Journal of Statistical Planning and Inference 58 (1997) 333–348

journal of statistical planning and inference

Optimal smoothing in adaptive location estimation[1]

Enno Mammen[a], Byeong U. Park[b,*]

[a] *Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany*

[b] *Department of Computer Science and Statistics, Seoul National University, Seoul 151-742, S.Korea*

Received 26 June 1995; revised 20 December 1995

Abstract

In this paper, we consider higher order performance of kernel based adaptive location estimates. We show how much one loses in efficiency without knowing the underlying translation density, and derive the optimal order of the bandwidths involved in kernel estimation of the efficient score function. The optimal order is obtained by minimizing the loss of efficiency in terms of estimating the location parameter. The main lesson here is that the optimal order of the bandwidths are different from those for optimal estimation of the score function. This implies that optimal estimation of the score function does not lead to second order optimal location estimation.

AMS Classification: 62F35; 62G05; 62G20

Keywords: Adaptive estimation; Bandwith; Kernel function; Semi-parametric location model

1. Introduction

In this paper we consider the problem of optimal bandwidth choice in kernel-based adaptive location estimation. The observations X_i's are generated by the model $X_i = \theta + \varepsilon_i$ $(1 \leqslant i \leqslant n)$ in which θ is an unknown parameter and the ε_i's are independent and identically distributed errors with an unknown common symmetric density f. Adaptivity means here that without knowing f one estimates θ asymptotically as well (to first order) as one could do when knowing f.

When f is known, the efficient maximum likelihood estimator $\hat{\theta}_n^{\mathrm{ML}}$ fulfills

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}_n^{\mathrm{ML}}) = 0.$$

* Corresponding author.

[1] This research was supported by KOSEF, the Korean–German Cooperative Science Program 1994, and by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse", Humboldt - Universität zu Berlin.

PII S0378-3758(96)00085-7

Here ψ denotes the efficient score function of f, i.e. $\psi(x) = -f'/f(x)$. A one-step approximation of $\hat{\theta}_n^{\mathrm{ML}}$ is given by

$$\hat{\theta}_n^{\mathrm{APPML}} = \bar{\theta}_n + n^{-1}\sum_{i=1}^{n}\psi(\hat{e}_i)\left[n^{-1}\sum_{i=1}^{n}\psi'(\hat{e}_i)\right]^{-1}, \tag{1.1}$$

where $\bar{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ and $\hat{e}_i = X_i - \bar{\theta}_n$ are the residuals $(1 \leqslant i \leqslant n)$. In most applications the term $[n^{-1}\sum_{i=1}^{n}\psi'(\hat{e}_i)]^{-1}$ in the definition of $\hat{\theta}_n^{\mathrm{APPML}}$ is replaced by I^{-1}, where $I = \int \psi^2 f$ is the Fisher information.

When f is unknown, as in the present case, one can get an adaptive estimate by plugging a kernel estimate $\hat{\psi}$ of ψ into the formula (1.1). More precisely, $\hat{\psi}$ could be some modification of $-\hat{f}'_g/\hat{f}_h$, where $\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - (X_i - \bar{\theta}_n))$ and $K_h(x) = h^{-1}K(x/h)$. Here K is the kernel function (usually a symmetric probability density function), and h and g are the bandwidths that control the smoothing amount of the function estimates. In most literature, the choice $h = g$ is taken and there has been no discussion on the issue of what is the best rate of convergence of $h(= g)$. Recent works include Stone (1975), Bickel (1982), Schick (1987), and Hsieh and Manski (1987), among others. The first three papers cited above focus on showing the first order equivalence, to $\hat{\theta}_n^{\mathrm{ML}}$ (adaptivity), of the kernel based method with deterministic choices of $h(= g)$, and the last one shows that the empirical performance of adaptive estimates is highly sensitive to the choice of $h(= g)$.

One of the main strength of this paper is that we derive the optimal orders of h and g in terms of estimation of θ, not ψ. It has been believed that the estimation of the efficient score function ψ is the crucial step toward adaptiveness of the location estimate. This has motivated many people to consider bandwidth choice in terms of estimating ψ. See, for example, Park (1993). However, through demonstrating higher order performance of the adaptive estimate, our results show that optimal adaptive estimation of θ yields different choices of h and g (even differing in the rate) from those for optimal estimation of the score function ψ. This illustrates the fact that optimal estimation of ψ does not lead to optimal adaptive location estimation. Since choice of the score function ψ could be interpreted as choice of the model, our results may be interpreted as an example where the following approach fails: first choose an "optimal" model, and then use an "optimal" procedure in this fitted model. The optimality of the procedure may get lost by the stochastic nature of the choice of the model.

Another important finding in this work is that the optimal rates of convergence of h and g are of different order. It turns out that, with h and g of the same order, the minimal relative loss of efficiency (in comparison with $\hat{\theta}_n^{\mathrm{ML}}$) is of order $n^{-2/5}$, but with h and g of different order, one can do a lot better: one can achieve $n^{-4/7}$. These rates may be compared with the approach based on the optimal estimation of ψ. This approach leads to much larger relative losses of efficiency. The corresponding rates are $n^{-2/7}$ (for the case $g = h$ and for the case that h and g are allowed to be different).

Other related works in adaptive estimation include Faraway (1992) and Jin (1992), among others. The cited papers consider spline based estimation of the score function, instead of kernel based method, and demonstrate empirical selection rules of the smoothing parameter.

Our assumptions are collected in Section 2. In particular, for technical reasons we assume that the density and its higher order derivatives have exponential tails. This allows uniform expansions of estimates of the score function. Densities with heavy tails (e.g. Cauchy density) are excluded by our assumptions. Section 3 contains some preliminary calculations for optimal bandwidth choices. There we make the theoretical choice $\bar{\theta}_n = \theta$. Section 4 shows that the results of Section 3 remain valid with $\sqrt{n}$-consistent estimates $\bar{\theta}_n$. In Section 5 we confirm the theoretical findings by a simulation study. Section 6 contains the proofs of our results.

2. Assumptions

(A.1) $X_1,\dots,X_n$ are independent and identically distributed with density $f(x-\theta)$. The density f is symmetric: $f(x) = f(-x)$.

(A.2) There exist positive constants C_0, C_1, C_2, C_3 with

$$f(x) \leqslant C_0 \exp(-C_1|x|),$$

$$|f^{(k)}(x)| \leqslant C_3(1+|x|^{C_2})\, f(x)$$

for $1 \leqslant k \leqslant 5$ and for all real x.

(A.3) $\bar{\theta}_n$ is a $\sqrt{n}$ consistent estimate of θ, i.e. $\bar{\theta}_n - \theta = O_P(n^{-1/2})$.

(A.4) K is a symmetric density function with compact support and it is three times continuously differentiable.

We use the convention that $C, C', C'', \dots$ denote universal constants (with different meanings at different places).

3. Optimal choice of smoothing parameters: Some preliminary calculations

Let f_X denote the density of the X_i's. Define $\hat{f}_{X,h,i}(x)$, the kernel estimate of $f_X(x)$ with leaving out the observation X_i, by

$$\hat{f}_{X,h,i}(x) = (n-1)^{-1} \sum_{j \neq i} K_h(x - X_j). \tag{3.1}$$

Here and after, we use the notation $K_h(x) = h^{-1}K(x/h)$, $K_h'(x) = h^{-2}K'(x/h)$, and so on. We consider the following estimate $\hat{\theta}_n$ of θ:

$$\hat{\theta}_n = \bar{\theta}_n + I_{h,g}(\bar{\theta}_n)^{-1} \Delta_{h,g}(\bar{\theta}_n),$$

where

$$\Delta_{h,g}(\theta) = -n^{-1}\sum_{i=1}^{n}\frac{\hat{f}'_{X,g,i}(X_i) - \hat{f}'_{X,g,i}(2\theta - X_i)}{s_n + \hat{f}_{X,h,i}(X_i) + \hat{f}_{X,h,i}(2\theta - X_i)},$$

$$I_{h,g}(\theta) = n^{-1}\sum_{i=1}^{n}[\hat{f}'_{X,g,i}(X_i) - \hat{f}'_{X,g,i}(2\theta - X_i)][\hat{f}'_{X,h,i}(X_i) - \hat{f}'_{X,h,i}(2\theta - X_i)]$$

$$\times[s_n + \hat{f}_{X,h,i}(X_i) + \hat{f}_{X,h,i}(2\theta - X_i)]^{-2}.$$

Here, s_n is a deterministic sequence converging to zero (with a rate discussed below). We write $\hat{\theta}_n(\bar{\theta}_n)$ for $\hat{\theta}_n$ to indicate the dependence of $\hat{\theta}_n$ on $\bar{\theta}_n$.

Our estimate $\hat{\theta}_n$ is nearly of the type (1.1) with $\psi(x)$ replaced by the estimate $\hat{\psi}(x) = -[\hat{f}'_g(x) - \hat{f}'_g(-x)][\hat{f}_h(x) + \hat{f}_h(-x)]^{-1}$ where $\hat{f}_h(x)$ is the kernel estimate of f, the density of ε_i's, defined by $\hat{f}_h(x) = n^{-1}\sum_{j=1}^{n} K_h(x - (X_j - \bar{\theta}_n))$. Note that this definition of $\hat{\psi}(x)$ preserves the anti-symmetry of $\psi(x)$. Also note that $\hat{f}_h(x - \bar{\theta}_n) = \hat{f}_{X,h}(x)$ and $\hat{f}_h(\bar{\theta}_n - x) = \hat{f}_{X,h}(2\bar{\theta}_n - x)$, where $\hat{f}_{X,h}(x)$ denotes the kernel estimate of $f_X(x)$ defined in the same way as (3.1) but with all observations.

With this estimate of $\psi(x)$, three modifications have been made in the definition of $\hat{\theta}_n$. First, the constant s_n is introduced in the denominator of $\hat{\psi}$ to avoid its erratic behaviour when the denominator has a very small value. The second modification concerns the definition of the kernel estimates. Note that in their definition (see (3.1)) terms corresponding to (i=j) are omitted in the summations. This modification is crucial. Otherwise we would get nonstochastic terms of the form $K(0)$ or $K'(0)$ in the definition of $I_{h,g}(\bar{\theta}_n)$ and $\Delta_{h,g}(\bar{\theta}_n)$. Finally, $\psi'(x)$ is estimated by $[\hat{f}'_g(x) - \hat{f}'_g(-x)][\hat{f}'_h(x) - \hat{f}'_h(-x)][s_n + \hat{f}_h(x) + \hat{f}_h(-x)]^{-2}$. This differs from $\hat{\psi}'(x)$ by a summand

$$-[\hat{f}''_g(x) + \hat{f}''_g(-x)][s_n + \hat{f}_h(x) + \hat{f}_h(-x)]^{-1}.$$

This modification is made only for simplification. All our calculations and conclusions would go through by inclusion of this additional term.

Discussion of the asymptotic performance of $\hat{\theta}_n$ is complicated by two facts.

(1) The summands of $\Delta_{h,g}$ and $I_{h,g}$ have random denominators.

(2) The estimate $\hat{\theta}_n$ depends on the preliminary estimate $\bar{\theta}_n$.

We proceed as follows: First, in the definition of $\hat{\theta}_n$, we replace the preliminary estimate $\bar{\theta}_n$ by the true underlying location parameter θ (i.e. we study $\hat{\theta}_n(\theta)$). Furthermore, for a stochastic approximation $\hat{\theta}_n^{\mathrm{APPR}}$ of $\hat{\theta}_n(\theta)$, we discuss appropriate choices of g, h and s_n. Then, in the next section, we show that, for these choices of g, h and s_n, the approximation $\hat{\theta}_n^{\mathrm{APPR}}$ is accurate enough for $\hat{\theta}_n(\bar{\theta}_n)$.

We put

$$\hat{\theta}_n^{\mathrm{APPR}} = \theta + \left[\int \frac{4f'_g f'_h}{[s_n + 2f_h]^2} f\right]^{-1}\left[-n^{-2}\Sigma^{\neq}_{i,j}\frac{K'_g(X_i - X_j) - K'_g(2\theta - X_i - X_j)}{s_n + 2f_h(X_i - \theta)}\right.$$

$$+n^{-2}I^{-1}\Sigma^{\neq}_{i,j}\left(\frac{4f'_g(X_i-\theta)f'_h(X_i-\theta)}{(s_n+2f_h(X_i-\theta))^2}-\int\frac{4f'_gf'_h}{[s_n+2f_h]^2}f\right)$$

$$\times\frac{2f'_g(X_j-\theta)}{s_n+2f_h(X_j-\theta)}+n^{-3}\Sigma^{\neq}_{i,j,k}[K'_g(X_i-X_j)-K'_g(2\theta-X_i-X_j)]$$

$$\left.\times[K_h(X_i-X_k)+K_h(2\theta-X_i-X_k)-2f_h(X_i-\theta)][s_n+2f_h(X_i-\theta)]^{-2}\right].$$

Here $f_h(x)=\int K_h(x-y)f(y)dy$ and $\Sigma^{\neq}$ denotes summation over pairwise different indices.

The asymptotics of $\hat{\theta}_n^{\text{APPR}}$ is described in the following theorem.

Theorem 1. *Suppose* (A.1), (A.2), (A.4) *and* $s_n\to 0$, $h\to 0$, $g\to 0$, $n^{-1}g^{-3}\to 0$, $n^{-1}h^{-1}\to 0$. *Then we have that*

$$\sqrt{n}(\hat{\theta}_n^{\text{APPR}}-\theta)\to N(0,I^{-1})\qquad (\textit{in distribution}).$$

Furthermore, for the first two moments of $\hat{\theta}_n^{\text{APPR}}$ *we get*

$$E(\hat{\theta}_n^{\text{APPR}})=\theta$$

$$\begin{aligned}n\text{Var}(\hat{\theta}_n^{\text{APPR}})={}&I^{-1}+2n^{-1}g^{-3}I^{-2}\int\frac{f^2}{(s_n+2f)^2}\int(K')^2(1+\text{o}(1))\\&+8n^{-1}h^{-1}I^{-2}\int\frac{(f')^2f^2}{(s_n+2f)^4}\int K^2(1+\text{o}(1))\\&+\frac{1}{4}g^4d_K^2I^{-2}\left[\int\frac{(f''')^2}{f}-\left(\int\frac{f'f'''}{f}\right)^2I^{-1}\right]\\&+h^2d_KI^{-2}\int\frac{(f'')^2}{f}+\text{o}(h^2+g^4)\\&+I^{-2}\left[I-\int\frac{4(f')^2}{(s_n+2f)^2}f\right](1+\text{o}(1)).\end{aligned}\qquad(3.2)$$

Here $d_K=\int t^2K(t)\text{d}t$.

For discussion of Theorem 1, let us first consider the case that f has compact support. In this case, $\int f^2(s_n+2f)^{-2}=\text{O}(1)$, and for the optimal choice of g we get $g\sim n^{-1/7}$. Under (A.2) we also have $\int(f')^2f^2(s_n+2f)^{-4}=\text{O}(1)$, therefore with $h\sim n^{-1/3}$ (for instance) and s_n small enough, the relative loss of efficiency $I^{-1}-n\text{Var}(\hat{\theta}_n^{\text{APPR}})$ is of order $n^{-4/7}$. If we choose g and h of the same order, however, the relative loss of efficiency would be of larger order. In fact, the optimal common order of g and h is then $n^{-1/5}$ and this would yield an order of $n^{-2/5}$ for the relative loss of efficiency.

It is important to note here that optimal kernel-based estimation of the score function ψ would lead to different choices of h and g (even differing in the rate). For instance, if we estimate ψ by $\hat{\psi}(x) = -[\hat{f}'_{X,g}(\theta+x) - \hat{f}'_{X,g}(\theta-x)][\hat{f}_{X,h}(\theta+x) + \hat{f}_{X,h}(\theta-x) + s_n]^{-1}$, under appropriate smoothness assumptions we arrive at the following asymptotic expansion for the mean integrated squared error $E\int(\hat{\psi}(x)-\psi(x))^2 f(x)\mathrm{d}x$:

$$\frac{1}{4}g^4 d_K^2 C_1 + \frac{1}{4}h^4 d_K^2 C_2 - \frac{1}{2}h^2 g^2 d_K^2 C_3$$

$$+\frac{2}{ng^3}\int\frac{f^2}{(s_n+2f)^2}\int(K')^2 + \frac{8}{nh}\int\frac{(f')^2 f^2}{(s_n+2f)^4}\int K^2$$

with $C_1 = \int(f'''/f)^2 f$, $C_2 = \int(f'/f)^2(f''/f)^2 f$ and $C_3 = \int[(f'f''f''')/f^3]f$. Here, if we insist on $h = g$, we arrive at optimal bandwidths $h, g \sim n^{-1/7}$. For the estimation of θ, this would give a relative loss of efficiency of order $n^{-2/7}$ (see Theorem 1). On the other hand, if we allow h and g to be different, then, if $C_3 > 0$, we get $h = C_3^{1/2}C_2^{-1/2}g(1+\mathrm{o}(1))$. The optimal choice of g would be $\left[6\int\{f^2/(s_n+2f)^2\}\int(K')^2/\{nd_K^2(C_1 - C_3^2/C_2)\}\right]^{1/7}$. Note that in this case $h \sim g \sim n^{-1/7}$ which leads again to a relative loss of order $n^{-2/7}$. These rates are much slower than those achieved above. This shows that (at least in our set-up) “optimal” estimation of the score function ψ does not lead to “optimal” adaptive location estimation.

We now come to the case of densities with not necessarily compact support. The rates of convergence of the terms in the formula of Theorem 1 which involves s_n are described in the next lemma.

Lemma 1. *Assume* $s_n \to 0$ *and* (A.2). *Then*

$$\int\frac{f^2}{(s_n+2f)^2} = \mathrm{O}(-\log s_n), \tag{3.3}$$

$$\int\frac{(f')^2 f^2}{(s_n+2f)^2} = \mathrm{O}((-\log s_n)^{2C_2+1}), \tag{3.4}$$

$$I - \int\frac{4(f')^2}{(s_n+2f)^2}f = \mathrm{O}(s_n(-\log s_n)^{2C_2+1}). \tag{3.5}$$

Suppose now that s_n is of order $n^{-\rho}$ with some $\rho > 0$. Then, using Theorem 1 and Lemma 1, we conclude that the optimal choice of g is of order $(\log n)^{\alpha}n^{-1/7}$ with an $\alpha \geqslant 0$. The terms with h^2 and $n^{-1}h^{-1}$ in the formula of Theorem 1 are the leading terms which include h. Thus a sensible choice of h is of order $n^{-1/3}(\log n)^{\beta}$ with a $\beta \geqslant 0$. The value of s_n may influence the robustness of $\hat{\theta}_n$. Therefore it may not only be chosen so that the loss of efficiency is small. Our calculations in the next section require that $s_n/(n^{-1}h^{-1}) \to +\infty$. Furthermore, the last term in the expansion of Theorem 1 should not dominate the g^4 and $n^{-1}g^{-3}$ terms, i.e. $s_n/n^{-4/7} \to 0$. An example of s_n that suffices these requirements would be $s_n \sim n^{-13/21}$. For simplicity let us fix this rate of convergence for s_n in the next section.

4. Accuracy of the stochastic approximation in Section 3

In this section we show that $\hat{\theta}_n^{\mathrm{APPR}}$ works well as an approximation for $\hat{\theta}_n = \hat{\theta}_n(\bar{\theta}_n)$. For s_n, h and g we assume the rates of convergence discussed in Section 3.
(A.5) $s_n \sim n^{-13/21}(\log n)^{\gamma}$, $h \sim n^{-1/3}(\log n)^{\beta}$ and $g \sim n^{-1/7}(\log n)^{\alpha}$ for some constants $\alpha, \beta, \gamma \geqslant 0$.

Theorem 2. *Assume* (A.1), (A.2), (A.4) *and* (A.5). *For every* $D > 0$, *the following holds*:

$$\sup_{|\theta'-\theta| \leqslant Dn^{-1/2}} n^{1/2}|\hat{\theta}_n(\theta') - \hat{\theta}_n^{\mathrm{APPR}} - I^{-1}\Gamma_{h,g}\,(\theta'-\theta)| = o_{\mathrm{P}}(n^{-4/7}), \tag{4.1}$$

where

$$\Gamma_{h,g} = n^{-1}\sum_{i=1}^{n} 2f_g'(X_i-\theta)[s_n + 2f_h(X_i-\theta)]^{-2}[\hat{f}'_{X,h,i}(X_i-\theta) + \hat{f}'_{X,h,i}(2\theta - X_i)].$$

$\Gamma_{h,g}$ *is a quadratic form with*

$$\begin{aligned} &E(\Gamma_{h,g}) = 0, \\ &\mathrm{Var}(\Gamma_{h,g}) = \mathrm{O}(n^{-2}h^3) = \mathrm{O}(n^{-1}(\log n)^{3\beta}). \end{aligned} \tag{4.2}$$

Theorems 1 and 2 imply that, under our assumptions, estimates $\hat{\theta}_n = \hat{\theta}_n(\bar{\theta}_n)$ are adaptive. We state this as a corollary.

Corollary 1. *Assume* (A.1)–(A.5). *Then it holds that*

$$\sqrt{n}(\hat{\theta}_n(\bar{\theta}_n) - \theta) \to N(0, I^{-1}) \quad (\textit{in distribution}).$$

We come now to the second order performance of adaptive estimates $\hat{\theta}_n = \hat{\theta}_n(\bar{\theta}_n)$. For asymptotically linear preliminary estimates $\bar{\theta}_n$, Theorem 2 has the following implication.

Corollary 2. *Assume* (A.1)–(A.5) *and there exists a function* χ *with*

$$\begin{aligned} &E\chi(\varepsilon_i) = 0, \quad E\chi^2(\varepsilon_i) < +\infty, \\ &\bar{\theta}_n = \theta + \frac{1}{n}\sum_{i=1}^{n}\chi(\varepsilon_i) + o_{\mathrm{P}}(n^{-4/7}(\log n)^{-3\beta/2}). \end{aligned}$$

Then it holds with a constant $\gamma > 0$ *that*

$$n^{1/2}(\hat{\theta}_n(\bar{\theta}_n) - \tilde{\theta}_n) = \mathrm{o}_\mathrm{p}(n^{-4/7}), \tag{4.3}$$

$$n\left[\mathrm{Var}(\tilde{\theta}_n) - \mathrm{Var}(\hat{\theta}_n^{\mathrm{APPR}})\right] = \mathrm{O}(n^{-11/14}(\log n)^\gamma) = \mathrm{o}(n^{-2/3}), \tag{4.4}$$

where

$$\tilde{\theta}_n = \hat{\theta}_n^{\mathrm{APPR}} + I^{-1}\Gamma_{h,g}\frac{1}{n}\sum_{i=1}^{n}\chi(\varepsilon_i).$$

Corollary 2 states that the asymptotic second order performance of the adaptive estimate $\hat{\theta}_n(\bar{\theta}_n)$ does not depend on the preliminary estimate $\bar{\theta}_n$ (under minimal regularity conditions on $\bar{\theta}_n$). And it says that $n^{1/2}(\hat{\theta}_n(\bar{\theta}_n) - \theta)$ admits a stochastic approximation, up to $\mathrm{o}_\mathrm{P}(n^{-4/7})$, which has the same asymptotic second order quadratic risk as described in Theorem 1.

5. Simulations

In this section, we confirm by simulation the theoretical findings that one can do better by choosing h and g of different order. We compare the two approaches, one using two different bandwidths h and g, and the other using $h = g$. For these comparisons, we use the theoretically optimal bandwidths obtained by the formula in Theorem 1. First, consider the case $h \neq g$. For each selected value of s_n, we use the optimal bandwidth g which trades off the two terms involving g in (3.2). Likewise, we use the optimal h obtained in the same way. For the case $h = g$, we trade off the $n^{-1}g^{-3}$ and the h^2 terms in (3.2) since the other terms involving h and g are negligible in this case.

Note that these optimal bandwidths are of the form (constant)$\times n^{-\alpha}(\alpha > 0)$, and the constant factor depends on the unknown error density f. For practical implementation the f-dependent constant should be estimated. The problem of estimating this constant would be interesting. However, we do not attempt to deal with this problem here, but simply use the theoretical constant, obtained by plugging the underlying f into the formula, since the main purpose in this section is to see how the asymptotic benefit of using different bandwidths comes into effect in finite sample cases.

The underlying error distributions chosen in this comparison are the standard normal $\mathrm{N}(0,1)$, the standard Cauchy $\mathrm{C}(0,1)$ and Student's t with 3 degrees of freedom $t(3)$ as an intermediate of the first two. For the trimming constants, the selected values of s_n are 0.001 and 0.01. In the simulations we have used an additional trimming constant d_n. This constant is used as follows. In the definitions of $\Delta_{h,g}(\theta)$ and $I_{h,g}(\theta)$, the summation runs only over the observations with $|X_i - \theta| \leqslant d_n$. The value of d_n used in the simulation is 2. In fact, for $d_n \to \infty$ as $n \to \infty$, one can show that the conclusions in Sections 3 and 4 remain valid by the arguments parallel to those without d_n. In particular, this additional trimming may be appropriate in case of heavier tailed distributions such as Cauchy which our assumption (A.2) excludes. As kernel function, we use the quartic kernel $K(x) = (15/16)(1-x^2)^2 I_{[-1,1]}(x)$. The preliminary estimator $\bar{\theta}_n$ is the sample median.

The means of comparison is the mean squared error (MSE):

$$\text{MSE} = E(\hat{\theta}_n - \theta)^2.$$

Table 1 contains the Monte Carlo approximation of MSE based on 500 pseudo samples of size 100 and 400.

In the table, MSE1 means the mean squared error of the estimator when the two different bandwidths h and g, which are optimally chosen, are used. MSE2 represents the mean squared error corresponding to the optimally chosen bandwidth $h(=g)$. For comparison of MSE1 and MSE2 the table gives the ratio MSE1/MSE2. Furthermore, the table contains the values of the optimal h and g, denoted by h_{opt} and g_{opt} respectively, and the values of the common optimal bandwidth $h(=g)$ denoted by $(h=g)_{\text{opt}}$. Note that the three bandwidths always satisfy the ordering $h_{\text{opt}} < (h=g)_{\text{opt}} < g_{\text{opt}}$.

From the table, one can see that the gains obtained by using $h \neq g$ are small in the standard normal case and the improvement is not great even when the sample size is increased to 400. However, in the cases of $t(3)$ and the standard Cauchy, one can find drastic changes in the ratios of MSE's as the sample size increases. This illustrates the fact that the benefit of using different bandwidths comes into effect rapidly as the sample size gets larger.

6. Proofs

Without loss of generality, for simplification of notation, we assume $\theta = 0$.

Proof of Theorem 1. We start by showing the asymptotic normality of $\hat{\theta}_n^{\text{APPR}}$. For this purpose note that

$$\hat{\theta}_n^{\text{APPR}} = \left[\int \frac{4f_g' f_h'}{[s_n + 2f_h]^2} f\right]^{-1} \left[-n^{-2} \sum_{i \neq j} \frac{K_g'(X_i - X_j) - K_g'(-X_i - X_j)}{s_n + 2f_h(X_i)}\right]$$

$$+ o_P(n^{-1/2})$$

$$= \left[\int \frac{4f_g' f_h'}{[s_n + 2f_h]^2} f\right]^{-1} \left[-n^{-1} \sum_{i=1}^{n} \frac{2f_g'(X_i)}{s_n + 2f_h(X_i)}\right]$$

$$+ o_P(n^{-1/2})$$

$$= -I^{-1} n^{-1} \sum_{i=1}^{n} \frac{f'(X_i)}{f(X_i)} + o_P(n^{-1/2}).$$

Here the equalities follow by calculation of second moments of the differences between the terms. For the second equality see formula (6.2). Asymptotic normality of the last term implies asymptotic normality of $\hat{\theta}_n^{\mathrm{APPR}}$.

For the proof of (3.2) we show here two details of the proof and we give a hint why the leading g term is of order g^4. The other calculations are of similar type. We prove here equations (6.1) and (6.2):

$$\begin{aligned}\int \frac{4f'_g f'_h}{[s_n+2f_h]^2} f &= I + g^2 2d_K \int f' f'''[s_n+2f]^{-2} f \\ &\quad + g^4 \frac{e_K}{24} \int f' f^{(5)}/f + h^2 d_K \left[\frac{1}{2}\int f' f'''/f - \int (f')^2 f''/f^2\right] \\ &\quad + \left[\int 4(f')^2 [s_n+2f]^{-2} f - I\right](1+\mathrm{o}(1)) \\ &\quad + \mathrm{o}(g^4+h^2), \end{aligned} \tag{6.1}$$

where $e_K = \int t^4 K(t)\mathrm{d}t$,

$$\mathrm{Var}(T) = n^{-1} g^{-3} \int \frac{f^2}{(s_n+2f)^2} 2 \int (K')^2 (1+\mathrm{o}(1)), \tag{6.2}$$

where

$$T = n^{-3/2} \Sigma^{\neq}_{i,j} \frac{K'_g(X_i - X_j) - K'_g(-X_i - X_j)}{s_n + 2f_h(X_i)}$$

$$- n^{-1/2} \sum_{i=1}^{n} \frac{2f'_g(X_i)}{s_n + 2f_h(X_i)}.$$

Proof of (6.1). Note that

$$\begin{aligned} f'_g(x) &= f'(x) + \frac{1}{2} g^2 d_K f'''(x) + \frac{1}{24} g^4 e_K f^{(5)}(\xi(x)), \\ f'_h(x) &= f'(x) + \frac{1}{2} h^2 d_K f'''(\eta(x)), \\ f_h(x) &= f(x) + \frac{1}{2} h^2 d_K f''(\zeta(x)), \end{aligned}$$

where $|\xi(x)-x| \leqslant Cg$, $|\eta(x)-x| \leqslant Ch$, and $|\zeta(x)-x| \leqslant Ch$. Now we put these expressions into the left-hand side of formula (6.1) and we expand the denominator of the integrand. For (6.1) it remains to show things of the following type:

$$\int 4f'(x) f^{(5)}(\xi(x))(s_n+2f)^{-2} f \longrightarrow \int f' f^{(5)}/f. \tag{6.3}$$

For (6.3) it suffices to show that the integrand of the left-hand side is absolutely bounded (uniformly in n) by an integrable function. This can be done easily by using (A.2) as follows:

$$\begin{aligned}
&|4f'(x)f^{(5)}(\xi(x))(s_n+2f(x))^{-2}f(x)|\\
&\leqslant C|f'(x)f^{(5)}(\xi(x))f^{-1}(x)|\\
&\leqslant C(1+|x|^{C_2})(1+|\xi(x)|^{C_2})\exp(-C_1|\xi(x)|)\\
&\leqslant C\exp(-C|x|).
\end{aligned}$$

Proof of (6.2). The variable T can be written as

$$T = n^{-3/2}\Sigma^{\neq}_{i,j}w(X_i,X_j) - n^{-3/2}\sum_{i=1}^{n}\frac{2f_g'(X_i)}{s_n+2f_h(X_i)}$$

where

$$w(X_i,X_j) = \frac{K_g'(X_i-X_j) - K_g'(-X_i-X_j) - 2f_g'(X_i)}{s_n+2f_h(X_i)}.$$

Now with $E(w(X_i,X_j)|X_j) = E(w(X_i,X_j)|X_i) = 0$ and $Ew(X_i,X_j)^2 = \int(K_g'(x-y) - K_g'(-x-y))^2(s_n+2f_h(x))^{-2}f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y(1+\mathrm{o}(1))$, the formula (6.2) follows.

Let us shortly comment why the g^2 terms cancel in (3.2). For seeing this it suffices to consider the variance of

$$\left[\int\frac{4f_g'f_h'}{[s_n+2f_h]^2}f\right]^{-1}n^{-1/2}\sum_{i=1}^{n}\frac{2f_g'(X_i)}{s_n+2f_h(X_i)}.$$

The variance of $n^{-1/2}\sum_{i=1}^{n}2f_g'(X_i)/[s_n+2f_h(X_i)]$ is $\int 4f_g'f_g'[s_n+2f_h]^{-2}f$. It can easily be seen that the g^2 terms cancel if this expression is divided by the square of the right-hand side of (6.1). □

Proof of Lemma 1. We will give here only the proof of (3.5). The formula (3.3) and (3.4) can be shown similarly. Let $1(A)$ denote the indicator defined by $1(A)=1$ if A is satisfied and $1(A)=0$ otherwise. Then we get

$$\begin{aligned}
\left|I-\int 4(f')^2(s_n+2f)^{-2}f\right| &= \left|\int(f')^2f^{-1}[s_n+2f]^{-2}[s_n^2+4s_nf]\right|\\
&\leqslant C\int(f')^2f^{-1}1(|x|\geqslant -D\log s_n)\\
&\quad +Cs_n\int(f')^2f^{-2}1(|x|<-D\log s_n)
\end{aligned}$$

with a constant D chosen below. Note that (A.2) implies

$$\begin{aligned}
&(f')^2f^{-1}(x)\leqslant C'(1+|x|^{2C_2})\exp(-C_1|x|)\leqslant C''\exp(-C'''|x|),\\
&(f')^2f^{-2}(x)\leqslant C'''(1+|x|^{2C_2}).
\end{aligned}$$

This gives

$$\left|I - \int 4(f')^2(s_n + 2f)^{-2}f\right| \leqslant C \cdot \int \exp(-C'''|x|)1(|x| \geqslant -D\log s_n) + Cs_n \int (1 + |x|^{2C_2})1(|x| < -D\log s_n) = \mathrm{O}(s_n(-\log s_n)^{2C_2+1})$$

for D large enough. □

Proof of Theorem 2. We divide the proofs into two parts. □

Proposition 1.

$$\sup_{|\theta'| \leqslant Dn^{-1/2}} n^{1/2}|\hat{\theta}_n(\theta') - \hat{\theta}_n(0) - I^{-1}\Gamma_{h,g}\theta'| = \mathrm{o_P}(n^{-4/7}).$$

Proposition 2.

$$n^{1/2}(\hat{\theta}_n(0) - \hat{\theta}_n^{\mathrm{APPR}}) = \mathrm{o_P}(n^{-4/7}).$$

Proposition 1 follows from the following three lemmas.

Lemma 2.

$$\sup_{|\theta'| \leqslant Dn^{-1/2}} |I_{h,g}(\theta') - I_{h,g}(0)| = \mathrm{o_P}(n^{-4/7}).$$

Lemma 3.

$$\sup_{|\theta'| \leqslant Dn^{-1/2}} n^{1/2}|\Delta_{h,g}(\theta') - \Delta_{h,g}(0) - \Delta'_{h,g}(0)\theta'| = \mathrm{o_P}(n^{-4/7}).$$

Lemma 4.

$$\Delta'_{h,g}(0) + I_{h,g}(0) = \Gamma_{h,g} + \mathrm{o_P}(n^{-4/7}).$$

For Proposition 2 it suffices to show the following two lemmas.

Lemma 5.

$$I_{h,g}(0) = n^{-1}\sum_{i=1}^{n} \frac{4f'_g(X_i)f'_h(X_i)}{[s_n + 2f_h(X_i)]^2} + n^{-1}\sum_{i=1}^{n} \frac{2f'_g(X_i)[\hat{f}'_{X,h,i}(X_i) - \hat{f}'_{X,h,i}(-X_i) - 2f'_h(X_i)]}{[s_n + 2f_h(X_i)]^2} + \mathrm{o_P}(n^{-4/7}).$$

Lemma 6.

$$\begin{aligned}
&n^{1/2}\Bigg(\Delta_{h,g}(0) + n^{-1} \sum_{i=1}^{n} \frac{\hat{f}'_{X,g,i}(X_i) - \hat{f}'_{X,g,i}(-X_i)}{s_n + 2f_h(X_i)} \\
&\quad - n^{-3} \Sigma^{\neq}_{i,j,k} [K'_g(X_i - X_j) - K'_g(-X_i - X_j)] \\
&\quad \times [K_h(X_i - X_k) + K_h(-X_i - X_k) - 2f_h(X_i)][s_n + 2f_h(X_i)]^{-2}\Bigg) \\
&\quad = \mathrm{o_P}(n^{-4/7}).
\end{aligned}$$

The proofs of Lemmas 2–6 are based on very lengthy calculations using higher order stochastic expansions of $\Delta_{h,g}(0)$, $\Delta'_{h,g}(0)$, and so on. Because of the similarities of the calculations we omit the proofs of Lemmas 5 and 6.

Proof of Lemma 2. We show

$$I'_{h,g}(0) = \mathrm{o_P}(n^{-1/14}), \tag{6.4}$$

$$\sup_{|\theta'| \leqslant Dn^{-1\,2}} |I''_{h,g}(\theta')| = \mathrm{o_P}(n^{3/7}). \tag{6.5}$$

First, note that $I'_{h,g}(0)$ consists of three summands. We treat only the term

$$\begin{aligned}
T = {} & n^{-3} \Sigma^{\neq}_{i,j,k} [K'_g(X_i - X_j) - K'_g(-X_i - X_j)]\, 2\, [-K''_h(-X_i - X_k)] \\
& \times [s_n + \hat{f}_{h,i}(X_i) + \hat{f}_{h,i}(-X_i)]^{-2}.
\end{aligned}$$

Expansion of the denominator gives: $T = T_1 + T_2 + T_3$, where

$$\begin{aligned}
T_1 = {} & n^{-3} \Sigma^{\neq}_{i,j,k} [K'_g(X_i - X_j) - K'_g(-X_i - X_j)] \\
& \times 2[-K''_h(-X_i - X_k)][s_n + 2f_h(X_i)]^{-2}, \\
T_2 = {} & -2n^{-3} \Sigma^{\neq}_{i,j,k} [K'_g(X_i - X_j) - K'_g(-X_i - X_j)] \\
& \times 2[-K''_h(-X_i - X_k)][s_n + 2f_h(X_i)]^{-3} \\
& \times [\hat{f}_{X,h,i}(X_i) + \hat{f}_{X,h,i}(-X_i) - 2f_h(X_i)], \\
T_3 = {} & 3n^{-3} \Sigma^{\neq}_{i,j,k} [K'_g(X_i - X_j) - K'_g(-X_i - X_j)] \\
& \times 2[-K''_h(-X_i - X_k)][s_n + 2f_h(X_i)]^{-4} \\
& \times w_i[\hat{f}_{X,h,i}(X_i) + \hat{f}_{X,h,i}(-X_i) - 2f_h(X_i)]^2,
\end{aligned}$$

with

$$\max_{1\leqslant i\leqslant n} |w_i - 1| \leqslant C\Delta,$$
$$\Delta = \max_{1\leqslant i\leqslant n} |[\hat{f}_{X,h,i}(X_i) + \hat{f}_{X,h,i}(-X_i) - 2f_h(X_i)][s_n + 2f_h(X_i)]^{-1}|.$$

One can show $T_1 = o_P(n^{-1/14})$ and $T_2 = o_P(n^{-1/14})$ by calculation of the second moments. For the treatment of T_3, note that for λ large enough

$$P(\max_{1\leqslant i\leqslant n} |X_i| \leqslant \lambda \log n) \to 1 \quad \text{as} \quad n \to \infty. \tag{6.6}$$

This implies, that with probability tending to one,

$$T_3 = 3n^{-1} \sum_{i=1}^{n} 1_i[\hat{f}'_{X,g,i}(X_i) - \hat{f}'_{X,g,i}(-X_i)](-2)\hat{f}''_{X,h,i}(-X_i)[s_n + 2f_h(X_i)]^{-4}$$
$$\times w_i[\hat{f}_{X,h,i}(X_i) + \hat{f}_{X,h,i}(-X_i) - 2f_h(X_i)]^2,$$

where $1_i = 1(|X_i| \leqslant \lambda \log n)$. We apply now that for every bandwidth b with $b(\log n) \to 0$ and $nb \to \infty$ the following two formulas hold:

$$\sup_{|x|\leqslant \lambda \log n, 1\leqslant i\leqslant n} |\hat{f}^{(j)}_{X,b,i}(x) - f^{(j)}_b(x)|[n^{-1}b^{-1}\log n \vee f(x)]^{-1/2}$$
$$= O_P(n^{-1/2}b^{-(2j+1)/2}(\log n)^{1/2}) \quad (j = 0, 1, \ldots), \tag{6.7}$$

where $c \vee d$ denotes the maximum of c and d,

$$\sup_{|x|\leqslant \lambda \log n} \left|\frac{f_b(x)}{f(x)} - 1\right| \to 0. \tag{6.8}$$

The statements (6.7) and (6.8) will be proved below. Note that (6.6) and (6.7) imply

$$\Delta = O_p(n^{-1/42}(\log n)^C). \tag{6.9}$$

With (6.7), (6.8) and (6.9) we get an upper bound of $|T_3|$ as follows:

$$|T_3| \leqslant Cn^{-1} \sum_{i=1}^{n} 1_i[|f'_g(X_i)| + U_n(f(X_i) \vee n^{-1}g^{-1}\log n)^{1/2} n^{-1/2} g^{-3/2}(\log n)^{1/2}]$$
$$\times[\,|f''_h(X_i)| + V_n(f(X_i) \vee n^{-1}h^{-1}\log n)^{1/2} n^{-1/2} h^{-5/2}(\log n)^{1/2}]$$
$$\times[s_n + 2f(X_i)]^{-4}[W_n(f(X_i) \vee n^{-1}h^{-1}\log n)^{1/2} n^{-1/2} h^{-1/2}(\log n)^{1/2}]^2,$$

where U_n, V_n, W_n are positive random variables of order $O_P(1)$. An evaluation of the summands gives now

$$|T_3| = o_P(n^{-1/14}(\log n)^{-1}) \cdot n^{-1} \sum_{i=1}^{n} 1_i \left[1 + \frac{1}{f(X_i)}\right].$$

Since $E(1_i[1 + 1/f(X_i)]) = O(\log n)$, this shows $T_3 = o_P(n^{-1/14})$.

It remains to show (6.7) and (6.8). First, (6.8) follows from the following inequality for x, δ:

$$\left| \log \frac{f(x+\delta)}{f(x)} \right| = |\log f(x+\delta) - \log f(x)| = |\delta f'(x+\delta^*)/f(x+\delta^*)|$$
$$\leqslant C|\delta|(1+|x+\delta^*|^{C_2}) \leqslant C|\delta|(1+|x|^{C_2}+|\delta^*|^{C_2}),$$

where δ^* is chosen such that $|\delta^*| \leqslant |\delta|$. Let us prove (6.7) now. Because $\hat{f}^{(j)}_{X,b,i}$ has a derivative which is bounded by a deterministic constant which increases polynomially in n, it suffices to prove (6.7) with the supremum taken over a grid with a polynomially increasing number of points. For simplicity, we consider here $\hat{f}_{X,b}$ instead of $\hat{f}_{X,b,i}$. Put, for L large enough,

$$c_n = Ln^{-1/2} b^{-(2j+1)/2} (\log n)^{1/2},$$
$$d_n = n^{1/2} b^{(2j+1)/2} (\log n)^{1/2},$$
$$b_n(x) = [n^{-1} b^{-1} (\log n) \vee f(x)]^{-1/2}.$$

Then $|n^{-1} d_n b_n(x) K_b^{(j)}(y)|$ is bounded by $\kappa_j = \sup_u |K^{(j)}(u)|$. One gets

$$P(b_n(x)(\hat{f}^{(j)}_{X,b}(x) - f_b^{(j)}(x)) \geqslant c_n)$$
$$= P\Big(b_n(x) n^{-1} \sum_{i=1}^{n} \big[K_b^{(j)}(X_i - x) - f_b^{(j)}(x)\big] \geqslant c_n\Big)$$
$$\leqslant E \exp\Big(b_n(x) d_n \cdot n^{-1} \sum_{i=1}^{n} \big[K_b^{(j)}(X_i - x) - f_b^{(j)}(x)\big]\Big) \exp(-d_n c_n)$$
$$= \{E \exp(b_n(x) d_n \cdot n^{-1} [K_b^{(j)}(X_1 - x) - f_b^{(j)}(x)])\}^n \exp(-d_n c_n)$$
$$\leqslant \Big\{ 1 + \frac{1}{2} b_n(x)^2 d_n^2 n^{-2} E[K_b^{(j)}(X_1 - x) - f_b^{(j)}(x)]^2$$
$$+ \frac{1}{6} \exp(\kappa_j) b_n(x)^3 d_n^3 n^{-3} E|K_b^{(j)}(X_1 - x) - f_b^{(j)}(x)|^3 \Big\}^n \exp(-d_n c_n)$$
$$\leqslant \exp(C \log n - L \log n) \leqslant n^{-\rho(L)},$$

with an increasing function ρ.

Proof of Lemma 3. It suffices to show

$$\sup_{|\theta'| \leqslant Dn^{-1/2}} |\Delta''_{h,g}(\theta')| = o_P(n^{-1/14}). \tag{6.10}$$

Claim (6.10) can be shown by calculations similar to the proof of Lemma 2.

Proof of Lemma 4. We will show that

$$n^{-1}\sum_{i=1}^{n}\frac{2\hat{f}''_{X,g,i}(-X_i)}{s_n+\hat{f}_{X,h,i}(X_i)+\hat{f}_{X,h,i}(-X_i)}=o_P(n^{-4/7}). \tag{6.11}$$

The other summands of $\Delta'_{h,g}(0)+I_{h,g}(0)$ can be treated similarly. Again, we expand the random denominators as in the proof of Lemma 2. An expansion of the left-hand side of (6.11) up to the Rth term involves terms of type

$$S_r=n^{-1}\sum_{i=1}^{n}\frac{\hat{f}''_{X,g,i}(-X_i)}{(s_n+2f_h(X_i))^{r+1}}[\hat{f}_{X,h,i}(X_i)+\hat{f}_{X,h,i}(-X_i)-2f_h(X_i)]^r,$$
$$(r=0,\dots,R-1).$$

The remainder in this expansion, denoted by S_R, suffices

$$|S_R|\leqslant Cn^{-1}\sum_{i=1}^{n}1_i\frac{|\hat{f}''_{X,g,i}(-X_i)|}{(s_n+2f_h(X_i))^{R+1}}|\hat{f}_{X,h,i}(X_i)+\hat{f}_{X,h,i}(-X_i)-2f_h(X_i)|^R$$

with probability tending to one. Here again $1_i=1(|X_i|\leqslant\lambda\log n)$ as in the proof of Lemma 2. An expansion of order $R=25(!)$ will work here. The fact that $S_r=o_P(n^{-4/7})$ for $r=0,\dots,R-1$ can be shown by calculation of the second moments. By use of (6.7) and (6.9), $|S_R|$ can be bounded as follows:

$$|S_R|\leqslant C\cdot O_P(1)\cdot n^{-1}\sum_{i=1}^{n}1_i f(X_i)^{-1}n^{-25/42}(\log n)^{C''}.$$

Now, $E1_i f(X_i)^{-1}=O(\log n)$ implies $S_R=o_P(n^{-4/7})$.

References

Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647–671.

Faraway, J.J. (1992). Smoothing in adaptive estimation. *Ann. Statist.* **20**, 414–427.

Hsieh, D.A. and C.F. Manski (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Ann. Statist.* **15**, 541–551.

Jin, K. (1992). Empirical smoothing parameter selection in adaptive estimation. *Ann. Statist.* **20**, 1844–1874.

Park, B.U. (1993). A cross-validatory choice of smoothing parameter in adaptive location estimation. *J. Amer. Statist. Assoc.* **88**, 848–854.

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference* **16**, 89–105.

Stone, C. (1975). Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3**, 267–284.