Data sharpening by estimating the derivative of log density

Byeong U. Park
Department of Statistics, Seoul National University
Shillim-Dong, Kwanak-Ku
Seoul, Korea
bupark@stats.snu.ac.kr

Kee-Hoon Kang
Department of Statistics, Hankuk University of Foreign Studies
Wangsan-Ri, Mohyun-Myeon
Yongin, Korea
khkang@hufs.ac.kr

Seok-Oh Jeong
Department of Statistics, Seoul National University
Shillim-Dong, Kwanak-Ku
Seoul, Korea
seokoh@stats.snu.ac.kr

1. Introduction

We suggest a simple approach to data sharpening for density estimation using local likelihood estimator of the derivative of log density. Data sharpening methods, which were recently introduced by Choi and Hall (1999) in density estimation, draw attention as a way of enhancing some properties of relatively conventional statistical procedures. One of the most prominent results of these methods is bias reduction. The idea is to make the data more clustered near peaks of the density and further apart near troughs, and then to compute the estimator in the usual way but with the sharpened data.

In this paper we introduce a new way of determining data perturbation amount. It is based on the local likelihood principle (Loader 1996, for example) to get a pilot estimator of $(\log f)' = f'/f$. Furthermore, we consider a modification, denoted by $\tilde{f}(x)$, of the ideal form to take into account the boundary effects. We show that plugging the pilot estimator into the modified $\tilde{f}(x)$ affords $O(h^4)$ bias in the interior, and $O(h^2)$ bias at the boundary of the density support.

2. Estimator and results

Let $X_1, \ldots, X_n$ be observations lying in a subset $\mathcal{X}$ of $\mathcal{R}$ from a distribution with an unknown density $f(x)$. The conventional kernel density estimator is $\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h \left(X_i - x\right)$, where $K_h(z) = h^{-1}K(h^{-1}z)$, K is a kernel function and h a bandwidth. It is well known that the bias and variance of $\hat{f}$ are $O(h^2)$ and $O((nh)^{-1})$, respectively. Samiuddin and El-Sayyad (1990) showed that $O(h^4)$ bias would be achieved by $n^{-1} \sum_{i=1}^{n} K_h((x - X_i - \frac{1}{2}h^2\kappa_2 f'(X_i)/f(X_i))$.

Let the support of the density f be $[a, b]$ with $-\infty$ and ∞ being allowed for the values of a and b, respectively. Suppose that we use a kernel with compact support, say $[-\tau, \tau]$. Define $\rho(t_1, t_2) = -\{K(t_2) - K(t_1)\}^{-1} \int_{t_1}^{t_2} vK(v)\, dv$ if t_1 or $t_2 \in (-\tau, \tau)$, and $\kappa_2/2$ if $t_1, t_2 \in (-\tau, \tau)^c$. Write $\gamma = (\log f)' = f'/f$. The target function $\tilde{f}(x)$, which is modified for boundary property, is then given by

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h} + h\, \gamma(X_i)\, \rho\Big(\frac{a - x}{h}, \frac{b - x}{h}\Big)\Big) \Big/ \int_a^b K\Big(\frac{y - x}{h}\Big) dy.$$

Then, the proposed estimator $\hat{f}$ uses local likelihood to estimate $\gamma(X_i)$. Under some regularity conditions on γ, K and f , we can show

$$\hat{f}(x) - \tilde{f}(x) = b(x)h^4 + D_n(x) + o_p(h^4 + n^{-1/2}h^{-1/2})$$

where $D_n(x)$ has zero mean and $O(n^{-1}h^{-1})$ variance.

3. Remarks

Several data sharpening methods in density estimation by plugging-in unknown quantities have similar performance in the interior but different in boundary regions of the support. The proposed estimator does not suffer from edge effects and it adapts automatically to estimation at the boundaries. This property at the boundaries is shared with the local linear smoothers in regression setting.

REFERENCES

Choi, E. and Hall, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika* **86**, 941–947.

Loader, C.R. (1996). Local likelihood density estimation. *The Annals of Statistics* **24**, 1602–1618.

Samiuddin, M. and el-Sayyad, G.M. (1990). On nonparametric kernel density estimates. *Biometrika* **77**, 865–874.

RESUME

Dans cet article, nous considérons le noyau type estimateur de densité.