Sequential Procedures for Nonparametric Kernel Density Estimation

Basil M. de Silva

Royal Melbourne Institute of Technology, Department of Statistics and Operations Research
GPO Box 2476V
Melbourne 3001, Australia
desilva@rmit.edu.au

Nitis Mukhopadhyay

University of Connecticut, Department of Statistics
Storrs, CT 06269-3120, U.S.A.
MUKHOP@UCONNVM.UCONN.EDU

1. Introduction

In this paper we employ sequential procedures to estimate the sample size, N required to obtain a fixed-width confidence interval for an unknown density function, $f(x)$ at a point $x = x_0$. We use the kernel density estimator

$$\widehat{f}_{n,h_n}(x) = (nh_n)^{-1}\sum_{i=1}^{n} K\left[(x - X_i)/h_n\right], \tag{1}$$

first introduced by Rosenblatt (1956). Here $X_1, X_2, \cdots, X_n$ is a set of independent and identically distributed (iid) random variables with a probability density function (pdf) $f(x)$, $K(.)$ is a bounded density function, known as a kernel, and h_n is the band-width. The goal is to construct a fixed-width confidence interval I_n for $f(x)$ at a given point $x = x_0$ with the preassigned coverage probability $1 - \alpha$, that is, to have $\mathcal{P}\left\{f(x_0) \in I_n\right\} \geq 1 - \alpha$ for $0 < \alpha < 1$. In view of Carroll (1976) and the treatment given in Isogai (1987), we take $h_n = n^{-r}$ for $\frac{1}{5} < r < 1$. We propose the confidence interval $I_n = \left[\widehat{f}_{n,h_n}(x_0) - d,\ \widehat{f}_{n,h_n}(x_0) + d\right]$ where x_0 is held fixed and we assume that $f(x_0) > 0$. Now for large n, one can prove that

$$\mathcal{P}\left(\widehat{f}_{n,h_n}(x_0) - d < f(x_0) < \widehat{f}_{n,h_n}(x_0) + d\right) \approx 1 - \alpha \tag{2}$$

if $n \geq n^*$ where

$$n^* = \left\{z^2_{\alpha/2} B f(x_0)/d^2\right\}^{\frac{1}{1-r}}, \tag{3}$$

$B = (1 + r)^{-1}\int_{-\infty}^{\infty} K^2(y)dy$ and $z_{\alpha/2}$ is the upper $50\alpha\%$ of the standard normal distribution.

2. Sequential Procedures

A brief description of the sequential procedures considered in this together with their stopping rules are given below. For a comprehensive details of these procedures are given in Ghosh, Mukhopadhyay and Sen (1997) and also Mukhopadhyay and Solanky (1994).

Purely Sequential Procedure: We refer to Carroll (1976) and Mukhopadhyay (1997) for this procedure. First take an initial sample of size m and then continue sampling until the sample size,

$$n \geq \left[\left\{z^2_{\alpha/2} B\left(\widehat{f}_{n,h_n}(x_0) + n^{-1}\right)\right\}/d^2\right]^{\frac{1}{1-r}} \text{ for the first time,} \tag{4}$$

and compute the estimator I_N.

Accelerated Sequential Procedure: This has three steps. First take an initial sample of size m, secondly take one sample at time until the sample size, n satisfy the following:

$$n \geq \rho \left[\left\{ z^2_{\alpha/2} B \left(\widehat{f}_{n,h_n}(x_0) + n^{-1} \right) \right\} / d^2 \right]^{\frac{1}{1-r}} \text{ for the first time,} \tag{5}$$

where $0 < \rho < 1$. Next estimate the final sample size N by $N = \langle n/\rho \rangle + 1$ and in the final step, we take a random sample of size $(N - n)$, and compute the estimator I_N.

Two-Stage Procedure: Stopping rule for a two-stage procedure from de Silva, Roy and Mukhopadhyay (2000) is given by

$$N = \max \left\{ m, \left\langle \left\{ z^2_{\alpha/2} B \widehat{f}_{m,h_m}(x_0)/d^2 \right\}^{\frac{1}{1-r}} \right\rangle + 1 \right\}. \tag{6}$$

For stage one, take a sample of size m and estimate final sample sample N using the above stopping rule. Next take a sample of size $N - m$ in the second stage and compute the interval I_N using all N observations.

Modified Two-Stage Procedure: This is a similar two-stage procedure, but the initial sample size m is determined using the following:

$$m = \max \left\{ 2, \left\langle \left(Z^2_{\alpha/2} B/d^2 \right)^{1/\{(1-r)(1+\gamma)\}} \right\rangle + 1 \right\}. \tag{7}$$

Three-Stage Procedure: This procedure starts with an initial fixed sample m whereas the sample sizes for the second and third stages are respectively $(N_1 - m)$ and $(N - N_1)$ where

$$N_1 = \max \left\{ m, \left\langle \rho \left\{ z^2_{\alpha/2} B \widehat{f}_{m,h_m}(x_0)/d^2 \right\}^{\frac{1}{1-r}} \right\rangle + 1 \right\}, \tag{8}$$

$$N = \max \left\{ N_1, \left\langle \left\{ z^2_{\alpha/2} B \widehat{f}_{N_1,h_{N_1}}(x_0)/d^2 \right\}^{\frac{1}{1-r}} \right\rangle + 1 \right\} \tag{9}$$

and $0 < \rho < 1$. Finally, we compute the interval I_N.

3. Simulation

A simulation study was conducted to compare the behaviors of the above sequential procedures for moderate sample sizes. The selection of a bandwidth is an important aspect for kernel density estimation. For given d, the sample size required to achieve the set confidence level depends on the choice of a bandwidth. We obtained the suitable bandwidth for each simulation by minimizing the bootstrap estimator of the mean integrated squared error. See Shao and Tu (1995) and de Silva, Roy and Mukhopadhyay (2000) for more detail. Due to the limitation of the space, numerical results based on 10,000 simulated replications and the bibliography will be provide during the presentation of the paper.

RÉSUMÉ

Dans cette présentation, l'efficacité d'exécution des procédures de traitement purement séquentielles, séquentielles accélérées, à deux niveaux, à deux niveaux modifiées et à trois niveaux d'un estimateur d'intervalle de confiance à largeur fixe d'une fonction de densité inconnue sont comparées. Une étude de simulation utilisant le noyau normal unité pour analyser ces rendements dans le cas d'échantillons moyens et petits fut entreprise. Une détermination optimale de largeurs de bande au moyen du "bootstrap" sera discutée au cours de la présentation