The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 911 - 915

A Note on Smoothing Distribution Function Estimation[1)]

In-Sun Chu[2)] and Jae-Ryong Choi[3)]

Abstract

The purpose of this paper is to consider the problem of selection of optimal smoothing parameter for kernel-type distribution function estimator, which asymptotically minimizes mean Hellinger distance.

1. Introduction

Let $X_1, X_2, \cdots, X_n$ be a random sample from an unknown continuous distribution function $F(x)$ whit density function $f(x)$ and denote by $F_n(x)$ corresponding to empirical distribution function at point x. The traditional nonparametric estimator of $F(x)$ is empirical distribution function. Despite the good statistical properties of $F_n(x)$, one could prefer in many applications a rather smooth estimate. A nonparametric alternative to $F_n(x)$ for smooth distribution function has been introduced in Nadaraya(1964). This estimate is defined as

$$\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)=\int_{-\infty}^{x}\hat{f}_n(y)dy$$

where $\hat{f}_n(x)$ is the well known Rosenblatt-Parzen density estimate defined by

$$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right)$$

with the kernel function $k(t)=K'(t)$ and the smoothing parameter h. We will classically assume in the following that the smoothing parameter $h\to 0$ and $nh\to\infty$ as $n\to\infty$ and the function k (or K) are to be chosen by the user.

Mainly for reasons of tractability, we assess how well $\hat{F}_n(x)$ estimates $F(x)$ by the mean intrgrated squared error (MISE) given by $\int_{-\infty}^{\infty} E\{\hat{F}_n(x)-F(x)\}^2 dx$ under the following two

1) This research was supported by the Basic Science Research Institute, Dong-A University, 1995.
2) Lecturer, Department of Mathematics, College of Natural Science, Dong-A University, Pusan, 604-714, Korea.
3) Professor, Department of Mathematics, College of Natural Science, Dong-A University, Pusan, 604-714, Korea.

conditions;

(A1) $\int_{-\infty}^{\infty} k(t)dt=1$, $\int_{-\infty}^{\infty} tk(t)dt=0$ and $0<\int_{-\infty}^{\infty} t^2 k(t)dt= x_2<\infty$

(A2) $F(x)$ is twice continuously differentiable with bounded $f'(x)$.

Then, if $h\to 0$ as $n\to\infty$, we have under the conditions (A1) and (A2) (see e.g., Swanepoel(1988) and Jones(1990)),

$$MISE(\hat{F}_n(x))=n^{-1}\int_{-\infty}^{\infty}F(x)\{1-F(x)\}dx-2hn^{-1}C_k + \frac{1}{4}h^4 x_2^2\int_{-\infty}^{\infty}f'(x)^2dx+O(h^4+hn^{-1})$$

where C_K is a constant depending only on the kernel function defined as

$$C_K=\int_{-\infty}^{\infty}tk(t)K(t)dt$$

We noted that $n^{-1}\int_{-\infty}^{\infty}F(x)\{1-F(x)\}dx$ is the MISE of the empirical distribution function $F_n(x)$ (essentially $\hat{F}_n(x)$ with $h=0$).

In this paper, we consider the selection of the optimal smoothing parameter that asymptotically minimizes mean Hellinger distance (MHD) between a kernel estimator $\hat{F}_n(x)$ and distribution function $F(x)$, as given by

$$MHD(\hat{F}_n(x))=\int_{-\infty}^{\infty}E\{\hat{F}_n(x)^{1/2}-F(x)^{1/2}\}^2dx$$

and compare with the case of $MISE(\hat{F}_n(x))$. The Hellinger distance was suggested by Pitman(1979) as an aid to study maximum likelihood. For MHD applied to kernel density estimation, see Kanazawa(1993).

2. Asymptotic Results

Throughout this section, we assume that the kernel function k and K satisfies (B1) $k(t)$ is a probability density such that it is bounded, symmetric around the origin and has a finite support. Thus, the function $k(t)$ satisfies

$$\int_{-\infty}^{\infty}k(t)dt=1,\ \int_{-\infty}^{\infty}tk(t)dt=0 \text{ and } \int_{-\infty}^{\infty}t^4k(t)dt<\infty.$$

Furthermore, $\int_{-\infty}^{\infty}t^2k(t)dt=x_2<\infty$ and $\int_{-\infty}^{\infty}t^4k(t)dt=x_4<\infty$.

We also assume that the underlying distribution function $F(x)$ satisfies (B2) $F(x)$ is four times continuously differentiable with bounded derivatives.

The main result of this paper is now stated in the following.

Theorem 1. Under the condition (*B*1) and (*B*2), assuming that the smoothing parameter $h \to 0$ as $n \to \infty$, we have

$$MHD(\hat{F}_n(x)) = n^{-1}\frac{1}{4}\int_{-\infty}^{\infty}\frac{1-F(x)}{F(x)}dx + hn^{-1}C_K\frac{1}{4}\int_{-\infty}^{\infty}\frac{f(x)}{F(x)}dx + h^4x_2^2\frac{1}{16}\int_{-\infty}^{\infty}\frac{f'(x)^2}{F(x)}dx + O(h^6+n^{-1}),$$

where C_K is defined in section 1.

Remarks. (1). From the above result, we can easily select the smoothing parameter in the sense of asymptotically minimizing $MHD(\hat{F}_n(x))$. The term $n^{-1}(1/4)\int_{-\infty}^{\infty}\{1-F(x)\}/F(x)dx$ is the MHD of the empirical distribution function $F_n(x)$.

(2). Next, we note that the constant C_K is positive for many kernel function K. For example, if the kernel function K is standard normal distribution, we have $C_K = 1/\sqrt{2\pi}$.

(3). Hence there does not exist essentially difference between the estimators by empirical distribution function and a kernel estimator $\hat{F}_n(x)$ in the sense of MHD. This result is the same as the case of $MISE(\hat{F}_n(x))$.

Proof of Theorem. Elementary calculations show that

$$E\{\hat{F}_n(x)\} = \frac{1}{n}\sum_{i=1}^{n}E\left\{K(\frac{x-X_i}{h})\right\} = \int_{-\infty}^{\infty}K(\frac{x-y}{h})dF(y) = \int_{-\infty}^{\infty}k(t)F(x-th)dt.$$

From a Taylor's expansion, we have

$$F(x-ht) = F(x) - htf(x) + \frac{1}{2}h^2t^2f'(x) - \frac{1}{6}h^3t^3f''(x) + \frac{1}{24}h^4t^4f'''(x) + O(h^5).$$

Therefore the expected value of the kernel estimator is written as

$$E\{\hat{F}_n(x)\} = F(x)\left[1 + \frac{1}{2}h^2\frac{f'(x)}{F(x)}k_2 + \frac{1}{24}h^4\frac{f'''(x)}{F(x)}k_4 + O(h^5)\right],$$

where

$$\begin{aligned} Var\{\hat{F}_n(x)\} &= \frac{1}{n}\sum_{i=1}^{n}Var\left\{K(\frac{x-X_i}{h})\right\} \\ &= \frac{1}{n}\int_{-\infty}^{\infty}K^2(\frac{x-y}{h})dF(y) - \int_{-\infty}^{\infty}K(\frac{x-y}{h})^2dF(y) \\ &= \frac{2}{n}\int_{-\infty}^{\infty}k(t)K(t)F(x-th)dt - \left\{\int_{-\infty}^{\infty}K(t)F(x-th)dt\right\}^2 \\ &= n^{-1}F(x)\{1-F(x)\} + hn^{-1}f(x)C_K + O(n^{-1}). \end{aligned}$$

where C_K is defined in section 1. The terms involving $\int_{-\infty}^{\infty}tk^2(t)dt$, $\int_{-\infty}^{\infty}t^3k^2(t)dt$,

$\int_{-\infty}^{\infty} t^5 k^2(t)dt$, ···disappear, because $k^2(t)$ is symmetric about 0, and Cauchy-Schwartz inequality along with the condition $\int_{-\infty}^{\infty} k^4(t)dt<\infty$ of (*B1*) requires

$$\int_{-\infty}^{\infty} t^{2m} k^2(t)dt \le \left\{\int_{-\infty}^{\infty} t^{4m}dt\right\}^{1/2}\left\{\int_{-\infty}^{\infty} k^4(t)\,dt\right\}^{1/2}<\infty.$$

Then,with a random variable $\xi = O_p(1)$ whose expectation is 0 and variance 1, we can write $\widehat{F_n}(x)$ as

$$\widehat{F_n}(x) = F(x)\Big[1+\frac{1}{2}h^2\frac{f'(x)}{F(x)}k_2+\frac{1}{24}h^4\frac{f'''(x)}{F(x)}k_4 + \left\{\frac{1}{n}\frac{1-F(x)}{F(x)}+\frac{h}{n}\frac{f(x)}{F^2(x)}C_K\right\}^{1/2}\cdot\xi\Big]+O((h^6)+O_p(n^{-1/2}),$$

where the $O(h^6)$ terms depend upon x. Using $(1+z)^{1/2}=1+\frac{1}{2}z-\frac{1}{8}z^2+O(z^3)$, we have

$$\widehat{F(x)}^{1/2} = F(x)^{1/2}\Big[1+\frac{1}{4}h^2\frac{f'(x)}{F(x)}x_2+\frac{1}{48}h^4\frac{f'''(x)}{F(x)}x_4 + \left\{\frac{1}{n}\frac{1-F(x)}{F(x)}+\frac{h}{n}\frac{f(x)}{F^2(x)}C_K\right\}^{1/2}\cdot\frac{1}{2}\xi-\frac{1}{16}h^4\left\{\frac{f'(x)^2}{F(x)}\right\}^2 x_2{}^2 - \left\{\frac{1}{n}\frac{1-F(x)}{F(x)}+\frac{h}{n}\frac{f(x)}{F^2(x)}C_K\right\}\cdot\frac{1}{8}\xi^2\Big]+O(h^6)+O_p(n^{-1}),$$

the term involving ξ vanishes upon taking expectation for $E(\xi)=0$. Hence we obtain

$$\begin{aligned}\int_{-\infty}^{\infty} E\{\widehat{F(x)}^{1/2}F(x)^{1/2}\}dx &= \int_{-\infty}^{\infty}F(x)dx+h^2x_2\frac{1}{4}\int_{-\infty}^{\infty}f'(x)dx+h^4x_4\frac{1}{48}\int_{-\infty}^{\infty}f'''(x)dx\\ &\quad -h^4x_2{}^2\frac{1}{16}\int_{-\infty}^{\infty}\frac{f'(x)^2}{F(x)}dx-n^{-1}\frac{1}{8}\int_{-\infty}^{\infty}\frac{1-F(x)}{F(x)}dx\\ &\quad +hn^{-1}C_K\frac{1}{8}\int_{-\infty}^{\infty}\frac{f(x)}{F(x)}dx+O(h^6+n^{-1}).\end{aligned}$$

Then, we have

$$\begin{aligned}MHD(\widehat{F_n}(x)) &= \int_{-\infty}^{\infty}E\{\widehat{F_n}(x)\}dx+\int_{-\infty}^{\infty}F(x)dx-2\int_{-\infty}^{\infty}E\{\widehat{F_n}(x)^{1/2}F(x)^{1/2}\}dx\\ &= n^{-1}\frac{1}{4}\int_{-\infty}^{\infty}\frac{1-F(x)}{F(x)}dx+hn^{-1}C_K\frac{1}{4}\int_{-\infty}^{\infty}\frac{f'(x)}{F(x)}dx\\ &\quad +h^4x_2{}^2\frac{1}{16}\int_{-\infty}^{\infty}\frac{f'(x)^2}{F(x)}dx+O(h^6+n^{-1}),\end{aligned}$$

as required.

References

[1] Jones, M. C. (1990). The Performance of Kernel Density Estimations in Kernel Distribution Function Estimation. Statistics and Probability Letters, 9, 129-132.

[2] Kanazawa, Y. (1993). Hellinger Distance and Kullback-Leibler Loss for the Kernel Density Estimator. Statistics and Probability Letters, 18, 315-321.

[3] Nadaraya, E. A. (1964). Some New Estimators for Distribution Function. Theory of Probability and Applications, 9, 479-500.

[4] Pitman, E. J. G. (1979). Some Basic Theory for Statistical Inference. Champman and Hall, London.

[5] Swanepoel, W. H. (1988). Mean Integrated Squared Error Properties and Optimal Kernels When Estimating a Distribution Function. Communication Statistics-Theory and Methods, 17, 3785-3799.