Report 99-017

ON THE MODE
OF AN UNKNOWN
PROBABILITY DISTRIBUTION

Novak S.Y.

ISSN: 1389-2355

ON THE MODE OF THE UNKNOWN PROBABILITY DISTRIBUTION

S. Y. NOVAK

EURANDOM
PO Box 513, Eindhoven 5600 MB, The Netherlands.

Abstract

The paper is devoted to the nonparametric estimation of the mode of an unknown probability distribution.

We formulate and solve the problem of choosing optimal parameters of the kernel estimator of the mode.

Keywords: mode of probability distribution, nonparametric estimation.

1 Introduction.

Suppose that the random variable (r.v.) X has an absolutely continuous distribution with density f. Given an i.i.d. sample $X_1, \dots, X_n$ from X, we want to estimate the mode

$$\theta = \inf\{x : f(x) = \sup_t f(t)\}.$$

A natural estimator of θ is the sample mode

$$\theta_n = \inf\{x : f_n(x) = \sup_t f_n(t)\}, \tag{1}$$

where f_n is the kernel density estimator

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} f_{x+h\gamma}(X_i) \tag{2}$$

with the smoothing parameter h and the kernel f_γ. Hereinafter the symbol f_ξ denotes the density of the distribution of a random variable ξ, and $\mathbf{D}$ denotes a variance.

Given the estimator (1), the question is which kernel f_γ to choose? We answer this question in theorem 1 below.

Conditions for consistency and asymptotic normality of the estimator (1) were found in [1–4], [8]. It was pointed out in [3] that, under appropriate assumptions on f and f_γ,

$$\mathbf{E}(\theta_n - \theta)^2 \sim \left[h^2 m_2 f^3(\theta)/2f''(\theta)\right]^2 + \nu f(\theta)/nh^3[f''(\theta)]^2 \tag{3}$$

as $h \equiv h(n) \to 0$, $nh^3 \to \infty$. Here

$$\nu \equiv \nu(\gamma) = \int f_\gamma'^2 , \; m_i = \mathbf{E}\gamma^i \equiv \int x^i f_\gamma(x)dx .$$

Relation (3) may be obtained on the basis of the following observation valid for smooth functions f and f_γ (see [1]):

$$0 \equiv f_n'(\theta_n) = f_n'(\theta) + (\theta_n - \theta) f_n''(\theta + \tau(\theta_n - \theta)) \qquad (0 \le \tau \le 1) .$$

This implies that

$$\theta_n - \theta = -f_n'(\theta) / f_n''(\theta + \tau(\theta_n - \theta)) . \tag{4}$$

Moreover, it is not hard to verify that

$$\mathbf{E} f_n'(x) = \mathbf{E} f'(x + h\gamma) \approx f'(x) + h^2 f'''(x) m_2/2 \; , \; \mathbf{D} f_n'(x) \sim \nu f(x)/nh^3$$

if the functions f and f_γ are "smooth" enough and h is "small" (see [12] for more information on ratios of type (4)).

Relation (3) entails that the mean squared error

$$\mathrm{MSE}\, \theta_n \equiv \mathbf{E}\,(\theta_n - \theta)^2 = O(n^{-4/7}) \tag{$*$}$$

if $h(n) \sim cn^{-1/7}$. A better rate of MSE may be achieved if f_γ is the density of a generalised distribution (unit measure) on $\mathbf{R}$. Note that the function f_n may take negative values in that case.

We deal with the problem of the optimal choice of the bandwidth h and the kernel f_γ .

Eddy [3] suggested to chose $h = h(n)$ in such a way that the 2-nd term on the right-hand side of (3) dominates the 1-st one, and then to look for a kernel f_ξ such that

$$\nu(\xi) = \min_\gamma \nu(\gamma) . \tag{5}$$

It is shown in [3,5] that Epanechnikov's kernel

$$f_\xi(x) = \frac{3}{4}(1 - x^2)\mathbf{1}\{\|x\| \le 1\}$$

is the solution to problem (5) in the class of absolutely continuous distributions.

The assumption that the 1-st term on the right-hand side of (3) is negligible with respect to the 2-nd one means that $h \ll n^{-1/7}$. In contrast to (*), this implies that $\mathbf{E}(\theta_n - \theta)^2 \gg n^{-4/7}$. Hence, Eddy's approach can hardly be regarded as a natural one.

In this article we suggest a different approach to the problem of choosing the optimal values of the parameters h and f_γ .

2 Optimal kernels.

Denote by h_* the value of h that minimises the right-hand side of (3). It is easy to see that

$$nh_*^7 = 3\nu f(\theta)[m_2 f'''(\theta)]^{-2} . \tag{6}$$

With $h = h_*$, the right-hand side of (3) becomes

$$\frac{7}{12}\,[R(\gamma)]^{2/7}\left[\frac{3f(\theta)}{n}\right]^{4/7}\frac{|f'''(\theta)|^{6/7}}{[f''(\theta)]^2}\,, \tag{7}$$

where

$$R(\gamma) = \nu^2(\mathbf{E}\gamma^2)^3\,. \tag{8}$$

Inspired by (3) and (7), we formulate the problem of minimising the functional $R(\gamma)$ over the class $\mathcal{D}$ of distributions with absolutely continuous densities.

A similar problem was treated in [5], but instead of $\mathcal{D}$, the authors of [5] considered the class of nonnegative kernels g (vanishing outside some symmetric interval $[-T;T]$) such that the derivative g' changes its sign only once.

Theorem 1. *Denote*

$$f_\xi(x) = \frac{15}{16}(1-x^2)^2\mathbf{1}\{|x|\le 1\}\,. \tag{9}$$

Then

$$\min_{\gamma\in\mathcal{D}} R(\gamma) = R(\xi)\,.$$

Remark. Note that the functional (8) is invariant under a replacement of γ by $c\gamma$. Thus, one can say that there is a family $\{f_{t\xi}, t>0\}$ of optimal kernels.

Now we compare the values of $R(\gamma)$ for some distributions:

No		$f_\gamma(x)$	supp f_γ	m_2	$\nu(\gamma)$	$R(\gamma)$
1	Normal	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	$(-\infty;\infty)$	1	$\frac{1}{4\sqrt{\pi}}$	$\frac{1}{16\pi}\approx 0.0199$
2	Epanechnikov's	$\frac{3}{4}(1-x^2)$	$[-1;1]$	1/5	3/2	0.018
3	Optimal	$\frac{15}{16}(1-x^2)^2$	$[-1;1]$	1/7	$\frac{15}{17}$	$225/7^5\approx 0.013$
4		$\frac{25}{32}(1-x^2)(1-x^2/5)$	$[-1;1]$	4/21	$\frac{85}{56}$	$\frac{289}{7^5}\cdot\frac{25}{27}\approx 0.016$
5		$1-\lvert x\rvert$	$[-1;1]$	1/6	2	$\frac{1}{54}\approx 0.0185$

Suppose that (3) holds. If the smoothing parameter h is chosen according to (6) and the kernel according to (9) then

$$\mathbf{E}(\theta_n-\theta)^2 \sim \frac{7}{12}n^{-4/7}\,[f''(\theta)]^{-2}\left[3^4\cdot 5^2\cdot 7^{-5}f^2(\theta)|f'''(\theta)|^3\right]^{2/7}$$

as $n \to \infty$. Under additional assumptions, this relation holds if one replaces θ and $f^{(3)}$ in (6) by their consistent estimators.

For every even k, denote

$$R_k(\gamma) = \nu^k m_k^3 \,. \tag{10}$$

Consider the problem of minimising the functional $R_k(\gamma)$ over the class $\mathcal{D}$.

The interest in this topic is facilitated by a tendency to use the so-called improved nonnegative density estimators (INDE) instead of kernel density estimators (2) with generalised kernels f_γ (for more information on INDE see [1, 7, 9, 13, 15, 16]). For every known INDE $\hat{f}_n$, the decay rate of the bias and of the MSE is the same as those of the kernel density estimator:

$$\mathbf{E}\hat{f}_n(x) - f(x) \to c_0(x)h^k m_k\,, \quad \mathbf{E}\left(\hat{f}_n(x) - f(x)\right)^2 = O\left((nh)^{-1} + h^{2k}\right)$$

as $n \to \infty$. If the distribution of X has a heavy tail then the MSE of INDE may be asymptotically smaller than the MSE of the classical kernel density estimator (see [13]).

Since the bias of $f_n'(x)$ decays like $O(h^k)$, we expect that $\mathbf{E}\hat{f}_n'(x) - f(x) \to c_1(x)h^k m_k$. The variance of any known estimator of $f'(x)$ decays (under appropriate assumptions) like $c_2(x)\nu/nh^3$. Therefore, arguments similar to those yielding (6) and (7) show that the optimal choice of h is $c_3(x)(\nu/nm_k^2)^{1/(3+2k)}$. In that case

$$\mathbf{E}\,(\hat{f}_n'(x) - f'(x))^2 \sim c(x) n^{-2k/(3+2k)}[R_k(\gamma)]^{2/(3+2k)}\,.$$

The problem is to find a density that minimises $R_k(\gamma)$.

Theorem 1*. *Denote*

$$f_\eta(x) = \frac{3(k+3)}{4k(k+2)}\left[k - (k+2)x^2 + 2x^{k+2}\right]\mathbf{1}\{|x| \le 1\}\,, \tag{11}$$

where $\eta \equiv \eta_k$. *Then for any even* $k \in \mathbf{N}$

$$\min_{\gamma \in \mathcal{D}} R_k(\gamma) = R_k(\eta)\,. \tag{12}$$

The assertion of Theorem 1 is a consequence of (12) with $k = 2$. If $k = 4$ then an optimal kernel is

$$f_{\eta_4}(x) = \frac{7}{16}(1 - x^2)^2(2 + x^2)\mathbf{1}\{|x| \le 1\}\,. \tag{13}$$

Eddy [3] considered the problem of minimising the functional (5) in a certain class $\mathcal{D}_4$ of generalised distributions satisfying the following conditions

$$\mathbf{E}\gamma^i \equiv \int x^i f_\gamma(x)dx = 0 \quad \text{for} \quad 1 \le i \le 3\,, \quad \mathbf{E}\gamma^4 \ne 0\,.$$

It is shown in [3,6] that $\min_{\gamma \in \mathcal{D}_4} \nu(\gamma) = \nu(\xi_*)$, where

$$f_{\xi_*} = \frac{15}{32}(1 - x^2)(3 - 7x^2)\mathbf{1}\{|x| \le 1\}\,, \tag{14}$$

and that $\min\limits_{\gamma\in\mathcal{D}_4} R_4(\gamma)=\nu(\eta_*)$, where

$$f_{\eta_*}=\frac{105}{64}(1-x^2)^2(1-3x^2)\mathbf{1}\{|x|\leq 1\}. \tag{15}$$

The biases of the estimators f_n and f_n' with kernels (14) and (15) decays like $O(h^4)$ as $h\to 0$.

*Proof of Theorem 1**. Note that $\mathbf{E}\eta^k=3/[(k+1)(2k+3)]$. Since the functional (10) is invariant under a replacement of γ by $c\gamma$, we may suppose that

$$\mathbf{E}\gamma^k=3/[(k+1)(2k+3)].$$

Denote by $\mathcal{D}_*$ the set of densities from $\mathcal{D}$ which obey the conditions

$$(i)\ \int g(x)dx=1\,,\qquad (ii)\ \int x^k g(x)dx=3/(k+1)(2k+3)\,.$$

Let us show that

$$r(g)\equiv\int g'^2\geq\int f'^2=r(f_\eta) \tag{16}$$

for every function $g\in\mathcal{D}_*$.

Put $h=g-f_\eta$. Observe that

$$f_\eta'(x)=\frac{3(k+3)}{2k}x(x^k-1)\,,\quad f_\eta''(x)=\frac{3(k+3)}{2k}\left((k+1)x^k-1\right).$$

Then, since $f_\eta'(\pm1)=0$, we have

$$\int g'^2\ \geq\ \int f_\eta'^2+2\int_{-1}^1 f_\eta'h'\ =\ \int f_\eta'^2-2\int_{-1}^1 f_\eta''h.$$

Because of (i) and (ii),

$$\int h(x)dx\ =\ \int x^k h(x)dx\ =\ 0.$$

Therefore,

$$\int h(x)\left((k+1)x^k-1\right)dx=0,$$

which implies that

$$\int g'^2\ \geq\ \int f_\eta'^2\ +\ \frac{3(k+3)}{2k}\int_{\mathbf{R}\setminus[-1;1]}((k+1)x^k-1)h(x)dx. \tag{17}$$

The function g is nonnegative. Hence $h(x)\geq 0$ if $x\notin[-1;1]$, and relation (16) immediately follows from (17). □

The approach suggested above is close to that of [2], ch. 5, where the problem of minimising the functional $(\mathbf{E}f_\gamma(\gamma))^k\mathbf{E}\gamma^k$ was considered. The optimal kernel (11) might also be found using tools of the calculus of variations [6] if we restricted ourselves to a class of absolutely continuous functions vanishing outside finite symmetric intervals and noticed that assumptions (i), (ii) for those functions could be rewritten in the form

$$(i')\ \int xg'(x)dx=-1\,,\quad (ii')\ \int x^{k+1}g'(x)dx=-3/(2k+3).$$

References

[1] Abramson I.S. On bandwidth variation in kernel estimatates — a squareroot law. — Ann. Statist. 1982. v. 10. No 4. p. 1217–1223.

[2] Devroye L. and Györfi L. *Nonparametric density estimation: the L_1 view.* New York: Wiley, 1985.

[3] Eddy W.F. Optimum kernel estimators of the mode. — Ann. Statist., 1980, v. 8, No 4, p. 870–882.

[4] Eddy W.F. The asymptotic distribution of kernel estimators of the mode. — Z. Wahr. verw. Geb., 1982, v. 59, p. 279–290.

[5] Gasser T., Müller H.-G., Mammitzsch V. Kernels for nonparametric curve estimation. — J. R. Statist Soc. B. 1985, v. 47, No 2, p. 238–252.

[6] Gelfand I.M. and Fomin S.V. *Calculus of variations.* Moscow: Fizmatgiz, 1961, 228 pp.

[7] Hall P. On global properties of variable bandwidth density estimators. — Ann. Statist., 1992, v. 20, No 2, p. 762–778.

[8] Konakov V.D. On asymptotic normality of the mode of multivariate distributions. — Theory Probab. Appl. 1973. V. 18. No 4. P. 836–842.

[9] Koshkin G.M. Improved nonnegative kernel density estimator. — Theory Probab. Appl. 1988. v. 33. No 4. p. 817–822.

[10] Mammitzsch V. Asymptotically optimal kernels for average derivative estimation. — Prace Naukove Academii Economiczneij imiena O. Langego we Wroclawiu, 1990, v. 513, p. 278–289.

[11] Müller H.-G. *Nonparametric regression analysis of longitudinal data.* Berlin: Springer, 1988, 199 pp.

[12] Novak S.Y. On the distribution of the ratio of sums of random variables. — Theory Probab. Appl., 1996, v. 41, No 3, 479–503.

[13] Novak S.Y. Generalised kernel density estimator. — Research Report No 98/10 SMS/CSSM, University of Sussex.

[14] Parzen E. On estimation of a probability density function and mode. — Ann. Math. Statist., 1962, v. 33, p. 1065–1076.

[15] Terrel G.R., Scott D.W. On improving convergence rates for nonnegative kernel density estimator. — Ann. Statist. 1980. v. 8. P. 1160–1163.

[16] Terrel G.R., Scott D.W. Variable kernel density estimation. — Ann. Statist., 1992, v. 20, No 3, p. 1236–1265.

[17] Woodroofe, M. On choosing a Delta-sequence. — Ann. Math. Statist., 1970, v. 41, No 5, 1665–1671.