k - m + 2 > 0, \\ \frac{3m}{2} - \frac{5k}{2} - 1 < 0, \\ -1 \leq k \leq 0, \\ m \leq 0 \end{cases} \text{ for case 1, } \begin{cases} m + 1 > 0, \\ m < 4, \\ \frac{6m}{5} - 2k - \frac{4}{5} > 0, \\ -1 \leq k \leq 0, \\ m \leq 0 \end{cases} \text{ for case 2, } \begin{cases} 3m - 5k - 2 = 0, \\ -1 < k \leq 0, \\ -1 < m \leq 0 \end{cases}$$

for case 3. The solutions of these sets are shown in figure 1 (left side): domain a for case 1, domain b for case 2, boundary between domains a and b for case 3.

Situation 2.

There is a realization of process X on time interval $[t,\, t+T]$. Then to construct a nonparametric estimation of density we use n observations of process X from this finite interval. Time Δ between observations may be determined as $\Delta = T/(n-1)$. It is obvious, that $\Delta \to 0$ as $n \to \infty$. Joint density function in definitions of α and β tends to marginal function which is multiplied by delta-function (at least, for finite values j), that is $f_j(x,\, y) \underset{n\to\infty}{\longrightarrow} f(x)\delta(y-x)$. Let's notice, that in definitions of α and β we have $y = x$. As a result α and β tend to infinity as $n \to \infty$. Therefore we have that $\alpha \cong An^k$, $k > 0$, and $\beta \cong Bn^m$, $m > 0$.

Let's find conditions which α and β should satisfy to provide (1). We can consider three cases the same as at Situation 1. The asymptotic expressions for h are identical. We have the following sets of inequalities:

$$\begin{cases} k-m<0, \\ k-m+2>0, \\ \frac{3m}{2}-\frac{5k}{2}-1<0, \\ k>0, \\ m>0 \end{cases} \text{ for case1, } \begin{cases} m+1>0, \\ m<4, \\ \frac{6m}{5}-2k-\frac{4}{5}>0, \\ k>0, \\ m>0 \end{cases} \text{ for case 2, } \begin{cases} 3m-5k-2=0, \\ 0<k<2, \\ 0<m<4 \end{cases}$$

for case 3. The solutions of these sets are shown in figure 1(right side): domain a for case 1, domain b for case 2, boundary between domains a and b for case 3.

3 Formula for Optimal Bandwidth

Thus, to find the optimal h in case 3 of Situation 2 it is necessary to solve the equation of the fifth degree, which results from (4). To simplify a problem, we will construct simple approximation formula for a case 3. We evaluate optimal parameter for function $g(h)$ (3) by use of expression $h^* = \left(\frac{\nu_2}{4n\beta}\right)^{1/5} + a\left(\frac{\nu_2}{4n\beta}\right)^{b}$. To find unknown constants a and b we substitute h^* in the equation of the necessary condition of extremum (4). Let's rewrite (4) as

$$\frac{4\beta n h^5}{\nu_2} = 1 + \frac{2\alpha n h^3}{\nu_2} \tag{6}$$

and decompose $(h^*)^5$ and $(h^*)^3$ by use the Binomial theorem. Substituting these decompositions in the equation (6) gives

$$\frac{4\beta n}{\nu_2}\left(\left(\frac{\nu_2}{4n\beta}\right) + 5\left(\frac{\nu_2}{4n\beta}\right)^{4/5+b} a + 10\left(\frac{\nu_2}{4n\beta}\right)^{3/5+2b} a^2 + \dots\right) =$$

$$= 1 + \frac{2\alpha n}{\nu_2}\left(\left(\frac{\nu_2}{4n\beta}\right)^{3/5} + 3\left(\frac{\nu_2}{4n\beta}\right)^{2/5+b} a + 3\left(\frac{\nu_2}{4n\beta}\right)^{1/5+2b} a^2 + \left(\frac{\nu_2}{4n\beta}\right)^{3b} a^3\right).$$

The first terms on the right hand side and on the left hand side are equal to unity. Equating second terms gives the following formula for optimal bandwidth:

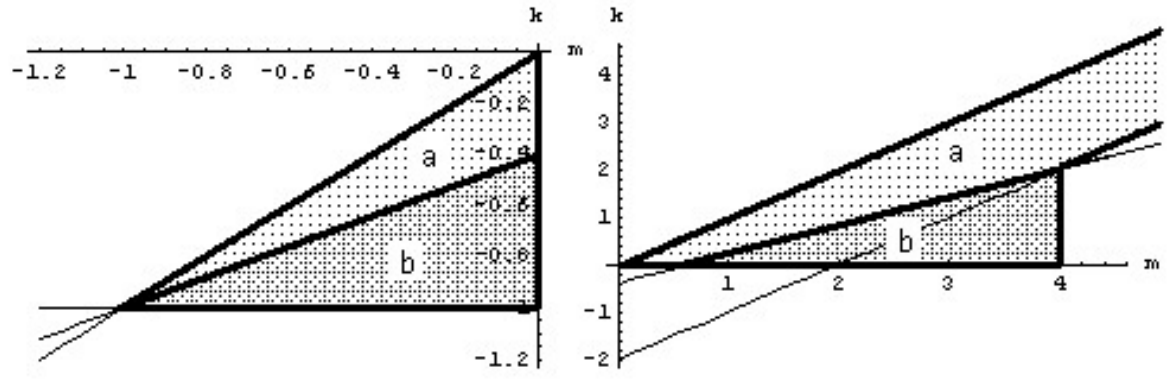

Figure 1: Admissible domains for orders of α and β. Left side for Situation 1. Right side for Situation 2.

$$h^* = \left(\frac{\nu_2}{4n\beta}\right)^{1/5} + \frac{\alpha}{10\beta}\left(\frac{\nu_2}{4n\beta}\right)^{-1/5}. \tag{7}$$

Let's determine orders of α and β to satisfy conditions (1). It is easy to show, that performance of the these conditions for the first and the second terms of (7) results in the following set of inequalities

$$\begin{cases} -1 < 2k - m < 0, \\ -1 < m < 4. \end{cases}$$

It is obvious that formula (7) may be applied not only to a case 3 of Situation 2.

Let's determine now as far as bandwidths determined under the formula (7) and as the solution of the equation (4) are close to each other. Using (6), it is easy to prove:

Proposition 1. *Let the condition $\alpha \le h_g^2\beta$, where h_g is the bandwidth obtained by minimization of function $g(h)$ (3), be held. Then the bandwidth h^* evaluated by formula (7) satisfies to the inequality $0.8h_g < h^* < 1.2\,h_g$.*

Proposition 2. *Let the conditions $\alpha > h_g^2\beta$ and $\sqrt{\frac{\alpha}{2\beta}} \le \sqrt[5]{\frac{\nu_2}{2\beta n}}$ be held. Then the bandwidth h^* evaluated by formula (7) satisfies to the inequality $0.7h_g < h^* < 1.2\,h_g$.*

Thus, we are convinced that for the cases, that were considered in propositions 1 and 2, formula (7) can be used as approximation of optimal bandwidth h_g.

References

[1] Engle R.F., McFadden D.L. (ed.) (1994). *Handbook of Econometrics*. Vol. **IV**, Chapter **38**. Elsevier Science.

[2] Turlach B.A. (1993). *Bandwidth selection in kernel density estimation: a review*. Technical report. Univ. Catholique de Louvain.