PROBABILITY
AND
MATHEMATICAL STATISTICS

Vol. 13, Fasc. 2 (1992), pp. 245–252

MINIMUM L_1-PENALIZED DISTANCE ESTIMATORS OF A DENSITY AND ITS DERIVATIVES

BY

LESŁAW GAJEK (ŁÓDŹ)

Abstract. Let F be an $(m+1)$-times differentiable distribution function (df) generating the data. Let f be the density of F. Let F_n denote the empirical df. The paper concerns fitting an $(m+1)$-times differentiable function G to the data by minimizing $d_n(G) = \|F_n - G\|_1 + \beta(n)\|G^{(m+1)}\|_1$, where $\|\cdot\|_p$, $p \geqslant 1$, denotes the L_p-norm and $\beta(n) > 0$ is a sequence of smoothing parameters. Let $\hat{F}_n$ be an (approximate) minimizer of the above problem. We establish an upper bound for $\mathrm{E}\|\hat{F}_n^{(i)} - F^{(i)}\|_1$, $i = 1, \ldots, m$, with respect to the choice of β. In particular, the choice of $\beta \sim n^{-1/(m+1)}$ results in the optimal L_1-rate of convergence of $\hat{F}_n'$ to f. The estimation $\mathrm{E}\|\hat{F}_n^{(i)} - F^{(i)}\|_2^2$ is also evaluated.

1. Introduction. Let $\mathscr{F}$ be some family of distribution functions (df's) and let d be a distance between df's. Let $R\colon \mathscr{F} \to \boldsymbol{R}_+$ be a penalty function and denote by F_n the empirical df. We say that $\hat{F}_n\colon \boldsymbol{R}^n \to \mathscr{F}$ is a *minimum penalized distance* (MPD) *distribution function estimator* if

$$d(\hat{F}_n, F_n) + \beta(n)R(\hat{F}_n) = \inf_{\mathscr{F}} \{d(F, F_n) + \beta(n)R(F)\} \tag{1}$$

for every sample point $x^n \in \boldsymbol{R}^n$, where $\beta(n) > 0$ is a sequence of smoothing parameters. Without loss of generality we assume that the infimum is achieved. If not, one can use any $\hat{F}_n$ that brings $d(\hat{F}_n, F_n) + \beta(n)R(\hat{F}_n)$ within ε_n decreasing quickly to zero.

The MPD estimator of a density is defined as a derivative of the MPD df estimator.

Given a distance d and a penalty for sharpness R, $\beta(n)$ plays a similar role to that of the bandwidth in the kernel estimation: to balance between the maximal smoothing and the maximal fitting the estimator to the data. So an important goal is to choose $\beta(n)$ properly to a given class $\mathscr{F}$ of df's.

In [9]–[11], the problem of strong consistency of MPD density estimators was considered when d was the norm sup, $\mathscr{F}$ was a subclass of $(m+1)$-times differentiable functions, and the penalty for roughness was $R(F) = \sup|F^{(m+1)}|$.

In [6] and [7], the mean integrated square error (MISE) of MPD estimators was investigated for d and R generated by the L_p-norm with $p=2$, while the strong consistency was treated for any $1 \leqslant p \leqslant \infty$. Moreover, in some classes of analytic functions the minimum distance estimators (defined by (1) with $\beta \equiv 0$) were shown to achieve extraordinary rates of L_1-, L_2- and L_∞-convergence.

The aim of this paper is to analyze the case where

$$d(F, F_n) = \int |F(t) - F_n(t)|\, dt \tag{2}$$

and

$$R(F) = \int |F^{(m+1)}(t)|\, dt$$

for $\mathscr{F}$ being a subclass of $(m+1)$-times differentiable functions.

In Section 2 we show that the MPD density estimators achieve, for a properly chosen sequence β, the best L_1-rate of convergence. However, for the L_2-convergence properties of the MPD density estimators defined via the distance (2), we were able to prove a weaker result. Theorem 2.3 implies that their MISE converges as $O(n^{-(2m-1)/(2m+1)})$ while the optimal rate is known to be $O(n^{-2m/(2m+1)})$. This presumable suboptimality can be explained in the way that fitting df to the data in the L_1-norm one assumes an importance of the distribution tails stronger than necessary when compared with the L_2-fitting. Further comments and comparisons can be found in Section 3.

All proofs are given in the Appendix. Somehow related results for regression function estimators can be found in [8].

2. The L_1- and L_2-rates of convergence of the MPD estimators. In order to establish the rates of L_1- and L_2-convergence of the MPD estimators we shall need that the following Lipschitz condition be satisfied:

There are L and $t > 0$ such that for all $|y| < t$

$$\int |F(x+y) - F(x)|^{1/2}\, dx \leqslant L|y|^{1/2}. \tag{3}$$

In Section 3 we give sufficient and necessary conditions for (3) to hold.

Throughout the paper we say that $\hat{F}_n$ is an *MPD type estimator* if $\hat{F}_n$ is a solution of the minimization problem (1) within the class $\mathscr{F}$ consisting (a) of df's for $m \leqslant 2$; (b) of measure generating functions for $m > 2$ (see [7]).

THEOREM 2.1. *Let $\hat{F}_n$ be an MPD type estimator of an $(m+1)$-times differentiable df for which (3) holds. Let $\beta(n)$ be a sequence of smoothing parameters tending to zero as $n \to \infty$. Then for every $i = 1, \ldots, m$*

$$\mathrm{E}\, \|\hat{F}_n^{(i)} - F^{(i)}\|_1 \leqslant \beta^{-i/(m+1)} \left\{ H_1 \left[\frac{\beta^{1/(m+1)}}{n} \right]^{1/2} + \beta H_2 \right\},$$

where H_1 and H_2 are some positive constants involving L and $\|F^{(m+1)}\|_1$ (see (17) and (18) in the Appendix below for their explicit values).

Theorem 2.1 enables one to choose $\beta(n)$ in an optimal way.

COROLLARY 2.2. *Let $\beta(n) = H_3 n^{-(m+1)/(2m+1)}$. Then*

$$\mathrm{E}\,\|\hat{F}_n^{(i)} - F^{(i)}\|_1 \leqslant n^{-(m+1-i)/(2m+1)}[H_1 H_3^{1/(m+1)} + H_2 H_3] H_3^{-i/(m+1)}.$$

The rate $n^{-m/(2m+1)}$ is known to be optimal for the L_1-convergence of density estimators in the class of m-times differentiable densities (see [1] and [3]). Thus Corollary 2.2 shows how to choose the sequence $\beta(n)$ of smoothing parameters to achieve the best possible rate of decreasing the expected L_1-error of MPD type estimators.

Since the L_1-distance puts more weight on the distribution tails than the L_2-distance does, the L_1-MPD estimators might be too "heavy" to achieve the best rate of decreasing their MISE. In fact, we have the following result:

THEOREM 2.3. *Let $\hat{F}_n$ be an MPD type estimator of an $(m+1)$-times differentiable df with a compact support. Let $\beta(n) \to 0$ as $n \to \infty$. Then for every $i = 1, \ldots, m$*

$$\mathrm{E}\,\|\hat{F}_n^{(i)} - F^{(i)}\|_2^2 \leqslant \beta^{-(2i+1)/(m+1)}\left[\frac{\beta^{1/(m+1)}}{n} H_4 + \beta^2 H_5\right] + \beta^{-2i/(m+1)}\left[\frac{\beta^{1/(m+1)}}{n} H_6 + \beta^2 H_7\right],$$

where H_4–H_7 are some positive constants which involve $\|F^{(m+1)}\|_1$ and $\|F^{(m+1)}\|_2$.

Let us notice that the rate of decreasing the MISE of the L_1-MPD estimators, following from Theorem 2.3, is slightly worse than the square of their L_1-rate of convergence. In fact, an optimal choice of β provided by Theorem 2.3 is again $\beta \sim n^{-(m+1)/(2m+1)}$.

COROLLARY 2.4. *If $\beta(n) = H_8 n^{-(m+1)/(2m+1)}$, then for $i = 1, \ldots, m$*

$$\mathrm{E}\,\|\hat{F}_n^{(i)} - F^{(i)}\|_2^2 \leqslant n^{-(2m+1-2i)/(2m+1)}[H_4 H_8^{1/(m+1)} + H_5 H_8^2 + o(1)].$$

From Corollary 2.4 and the formulas on H_4 and H_5 one could find an asymptotically optimal choice of H_8 which, however, involves $\|F^{(m+1)}\|_1$ and $\|F^{(m+1)}\|_2$ being unknown.

The optimal rate of decreasing the MISE for the density estimators in the class considered is known to be $n^{-2m/(2m+1)}$ while Corollary 2.4 gives a slower rate $n^{-(2m-1)/(2m+1)}$. This corresponds somehow to the known property that the minimum distance method in a parametric setup is very sensitive to changing the distance of fitting the model to the data (cf. [4] and [5]).

Let $\mathscr{F}(L, C)$ denote the class of all df's F with the Lipschitz constant not greater than L and $\|F^{(m+1)}\|_1 \leqslant C$. It is easy to see that the bounds given in Theorems 2.1 and 2.3 are uniform over the class $\mathscr{F}(L, C)$ whenever H_1 and H_2 are properly modified. A similar remark concerns the rates of convergence of the MPD estimators.

3. Some comments. To avoid a slow convergence phenomenon (see [3], p. 36, Theorem 1) one should impose a combination of continuity and tail conditions on the density f. For good reasons the quantity

$$D_m(f) = \|f^{(m)}\|_1^{1/(2m+1)}\left(\int \sqrt{f}\right)^{2m/(2m+1)}$$

can be used as a proper criterion that measures how long-tailed or unsmooth f is. Theorems 2.1 and 2.3 involve $\|f^{(m)}\|_1$ in H_2 and H_4, respectively. Seemingly, $\int \sqrt{f}$ does not appear but the following lemma shows that it is hidden in the Lipschitz condition (3).

LEMMA 3.1. *If* (3) *holds with the Lipschitz constant* L, *then* $\int \sqrt{f} \leqslant L$. *Conversely, if* f *is a unimodal and bounded density for which* $\int \sqrt{f} < \infty$, *then* (3) *is satisfied.*

It is of interest to compare the minimum distance method presented here with the minimum distance approach of Yatracos [12] (see also Devroye [2]). The latter method, which is applicable only to L_1 totally bounded families of densities, is a kind of the method of sieves. It has a disadvantage that one must construct an ε-cover of the family of densities $\mathscr{F}'$ *before* sampling from $f \in \mathscr{F}'$. Our method copes with this problem since it relies on finding the best approximation of the empirical df F_n but *after* sampling from f. So, only a neighbourhood of F_n has to be known when we construct an MPD estimator from a given sample. For this reason our method can be immediately applied to such families as the translation class or the scale class which are not totally bounded (cf. [2], p. 98). The problems discussed above can be also overcome following Yatracos [13].

APPENDIX

Proof of Theorem 2.1. Let k be an $(m+1)$-times continuously differentiable function vanishing outside an interval with the properties

$$\int k(x)dx = 1 \quad \text{and} \quad \int x^i k(x)dx = 0 \text{ for } i = 1, \ldots, m.$$

Let F_h be the kernel estimator

$$F_h(x) = h^{-1}\int F_n(t)k\left(\frac{x-t}{h}\right)dt,$$

where $h = h(n)$. Let $\hat{F}_n$ be the MPD type estimator corresponding to the sequence of smoothing parameters $\beta(n)$. From Theorem 2.1 of Gajek [7] we infer that if $h(n) = C_1\beta(n)^{1/(m+1)}$ with

$$C_1 = \left[i\,\|k^{(i)}\|_1(m-i)!/\int |v|^{m+1-i}|k(v)|\,dv\right]^{1/(m+1)}, \tag{4}$$

then for $i = 1, \ldots, m-1$

$$\mathrm{E}\|\hat{F}_n^{(i)} - F_h^{(i)}\|_1 \leqslant C_2 \beta(n)^{-i/(m+1)} \mathrm{E} d_n(\hat{F}_n),$$

where $d_n(F) = \|F_n - F\|_1 + \beta(n)\|F^{(m+1)}\|_1$ and C_2 is a constant independent of both n and F, involving the kernel k in the following way:

$$C_2 = \frac{m+1}{m+1-i}\|k^{(i)}\|_1 C_1^{-i}. \tag{5}$$

Hence, applying the triangle inequality, we get

$$\mathrm{E}\|\hat{F}_n^{(i)} - F^{(i)}\|_1 \leqslant C_2 \beta(n)^{-i/(m+1)} \mathrm{E} d_n(\hat{F}_n) + \mathrm{E}\|F_h^{(i)} - F^{(i)}\|_1.$$

Since $d_n(\hat{F}_n) \leqslant d_n(F_h)$, we have

$$\mathrm{E}\|\hat{F}_n^{(i)} - F^{(i)}\|_1 \leqslant C_2 \beta(n)^{-i/(m+1)} \mathrm{E} d_n(F_h) + \mathrm{E}\|F_h^{(i)} - F^{(i)}\|_1. \tag{6}$$

We shall evaluate the right-hand side of (6). Let us observe that, under the conditions imposed on k, the following identities hold:

$$F_h^{(i)}(x) = h^{-i-1}\int k^{(i)}\left(\frac{x-t}{h}\right)F(t)dt + \int\int_0^{hv}\frac{(z-hv)^{m-i}}{(m-i)!}F^{(m+1)}(x-z)k(v)dz\,dv \tag{7}$$

and

$$F_h^{(i)}(x) = h^{-i-1}\int k^{(i)}\left(\frac{x-t}{h}\right)F_n(t)dt. \tag{8}$$

Since k is $(m+1)$-times differentiable and vanishes outside some interval, it follows from (7) and (8) that

$$\begin{aligned}\mathrm{E}|F^{(i)}(x) - F_h^{(i)}(x)| \leqslant\ & h^{-1}\mathrm{E}\left|\int [F(x-hv) - F(x) - F_n(x-hv) + F_n(x)]k^{(i)}(v)dv\right| \\ & + \left|\int\int_0^{hv}\frac{(z-hv)^{m-i}}{(m-i)!}F^{(m+1)}(x-z)k(v)dz\,dv\right|.\end{aligned} \tag{9}$$

Now, observe that

$$\begin{aligned}\mathrm{E}|F_n(x-hv) - F_n(x) - F(x-hv) + F(x)| &\leqslant \{\mathrm{Var}[|F_n(x-hv) - F_n(x)|]\}^{1/2} \\ &\leqslant n^{-1/2}|F(x-hv) - F(x)|^{1/2}.\end{aligned} \tag{10}$$

From (9), (10) and (3) we get

$$\begin{aligned}\int \mathrm{E}|F^{(i)}(x) - F_h^{(i)}(x)|dx \leqslant\ & h^{-i+1/2}n^{-1/2}L\int |v|^{1/2}|k^{(i)}(v)|dv \\ & + h^{m+1-i}\|F^{(m+1)}\|_1\frac{\int |v|^{m+1-i}|k(v)|dv}{(m+1-i)!}.\end{aligned} \tag{11}$$

Now, we evaluate $\mathrm{E}d_n(F_h)$. Since k has m vanishing moments, using Taylor's series expansion, we get

$$\int [F(x-hv)-F(x)]k(v)dv = -\int \int_0^{hv} \frac{(z-hv)^m}{m!} F^{(m+1)}(x-z)k(v)dz\,dv,$$

and therefore

$$(12)\qquad |F_h(x)-F_n(x)| = \Big|\int [F_n(x-hv)-F_n(x)-F(x-hv)+F(x)]k(v)dv + \int [F(x-hv)-F(x)]k(v)dv\Big|$$
$$\leqslant \int |F_n(x-hv)-F_n(x)-F(x-hv)+F(x)|\,|k(v)|dv + \left|\int \int_0^{hv} \frac{(z-hv)^m}{m!} F^{(m+1)}(x-z)k(v)dz\,dv\right|.$$

Now, using (10) and (3), we get

$$(13)\qquad \int \mathrm{E}|F_h(x)-F_n(x)|dx \leqslant n^{-1/2}h^{1/2}L\int |v|^{1/2}|k(v)|dv + h^{m+1}\|F^{(m+1)}\|_1 \frac{\int |v|^{m+1}|k(v)|dv}{(m+1)!}.$$

Observe that

$$(14)\qquad F_h^{(m+1)}(x) = h^{-m-1}\int F_n(x-hv)k^{(m+1)}(v)dv$$
$$= h^{-m-1}\int [F_n(x-hv)-F_n(x)-F(x-hv)+F(x)]k^{(m+1)}(v)dv + h^{-m-1}\int [F(x-hv)-F(x)]k^{(m+1)}(v)dv.$$

Since k vanishes outside some interval and F and k are $(m+1)$-times differentiable functions, we obtain

$$(15)\qquad \int [F(x-hv)-F(x)]k^{(m+1)}(v)dv = h^{m+1}\int F^{(m+1)}(x-hv)k(v)dv.$$

From (14), (15) and (10) it follows that

$$\int \mathrm{E}|F_h^{(m+1)}(x)|dx \leqslant h^{-m-1}n^{-1/2}\int\int |F(x-hv)-F(x)|^{1/2}|k^{(m+1)}(v)|dv\,dz + \|F^{(m+1)}\|_1 \int |k(v)|dv.$$

Hence, applying (3), we get

$$(16)\qquad \int \mathrm{E}|F_h^{(m+1)}(x)|dx \leqslant n^{-1/2}h^{-m-1/2}L\int |v|^{1/2}|k^{(m+1)}(v)|dv + \|F^{(m+1)}\|_1 \int |k(v)|dv.$$

Finally, from (6), (11), (13) and (16) it follows that

$$\mathrm{E}\|\hat{F}_n^{(i)}-F^{(i)}\|_1 \leqslant C_2\beta^{-i/(m+1)}\Big[n^{-1/2}h^{1/2}L\int|v|^{1/2}|k(v)|\,dv + h^{m+1}\|F^{(m+1)}\|_1\frac{\int|v|^{m+1}|k(v)|\,dv}{(m+1)!} + \beta\big(n^{-1/2}h^{-m-1/2}L\int|v|^{1/2}|k^{(m+1)}(v)|\,dv+\|F^{(m+1)}\|_1\int|k(v)|\,dv\big)\Big] + n^{-1/2}h^{-i+1/2}L\int|v|^{1/2}|k^{(i)}(v)|\,dv+h^{m+1-i}\|F^{(m+1)}\|_1\frac{\int|v|^{m+1-i}|k(v)|\,dv}{(m+1-i)!}.$$

Since $h = C_1\beta^{1/(m+1)}$, we get

$$\mathrm{E}\|\hat{F}_n^{(i)}-F^{(i)}\|_1 \leqslant \beta^{-i/(m+1)}[H_1(\beta^{1/(m+1)}/n)^{1/2}+H_2\beta],$$

where

(17) $$H_1 = LC_1^{1/2}\int|v|^{1/2}[C_2|k(v)|+C_1^{-i}C_2|k^{(i)}(v)|+C_1^{-m-1}|k^{(m+1)}(v)|]\,dv$$

and

(18) $$H_2 = \|F^{(m+1)}\|_1\left(\frac{C_1^{m+1}\int|v|^{m+1}|k(v)|\,dv}{(m+1)!}+\int|k(v)|\,dv\right)C_2 + \frac{C_1^{m+1-i}\int|v|^{m+1-i}|k(v)|\,dv}{(m+1-i)!},$$

with C_1 and C_2 given by (4) and (5). ∎

Since Theorem 2.3 can be proved in a similar way, its proof is omitted.

Proof of Lemma 3.1. Applying the Cauchy inequality, we get

$$\int|F(x+z)-F(x)|^{1/2}dx = \int\Big|\int_0^y f(x+z)\,dz\Big|^{1/2}dx \geqslant |y|^{-1/2}\int\Big|\int_0^y f^{1/2}(x+z)\,dz\Big|dx \geqslant |y|^{-1/2}\Big|\int_0^y\big(\int f^{1/2}(x+z)\,dx\big)dz\Big| = |y|^{1/2}\int\sqrt{f}.$$

Hence, if (3) holds for some L, then $\int\sqrt{f}\leqslant L$. To prove the converse, let us notice that

$$\int|F(x+y)-F(x)|^{1/2}dx = \int_{|x|\leqslant T}\Big|\int_0^y f(x+z)\,dz\Big|^{1/2}dx + \int_{|x|>T}\Big|\int_0^y f(x+z)\,dz\Big|^{1/2}dx$$
$$\leqslant (2T)^{1/2}\Big(\int_{|x|\leqslant T}\Big|\int_0^y f(x+z)\,dz\Big|dx\Big)^{1/2} + \int_{|x|>T}\Big|\int_0^y \sup_{|x-v|<|y|} f(v)\,dz\Big|^{1/2}dx$$
$$\leqslant |y|^{1/2}\Big[(2T)^{1/2}+\int_{|x|>T}\sqrt{\sup_{|x-v|<|y|} f(v)}\,dx\Big].$$

So, if for some $T > 0$ and $t > 0$

$$\int_{|x|>T} \sqrt{\sup_{|x-v|<t} f(v)}\, dx < \infty, \tag{19}$$

then (3) holds with

$$L = \sqrt{2T} + \int_{|x|>T} \sqrt{\sup_{|x-v|<t} f(v)}\, dx.$$

Now, observe that if f is unimodal, bounded and $\int \sqrt{f} < \infty$, then (19) holds true for all positive t and T. ∎

REFERENCES

[1] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation*, Z. Wahrsch. Verw. Gebiete 65 (1983), pp. 181–237.

[2] L. Devroye, *A Course in Density Estimation*, Birkhäuser, Boston 1987.

[3] – and L. Gyorfi, *Nonparametric Density Estimation. The L_1 View*, Wiley, New York 1985.

[4] D. L. Donoho and R. C. Liu, *The "authomatic" robustness of minimum distance functionals*, Ann. Statist. 16 (1988), pp. 552–587.

[5] – *Pathologies of some minimum distance estimators*, ibidem 16 (1988), pp. 587–608.

[6] L. Gajek, *Nonparametric estimation of a density function and its derivatives via the minimum distance method* (in Polish), Thesis 103, Scient. Bull. Łódź Techn. University, No 533 (1987).

[7] – *Estimating a density and its derivatives via the minimum distance method*, Probab. Theory Related Fields 80 (1989), pp. 601–617.

[8] – and M. Kałuszka, *Upper bounds for the L_1-risk of the minimum L_1-distance regression estimator*, Ann. Inst. Statist. Math. (1991).

[9] R.-D. Reiss, *Sharp rates of convergence of minimum penalized distance estimators*, Sankhyā, Ser. A, 48 (1986), pp. 59–68.

[10] A. P. Stefanyuk, *Convergence rate of a class of probability density estimates* (in Russian), Avtomat. i Telemekh. 11 (1979), pp. 187–192.

[11] V. Vapnik and A. P. Stefanyuk, *Nonparametric methods of probability density recovery* (in Russian), ibidem 8 (1978), pp. 38–52.

[12] Y. G. Yatracos, *Rates of convergence of minimum distance estimators and Kolmogorov's entropy*, Ann. Statist. 13 (1985), pp. 768–774.

[13] – *A note on L_1-consistent estimation*, Canad. J. Statist. 16 (1988), pp. 283–292.

Technical University of Łódź
Institute of Mathematics
Al. Politechniki 11
90-924 Łódź, Poland

Received on 29.5.1991;
revised version on 5.3.1992