Journal of the Nigerian Association of Mathematical Physics
***Volume* 22 (November, 2012), pp 195 – 198**
© J. of NAMP

On the generalized optimal bandwidth for multivariate higher-order nonsymmetic kernels

[1]Afere B. A. and [2]Oyegue F. O.

[1]Department of Mathematics and Industrial Mathematics,
Adekunle Ajasin University, Akungba Akoko, Nigeria.
E – mail: baafere@gmail.com
[2]Department of Mathematics,
University of Benin, Benin City, Nigeria.

Abstract

We present a novel generalized scheme for optimal bandwidth of higher-order multivariate nonsymmetric kernels in kernel density estimation. The conventional way of first specifying the dimension and the order of the bandwidth h before obtaining the global error is removed. The paper however reveals that the convergence rates for this scheme would be faster than those existing in the literature.

Keywords: Multivariate nonsymmetric kernels, density estimation, optimal bandwidth, higher-order kernels, global error.

1.0 Introduction

Kernel density estimation methods are very widely and increasingly used smoothing methods for statistical analysis and its related problems. Basically, it is the construction of an estimate of the true probability density function from the observe data. Since its inception in earlier papers [1, 2]; the focus of attention by most authors has been on symmetric kernels [3, 4] with little attention on nonsymmetric kernels [5, 6]. However, in Afere and Adaja [7], the higher-order version of the nonsymmetric kernels was discussed. All these works have been centred on univariate nonsymmetric kernels. In this paper, we thus extend this to the multivariate setting.

Generally, given *N* samples $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$ drawn from a population with density $f(\mathbf{x})$, the *d* – dimensional multivariate kernel density estimator with the parameterization $\mathbf{H} = h^2.\mathbf{I}_d$ is given by

$$\hat{f}_h(\mathbf{x}) = \frac{1}{Nh^d}\sum_{i=1}^{N} K(\frac{\mathbf{x}-\mathbf{X}_i}{h}) \qquad (1)$$

where $\mathbf{x} = (x_1, x_2, ..., x_d)^{\mathrm{T}}$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{id})^{\mathrm{T}}$, $i = 1, 2, ..., n$; *h* is the bandwidth and *K* is the *d* – variate kernel function which satisfies $\int_{R^d} K(\mathbf{x})d\mathbf{x} = 1$ [8]. Conventionally, *K* is supposed to be symmetric; that is $K(\mathbf{t}) = K(-\mathbf{t})$. This assumption shall be dropped in this paper. In [7], *K* is considered to be higher-order nonsymmetric kernels. The extension of this paper to the multivariate setting shall be the main thrust of this paper.

The bandwidth *h* is the most important characteristics of density estimate [3, 9]. In this case, one computes the bandwidth *h* via the asymptotic mean integrated squared error (AMISE) given by

$$\mathrm{AMISE}\hat{f}_h(\mathbf{x}) = \mathrm{AISB}\hat{f}_h(\mathbf{x}) + \mathrm{AIV}\hat{f}_h(\mathbf{x}) \qquad (2)$$

where $\mathrm{AISB}\hat{f}_h(\mathbf{x}) = (\int \mathrm{E}\hat{f}_h(\mathbf{x})d\mathbf{x} - f(\mathbf{x}))^2$ and $\mathrm{AIV}\hat{f}_h(\mathbf{x}) = \int(\mathrm{E}\hat{f}_h^{\,2}(\mathbf{x}) - \mathrm{E}^2\hat{f}_h(\mathbf{x}))$. On optimizing (2), Silverman [3] showed that for any *d* – dimensional symmetric kernel of order two, (2) can be expressed as

$$\mathrm{AMISE}\hat{f}_h(\mathbf{x}) \cong \frac{1}{4}h^4\alpha^2\int_{R^d}\{\nabla^2 f(\mathbf{x})\}^2 d\mathbf{x} + n^{-1}h^{-d}\beta \qquad (3)$$

where $\alpha = \int_{R^d} t_1^2 K(\mathbf{t})d\mathbf{t}$ and $\beta = \int_{R^d} K(\mathbf{t})^2 d\mathbf{t}$, and Osemwenkhae [10] extended this to order four and obtained

Corresponding author: ***Afere B. A.,*** E-mail: baafere@gmail.com, Tel. +234 8050502000

$$\text{AMISE}\hat{f}_h(x) \cong (\frac{d+8}{8d})(\frac{8}{(4!)^2})^{\frac{d}{d+8}}(d(K_4.I_d)^2\left\|\nabla^2 f\right\|_2^2)^{\frac{8}{d+8}} n^{-\frac{8}{d+8}} \quad (4)$$

These were extended to the higher-order kernel of order ($2m + 2$) by Afere [11] and obtained the generalized expression as

$$\text{AMISE}\hat{f}_h(\mathbf{x}) \cong (\frac{d+4m+4}{d(4m+4)})[\frac{(4m+4)}{((2m+2)!)^2}]^{\frac{d}{d+4m+4}} \times \quad (5)$$

$$(d \cdot \|K\|_2^2)^{\frac{4m+4}{d+4m+4}}((K_{2m+2}\cdot \mathrm{I}_d)^2)^{\frac{d}{d+4m+4}} \cdot (\left\|\nabla^{2m+2} f\right\|_2^2)^{\frac{d}{d+4m+4}} \cdot n^{-\frac{4m+4}{d+4m+4}}$$

The simplified extension expressions in (4) and (5) were achieved by the symmetry assumption of K [3, 4, 11, 12].

In this paper, as earlier said; we shall focus attention on nonsymmetric kernels. Thus, the assumption of symmetricity of kernel function that has simplified most of the work in literature shall be dropped. Hence, K can take any of the nonsymmetric kernel – exponential, Weibull, Gamma, chi-square, e.t.c. that are available in the literature [13]. The aim of this work, however, is to extend the earlier work on the higher-order nonsymmetric kernels by Afere and Adaja [7]. This shall be done by obtaining the generalized asymptotic optimal bandwidths via the asymptotic mean integrated squared error.

The structure of the paper is as follows. In Section 2, we established the fundamental case of order one. The result is extended to handle the higher-order case in Section 3. Section 4 is dedicated to the proposed generalized scheme. Its methodology and the convergence scheme are also discussed in this section. The discussion of findings is given in Section 5. Finally, a conclusion is provided in Section 6.

2.0 The *d* – dimensional nonsymmetric kernel of order one

Here, we shall first consider when the optimal bandwidth h is of order one. Suppose (1) is defined for the following nonsymmetric regualarity condtions

$$\left.\begin{array}{ll} i. & \int_{R^d} K(\mathbf{t})d\mathbf{t} \\ ii. & \int_{R^d} (\mathbf{t}^{\mathrm{T}}\mathbf{t})K(\mathbf{t})d\mathbf{t} = K_1 < \infty \end{array}\right\} \quad (6)$$

Assuming also that f' and f'' are both continuous and squared integrable as well as $\lim_{N\to\infty} h = 0$ and $\lim_{N\to\infty} Nh = \infty$. Then, it follows from definition (6); that the AMISE of $\hat{f}_h(\mathbf{x})$ in (1) is

$$\text{AMISE}\hat{f}_h(\mathbf{x}) = (\int \mathrm{E}\hat{f}_h(\mathbf{x})d\mathbf{x} - f(\mathbf{x}))^2 + \int (\mathrm{E}\hat{f}_h{}^2(\mathbf{x}) - \mathrm{E}^2\hat{f}_h(\mathbf{x}))d\mathbf{x} \quad (7)$$

On substituting (1) in (7), and by employing the multivariate Taylor series expansion and then applying appropriate simplifications by using the conditions in (6), (7) becomes

$$\text{AMISE}\hat{f}_h(x) \cong h\|\nabla f\|_2^2 (K_1)^2 + N^{-1}h^{-d}\|K\|_2^2 \quad (8)$$

The minimization of (8) with respect to h, results in

$$h_{1,\text{nonsym}} \cong 2^{-\frac{1}{d+2}}(d)^{\frac{1}{d+2}}(\|K\|_2^2)^{\frac{1}{d+2}}(\|\nabla f\|_2^2(K_1)^2)^{-\frac{1}{d+2}} N^{-\frac{1}{d+2}} \quad (9)$$

Equation (9) is the asymptotic optimal bandwidths for a multivariate nonsymmetric kernel of order one. In the next section, we shall extend this to higher-order kernel.

3.0 The higher-order nonsymmetric multivariate kernel

We take the order of the optimal bandwidth h to three. Thus, we modify the regualarity conditions in (6) by imposing the conditions below on (1)

$$\left.\begin{array}{ll} i. & \int_{R^d} K(\mathbf{t})d\mathbf{t} \\ ii. & \int_{R^d} \mathbf{t}^{\mathrm{T}}K(\mathbf{t})d\mathbf{t} = \int_{\square^d} (\mathbf{t}^{\mathrm{T}}\mathbf{t})K(\mathbf{t})d\mathbf{t} = 0 \\ iii. & \int_{R^d} \mathbf{t}^{\mathrm{T}}(\mathbf{t}^{\mathrm{T}}\mathbf{t})K(\mathbf{t})d\mathbf{t} = K_3 < \infty \end{array}\right\} \quad (10)$$

Also, supposing that f', f'', f''' and $f^{(4)}$ are both continuous and squared integrable and that $\lim_{N\to\infty} h = 0$ and $\lim_{N\to\infty} Nh = \infty$ are satisfied; then on substituting (1) into (7) and taking the multivariate Taylor series expansion up to the fifth term; and with the modified regularity conditions in (10), the asymptotic MISE for the d – dimensional nonsymmetric kernel of order three is given as

$$\text{AMISE}\hat{f}_h(\mathbf{x}) \cong \frac{h^6}{(3!)^2}(K_3)^2 \left\|\nabla^3 f\right\|_2^2 + N^{-1}h^{-d}\left\|K\right\|_2^2 \tag{11}$$

If we minimize (11), we have the asymptotic optimal bandwidths h of the multivariate nonsymmetric kernel of order three to be

$$h_{3,\text{nonsym}} \cong (d(3!)^2)^{\frac{1}{d+6}}(6)^{-\frac{1}{d+6}}(\left\|K\right\|_2^2)^{\frac{1}{d+6}}(\left\|\nabla^3 f\right\|_2^2 (K_3)^2)^{-\frac{1}{d+6}} N^{-\frac{1}{d+6}} \tag{12}$$

The nomenclature $h_{3,\text{nonsym}}$ is used for the optimal bandwidth of the multivariate higher-order nonsymmetric kernel of order three. Equation (12) is thus the fundamental equation for the optimal bandwidth of any higher-order nonsymmetric multivariate kernel of order three. In the next section, we shall give the proposed generalized scheme for any d - dimensional higher-order multivariate nonsymmetric kernels.

4.0 The proposed generalized optimal bandwidth scheme

Now, supposing we impose the regularity conditions (13) in (1) by modifying (10), we have

$$\left.\begin{array}{ll} i. & \int_{R^d} K(\mathbf{t})d\mathbf{t} \\ ii. & \int_{R^d} \mathbf{t}^{\mathrm{T}}K(\mathbf{t})d\mathbf{t} = \int_{R^d}(\mathbf{t}^{\mathrm{T}}\mathbf{t})K(\mathbf{t})d\mathbf{t} = \ldots = \int_{R^d}\mathbf{t}^{\mathrm{T}}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{2m-1}K(\mathbf{t})d\mathbf{t} = \int_{R^d}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{2m}K(\mathbf{t})d\mathbf{t} = 0 \\ iii. & \int_{R^d}\mathbf{t}^{\mathrm{T}}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{2m+1}K(\mathbf{t})d\mathbf{t} = K_{2m+1} < \infty \end{array}\right\} \tag{13}$$

Now, on substituting (1) in (7) and taking the multivariate Taylor series expansion of the resultant equation up to the term $(2m + 1)$ and using condition (i) in (13), (7) becomes

$$\begin{aligned} \text{AMISE}\hat{f}_h(\mathbf{x}) &\cong \text{tr}\{(\int_{R^d}\mathbf{t}^{\mathrm{T}}K(\mathbf{t})d\mathbf{t})(D_f h.I_d)\} \\ &+\frac{1}{2!}\text{tr}\{(\int_{R^d}(\mathbf{t}^{\mathrm{T}}\mathbf{t})K(\mathbf{t})d\mathbf{t})((h.I_d)\mathrm{H}_f(h.I_d))\} \\ &-\frac{1}{3!}\text{tr}\{(\int_{R^d}\mathbf{t}^{\mathrm{T}}(\mathbf{t}^{\mathrm{T}}\mathbf{t})K(\mathbf{t})d\mathbf{t})(D_f\left(h.I_d\right))((h.I_d)\mathrm{H}_f(h.I_d))\} \\ &+\frac{1}{4!}\text{tr}\{(\int_{R^d}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^2K(\mathbf{t})d\mathbf{t})((h.I_d)\mathrm{H}_f(h.I_d))^2\} \\ &-\frac{1}{5!}\text{tr}\{(\int_{R^d}\mathbf{t}^{\mathrm{T}}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^2K(\mathbf{t})d\mathbf{t})(D_f(h.I_d))((h.I_d)\mathrm{H}_f(h.I_d))^2\} \\ &+\frac{1}{6!}\text{tr}\{(\int_{R^d}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^3K(\mathbf{t})d\mathbf{t})((h.I_d)\mathrm{H}_f(h.I_d))^3\} - \cdots + \cdots - \\ &\frac{1}{(2m+1)!}\text{tr}\{(\int_{R^d}\mathbf{t}^{\mathrm{T}}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^m K(\mathbf{t})d\mathbf{t})(D_f(h.I_d))((h.I_d)\mathrm{H}_f(h.I_d))^m\})d\mathbf{t} \end{aligned} \tag{14}$$

On simplifying (14) and imposing condition (ii) in (13), we have that $\text{AMISE}\hat{f}_h(\mathbf{x})$ in (14) reduces to

$$\text{AMISE}\hat{f}_h(\mathbf{x}) \cong \frac{h^{4m+2}}{((2m+1)!)^2}(K_{2m+1}.I_d)^2\left\|\nabla^{2m+1}f\right\|_2^2 + N^{-1}h^{-d}\left\|K\right\|_2^2 \tag{15}$$

On differentiating (15), we have

$$\frac{\delta\text{AMISE}\hat{f}_h(\mathbf{x})}{\delta h} = \frac{(4m+2)h^{4m+1}}{((2m+1)!)^2}(K_{2m+1}.I_d)^2\left\|\nabla^{2m+1}f\right\| - dN^{-1}h^{-(d+1)}\left\|K\right\|_2^2$$

But, at the minimum or maximum point, $\dfrac{\delta\text{AMISE}\hat{f}_h(\mathbf{x})}{\delta h} = 0$

$$\therefore \quad \frac{(4m+2)h^{4m+1}}{((2m+1)!)^2}(K_{2m+1}.I_d)^2\left\|\nabla^{2m+1}f\right\| = dN^{-1}h^{-(d+1)}\left\|K\right\|_2^2$$

$$\Rightarrow \quad h^{d+4m+2} = \frac{((2m+1)!)^2\left\|K\right\|_2^2 dN^{-1}}{(4m+2)(K_{2m+1}.I_d)^2\left\|\nabla^{2m+1}f\right\|}$$

$$\Rightarrow \quad h = \frac{(((2m+1)!)^2 \|K\|_2^2 \, dN^{-1})^{\frac{1}{d+4m+2}}}{((4m+2)(K_{2m+1}.I_d)^2 \left\|\nabla^{2m+1} f\right\|)^{\frac{1}{d+4m+2}}}$$

Therefore,

$$h_{(d+4m+2),\text{nonsym}} \cong (d((2m+1)!)^2 \|K\|_2^2)^{\frac{1}{d+4m+2}} \times ((4m+2)(K_{2m+1}.I_d)^2 \left\|\nabla^{2m+1} f\right\|)^{-\frac{1}{d+4m+2}} N^{-\frac{1}{d+4m+2}} \tag{16}$$

Equation (16) is thus the required expression for the generalized asymptotic optimal bandwidth corresponding to the regularity conditions in (13) for any higher-order multivariate nonsymmetric kernels.

5.0 Findings and discussion

On comparing (12) with (16), we found out that the convergence rates of the asymptotic optimal bandwidth for any higher-order multivariate nonsymmetric kernels has increased from $N^{-\frac{1}{d+6}}$ to $N^{-\frac{1}{d+4m+2}}$ $\forall\, d = 1, 2, ..., < \infty;\, m = 1, 2, ..., < \infty$. This convergence rate is made possible with the aid of the regularity conditions in (13). The proposed scheme however has not only given us the generalized optimal bandwidth but also enhances the rates of convergence of the global error for any higher-order multivariate nonsymmetric kernels.

6.0 Concluding remarks

In this paper, we have been able to give the generalized form of the optimal bandwidth for the *d*-dimensional higher-order multivariate nonsymmetric kernels. This will surely give helping hands to the researchers in this area of mathematical statistics since it has completely removed the problem of first specifying the order of *h* and its dimension (*d*) before obtaining the optimal bandwidth when calculating the global error of any higher-order multivariate nonsymmetric kernels.

References

[1] Fix, E. and Hodges, J. L. (1951), 'Discriminatory analysis:-nonparametric discrimination: consistency properties', Report No.4, Project No. 21 – 29 – 004, USAF School of Aviation Medicine, Randolph Field, Texas.

[2] Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *Annals of Mathematical Statistics* **27**, 832 – 837.

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

[4] Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

[5] Osemwenkhae, J. E. and Izevbizua, O. (2005), 'Generalized forms in nonsymmetric univariate kernels', *Nigerian Journal of Applied Science*, **23**, 226 – 230.

[6] Osemwenkhae, J. E. and Orhionkpaiyo, B. C. (2006), 'The nonsymmetric univariate kernels', *Global Journal of Mathematical Sciences*, **5**, 175 – 178.

[7] Afere, B. A. and Adaja, J. I. (2011), 'On higher-order nonsymmetric univariate kernels', Proceedings of September 2011 Annual National Conference of Mathematical Association of Nigeria (M.A.N); Edited by Profs. J. S. Sadiku, M. O. Fajemidagba and Dr. R. B. Adeniyi. Pp 81 – 86.

[8] Cacoullos, T. (1966), 'Estimation of a multivariate density', *Annals of the Institute of Statistical Mathematics* **18**, 179 – 189.

[9] Raudys, S. (1991), 'On the effectiveness of Parzen window classifier', *Information*, Vol 2, No. 3, 434 – 454.

[10] Osemwenkhae, J. E. (2003), 'Higher-order forms in kernel density estimation', Ph.D Thesis, University of Benin, Benin City, Nigeria.

[11] Afere, B. A. (2010), 'A family of multivariate higher-order hybrid polynomial kernels in kernel density estimation', Ph. D Thesis, University of Benin, Benin City, Nigeria.

[12] Scott, D. W. (1992), *Multivariate Density Estimation*: *Theory, practice and visualization*, John Wiley & Sons Inc. New York.

[13] Rohatgi, V. J. (1984), *Statistical Inference*, John Wiley & Sons, New York.