The Region of Positivity and Unimodality in the Truncated Series of a Nonparametric Kernel Density Estimator

A.K. Gupta* & B.K.K im**

Abstract

This paper approximates to a kernel density estimate by a truncated series of expansion involving Hermite polynomials, since this could ease the computational burden involved in the kernel-based density estimation. However, this truncated series may give negative values particularly near the tails of distribution, or may give a multimodal estimate when we are estimating unimodal density. In this paper we will show a way to insure the truncated series to be positive and unimodal so that the approximation to a kernel density estimator would be meaningful.

1. Introduction

It is very interesting to find that when we expand the nonparametric kernel density estimator of Parzen (1962) with the standard normal kernel, we obtain an infinite series of expansion involving the sample moments and the usual Hermite polynomials. The sum of a finite number of terms of this series may then be used to approximate the kernel density estimator. This seems a potentially useful device for easing the computational burden involved in the kernel-based density estimation. However, care would have to be taken that the approximation afforded by the truncation is satisfactory since the sum of a finite number of terms of the series may give negative values particularly near the tails, may behave badly in the sense that the sum of k terms may give a worse fit than the sum of $(k-1)$ terms, or may give a multimodal density estimator when true density is unimodal. See examples in Specht (1971).

The parallel with Barton and Dennis (1952) may then be exploited to define positive and unimodal regions for the approximated (truncated) kernel density estimator so that we may assess positivity and unimodality of the approximated estimator by checking where the approximated estimator lies relative to these regions. In Barton and Dennis

* Professor in the Department of Mathematics and Statistics, Bowling Green State University, Ohio, U.S.A.

** Professor in the Department of Mathematics and Statistics, Memorial University of Newfoundland, Canada. Research partially supported by National Sciences and Engineering Research Council of Canada Grant No. 4527, 1982. (Visiting Proressor in the Department of Statistics, Korea University)

(1952) the positivity and unimodality regions for the Gram-Charlier and the Edgeworth series estimator involving the first four known moments were plotted in the two dimensional space of (β_1, β_2) where $\beta_1=\mu_3^2/\mu_2^3$, and $\beta_2=\mu_4/\mu_2^2$.

In this paper it is proposed to assess unimodality and positivity of the approximated density estimator by seeing where the estimate lies relative to the regions plotted in the two dimensional space of $(\hat{\beta}_1, \hat{\beta}_2)=(M_3^2/M_2^3,\ M_4/M_2^2)$. for example, when the truncated series involves only the first four sample moments, M_1, M_2, M_3, and M_4.

2. The Series Expansion of the Kernel Density Estimator.

In an attempt to investigate whether an unknown denisty is unimodal or not, it seems to be natural to examine the first derivative of a nonparametric kernel density estimator of Parzen (1962).

Let $X_1, X_2, \cdots, X_n$ be n independent and identically distributed random variables whose common density $f(x)$ is continuous, but unknown. Parzen(1962) proposed the nonparametric kernel density estimator $f_n(x)$ of the form,

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

where x is a point of estimation, $\{h_n\}$ is a suitably chosen sequence of positive numbers, and $K(\cdot)$ is a symmetric, nonnegative kernel (weight) function such that $\int_{-\infty}^{\infty}K(u)\,du=1$. Here in this note, we use the standard normal kernel,

$$K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \tag{2}$$

Using (2) in (1) we obtain

$$f_n(x)=\frac{1}{\sqrt{2\pi}nh_n^2}e^{-\left(\frac{x}{2h_n}\right)^2}\sum_{i=1}^{n}e^{\left(\frac{X_i}{h_n}\right)\frac{x}{h_n}-\frac{1}{2}\left(\frac{X_i}{h_n}\right)^2} \tag{3}$$

Now substituting the identity,

$$e^{\frac{X_i}{h_n}\frac{x}{h_n}-\frac{1}{2}\left(\frac{X_i}{h_n}\right)^2}=\sum_{k=0}^{\infty}\frac{\left(\frac{X_i}{h_n}\right)^k}{K!}H_k\left(\frac{x}{h_n}\right)$$

in (3) whree $H_k\left(\frac{x}{h_n}\right)$ is the Hermite polynomial of degree k in $\frac{x}{h_n}$,

we get,

$$f_n(x)=\frac{1}{\sqrt{2\pi}h_n^2}e^{-\frac{1}{2}\left(\frac{x}{h_n}\right)^2}\sum_{k=0}^{\infty}\frac{M_k'}{k!h_n^k}H_k\left(\frac{x}{h_n}\right) \tag{4}$$

where $M_k'=\frac{1}{n}\sum_{i=1}^{n}X_i^k$, the kth sample moment.

3. The Positivity and Unimodality Regions of $f_n(x)$.

In practical application one makes use of only the first four sample moments from which the estimates of the mean, variance, skewness, and kurtosis are obtained. This

is especially true since the sampling variances of moments higher than the fourth are large, and furthermore it is true that often, the sum of k terms may give a worse fit than the sum of $(k-1)$ terms in the expansion. Hence, here we shall limit our attention to the series expansion of $f_n(x)$ in (4) with the terms involving only the first m sample moments. Therefore the resulting expansion will be as follows :

$$f_n(x)=\frac{1}{\sqrt{2\pi}h_n}e^{-\frac{1}{2}\left(\frac{x}{h_n}\right)^2}\left[1+\sum_{k=1}^{m}b_k'\,H_k\left(\frac{x}{h_n}\right)\right] \tag{5}$$

where $b_k'=\frac{M_k'}{k!h_n{}^k}$

Now replacing M_k' in (5) by the moments about the sample mean

$$M_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-M_1')^k,$$

(5) becomes

$$f_n(x)=\frac{1}{\sqrt{2\pi}h_n}e^{-\frac{1}{2}\left(\frac{x}{h_n}\right)^2}\left[1+\sum_{k=2}^{m}b_kH_k\left(\frac{x}{h_n}\right)\right] \tag{6}$$

where $b_k=\frac{M_k}{k!h_n{}^k}$.

Hence, the condition $\frac{df_n(x)}{x}=0$ is equivalent to the condition that

$$H_1\left(\frac{x}{h_n}\right)+h_n\sum_{k=2}^{m}b_kH_{k+1}\left(\frac{x}{h_n}\right)=0 \tag{7}$$

since $\frac{x}{h_n}H_k\left(\frac{x}{h_n}\right)-kH_{k-1}\left(\frac{x}{h_n}\right)=H_{k+1}\left(\frac{x}{h_n}\right)$, and the condition $\frac{d^2f_n(x)}{dx^2}\leqq 0$ is equivalent to the condition that

$$H_2\left(\frac{x}{h_n}\right)+\sum_{k=2}^{m}b_kH_{k+2}\left(\frac{x}{h_n}\right)=0 \tag{8}$$

by the virtue of the identity that

$$H_{k+2}=H_2H_k-2kH_1H_{k-1}+k(k-1)H_{k-2}$$

for the Hermite polynomials.

Conseqently, by (7) and (8), the unimodality region is given parametrically by

$$H_1\left(\frac{x}{h_n}\right)+\sum_{k=2}^{m}b_kH_{k+1}\left(\frac{x}{h_n}\right)\equiv 0\equiv H_2\left(\frac{x}{h_n}\right)+\sum_{k=2}^{m}b_kH_{k+2}\left(\frac{x}{h_n}\right) \tag{9}$$

Now for the positivity region of the truncated series, note that $f_n(x)$ of (5) will be nonnegative if the quantity,

$$\left[1+\sum_{k=1}^{m}b_k'H_k\left(\frac{x}{h_n}\right)\right]\geqq 0 \text{ for all } x. \tag{10}$$

If we consider $(b_1', b_2', \cdots, b_m')$ as a point in m dimensional space, then the relation (10) implies that $(b_1', b_2', \cdots, b_m')$ lies on the same side as $(0, 0, \cdots, 0)$, and by proceeding parallel with Barton and Dennis (1952) we will find that the positivity region of $f_n(x)$ is given by points $(b_1', b_2', \cdots, b_m')$ on the surface defined parametrically by

$$1+\sum_{k=1}^{m} b_k' H_k\left(\frac{x}{h_n}\right)\equiv 0\equiv \sum_{k=1}^{m} b_k'\ kH_{k+1}\left(\frac{x}{h_n}\right). \tag{11}$$

Solution loci for the two parametric equations (9) and (11) may then be plotted to give the regions of unimodality and positivity, respectively, of the truncated series of the kernel density estimator. In particular, if $m=4$ the parametric equations (9) and (11) replacing $M_k'$by M_k like in (6) will involve only the three sample moments M_2, M_3, and M_4 and we may therefore plot two dimensional regions of unimodality and positivity of $f_n(x)$ in (6) on the $(\hat{\beta}_1, \hat{\beta}_2)=(M_3^2/M_2^3, M_4/M_2^2)$ plane by proceeding parallel with Barton and Dennis (1952).

The regions of Barton and Dennis (1952) were replotted by Draper and Tierney (19 72) by a nonlinear least square fitting subroutine, which had an advantage of being extendable to more complicatd cases when more terms are added in the approximation, that is, when $m>4$ in (9) and (11).

We note here that the regions in our case will depend on choice of the sequence $\{h_n\}$, the "smoothing parameter" for the kernel density estimator.

4. Concluding Remarks.

From a practical viewpoint, the important consideration is not whether an infinite series of the kernel density estimate can satisfactorily estimate a density function, but whether the sum of a finite number of terms can also satisfactorily approximate an unknown density. The regions of unimodality and positivity in the truncated series of the kernel density estimator, in this sense, should be of great importance in density estimation in order to determine the range of x over which the approximation will be meaningful.

In Specht (1971), an example plotted of the truncated series in terms of the Hermite polynomials gave a substantial amount of negative values for the density estimator near the left hand tail of the normal density. It could also be noted in Specht (1971) that a truncated estimator gave a multimodal estimator when it was estimating a normal density function.

The finite series may be successful in cases of moderate skewness. For a value of (β_1, β_2) distant from that of the normal density $(0, 3)$, the approximation may very well be negative at least for some part of x. For many statistical purposes we are often interested in the tails of a distribution, and it is here that such inadequacies occur. The regions of positivity and unimodality we consider here may come in handy when such inadequacies are serious.

5. Some Further Work.

As a further work, such regions for unimodality and positivity with various choice of the smoothing parameter $\{h_n\}$ in kernel-based density estimation, should actually be

plotted.

With such a trucation in the "orthogonal series" density estimators such as in Tarter and Kronmal (1968), we would not expect $f_n(x)$ to be nonnegative for all x. Consequently, a similar method of insuring positivity, or unimodality should be proposed for many orthogonal series density estimators available in density estimation.

References

(1) Barton, D.E., aud Dennis, K.E. (1952). The conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. *Biometrika*, 39, 425~427.

(2) Draper, N.R., and Tierney, D. E. (1972). Regions of positive and unimodal series expansion of the Edgeworth and Gram-Charlier approximations. *Biometrika*, 59, 463~465.

(3) Kendall, M.G., and Stuart, A. (1963). *The Advanced Theory of Statistics*, Vol. 1, 2nd ed., Griffin, London.

(4) Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.*, 33, 1065~1076.

(5) Specht, D.F. (1971). Series estimation of a probability density function. Technometrics, vol. 13, No. 2, 409~424.

(6) Tarter, M.E. and Kronmal, R.A. (1968). The estimation of probability densities and cumulatives by Fouier series methods. Journal of the American Statistical Association, Vol. 63, 925~952.