T E C H N I C A L

R E P O R T

0304

LINEAR PROGRAMMING PROBLEMS FOR FRONTIER ESTIMATION

G. BOUCHARD, S. GIRARD, A. IOUDITSKI and A. NAZIN

I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Abstract

We propose new estimates for the frontier of a set of points. They are defined as kernel estimates covering all the points and whose associated support is of smallest surface. The estimates are written as linear combinations of kernel functions applied to the points of the sample. The coefficients of the linear combination are then computed by solving a linear programming problem. In the general case, the solution of the optimization problem is sparse, that is, only a few coefficients are non zero. The corresponding points play the role of support vectors in the statistical learning theory. The L_1 error between the estimated and the true frontiers is shown to be almost surely converging to zero, and the rate of convergence is provided. The behaviour of the estimates on finite sample situations is illustrated on some simulations.

Contact information

Guillaume Bouchard: IS2, INRIA, ZIRST, 655, av. de l'Europe, Montbonnot,
38334 Saint-Ismier cedex, France. `Guillaume.Bouchard@inrialpes.fr`

Stéphane Girard and Anatoli Iouditski: SMS/LMC, Université Grenoble I, BP 53,
38041 Grenoble cedex 9, France. `Stephane.Girard@imag.fr`, `Anatoli.Iouditski@imag.fr`

Alexander Nazin: Institute of Control Sciences, RAS, Profsoyuznaya str., 65,
117997 Moscow, Russia. `nazine@ipu.rssi.ru`

Acknowledgements

Financial support from the IAP research network nr P5/24 of the Belgian Government (Federal Office for Scientific, Technical and Cultural Affairs) is gratefully acknowledged.

The work of A. Nazin has been carried out during his stay in INRIA Rhône-Alpes as invited professor, November-December 2002.

1 Introduction

Many proposals are given in the literature for estimating a set S given a finite random set of points drawn from the interior. This problem of edge or support estimation arises in classification (Hardy & Rasson [24]), clustering problems (Hartigan [25]), discriminant analysis (Baufays & Rasson [3]), and outliers detection. Applications are found in medical diagnosis (Tarassenko *et al* [32]) as well as in condition monitoring of machines (Devroye & Wise [11]). In image analysis, the segmentation problem can be considered under the support estimation point of view, where the support is a convex bounded set in $\mathbb{R}^2$ (Korostelev & Tsybakov [30]). We also point out some applications in econometrics (e.g. Deprins, *et al* [10]). In such cases, the unknown support can be written

$$S \triangleq \{(x, y) : \quad 0 \leq \ x \leq \ 1 \ \ ; \quad 0 \leq \ y \ \leq \ f(x)\}, \tag{1}$$

where f is an unknown function. Here, the problem reduces to estimating f, called the production frontier (see for instance Härdle *et al* [21]). The data consist of pair (X, Y) where X represents the input (labor, energy or capital) used to produce an output Y in a given firm. In such a framework, the value $f(x)$ can be interpreted as the maximum level of output which is attainable for the level of input x. Korostelev *et al* [29] suppose f to be increasing and concave, from economical considerations, which suggests an adapted estimator, called the DEA (Data Envelopment Analysis) estimator. It is the lowest concave monotone increasing function covering all the sample points. Therefore it is piecewise linear and, up to our knowledge, it is the first frontier estimate computed thanks to a linear programming technique (Charnes *et al* [7]). Its asymptotic distribution is established by Gijbels *et al* [15].
An early paper was written by Geffroy [13] for independent identically distributed observations from a density ϕ. The proposed estimator is a kind of histogram based on the extreme values of the sample. This work was extended in two main directions.

On the one hand, piecewise polynomials estimates were introduced. They are defined locally on a given slice as the lowest polynomial of fixed degree covering all the points in the considered slice. Their optimality in an asymptotic minimax sense is proved under weak assumptions on the rate of decrease α of the density ϕ towards 0 by Korostelev & Tsybakov [30] and by Härdle *et al* [22]. Extreme values methods are then proposed by Hall *et al* [20] and by Gijbels & Peng [14] to estimate the parameter α.

On the other hand, different propositions for smoothing Geffroy's estimate were made in the case of a Poisson point process. Girard & Jacob [18] introduced estimates based on kernel regressions and orthogonal series method [16, 17]. In the same spirit, Gardes [12] proposed a Faber-Shauder estimate. Girard & Menneteau [19] introduced a general framework for studying estimates of this type and generalized them to supports writting

$$S = \{(x, y) : \quad x \in E \ \ ; \quad 0 \leq \ y \ \leq \ f(x)\},$$

where f is an unknown function and E an arbitrary set. In each case, the limit distribution of the estimator is established.
We also refer to Abbar [1] and Jacob & Suquet [28] who used a similar smoothing approach, although their estimates are not based on the extreme values of the Poisson process.

The estimate proposed in this paper can be considered to belong to the intersect of these two directions. It is defined as a kernel estimate obtained by smoothing some selected points of the sample. These points are chosen automatically by solving a linear programming problem to obtain an estimate of the support covering all the points and with smallest surface. Its advantages are the following: it can be computed with standard optimization algorithms (see e.g. BONNANS *et al* [5], chapter 4), its smoothness is directly linked to the smoothness of the chosen kernel and it benefits from interesting theoretical properties. Here, we prove that it is almost surely convergent for the L_1 norm. The estimate is defined in Section 2. Its theoretical properties are established in Section 3. The behaviour of the estimate is illustrated in Section 4 on finite sample situations. Its compared to a similar proposition found in BARRON *et al* [2]. Proofs are postponed to Section 5.

2 Boundary estimates

2.1 A linear programming problem

Let all the random variables be defined on a probability space $(\Omega, \mathcal{F}, P)$. The problem under consideration is to estimate an unknown positive function $f : [0,1] \to (0,\infty)$ on the basis of observations $Z_N = (X_i, Y_i)_{i=1,\dots,N}$. The former represents an i.i.d. sequence with pairs (X_i, Y_i) being uniformly distributed in the set S defined as in (1). For the sake of simplicity, we consider in the following the extension of f on all $\mathbb{R}$ by introducing $f(x) = 0$ for all $x \notin [0,1]$. Letting

$$C_f \triangleq \int_0^1 f(u)\,du = \int_{\mathbb{R}} f(u)\,du,$$

each variable X_i is distributed in $[0,1]$ with p.d.f. $f(\cdot)/C_f$ while Y_i has the uniform conditional distribution with respect to X_i in the interval $[0, f(X_i)]$.

The considered estimate of the frontier is chosen from the family of functions:

$$\begin{cases} \widehat{f}_N(x) = \sum_{i=1}^{N} K_h(x - X_i)\alpha_i\,, \qquad K_h(t) = h^{-1}K(t/h), \\ \alpha_i \geq 0, \qquad i = 1,\dots,N, \end{cases} \tag{2}$$

where K is a kernel function $K : \mathbb{R} \to [0,\infty)$ integrating to one and with bandwidth $h > 0$. Each coefficient α_i represents the importance of the point (X_i, Y_i) in the estimation. In particular, if $\alpha_i \neq 0$, the corresponding point (X_i, Y_i) can be called a support vector by analogy with Support Vector Machines (SVM). We refer to CRISTIANINI & SHAWE-TAYLOR [9] for a review on this topic and to SCHÖLKOPF & SMOLA [31], chapter 8, for examples of application of SVM to quantile estimation. The constraint $\alpha_i \geq 0$ for all $i = 1,\dots,N$ ensures that $\widehat{f}_N(x) \geq 0$ for all $x \in \mathbb{R}$ and prevents the estimator from being too irregular (see equation (50)). Let us remark that the surface of the estimated support is given by

$$\int_{\mathbb{R}} \widehat{f}_N(x)\,dx = \sum_{i=1}^{N} \alpha_i.$$

This suggests to define the vector parameter $\alpha = (\alpha_1, \dots, \alpha_N)^T$ from a linear program as follows

$$\begin{aligned} J_P^* &\triangleq \min_\alpha \mathbf{1}^T \alpha && (3) \\ \text{subject to} & && \\ & A\alpha \geq Y && (4) \\ & \alpha \geq 0. && (5) \end{aligned}$$

The following notations have been introduced:

$$\begin{aligned} \mathbf{1} &\triangleq (1, 1, \dots, 1)^T \in \mathbb{R}^N \\ A &\triangleq \|K_h(X_i - X_j)\|_{i,j=1,\dots,N} \\ Y &\triangleq (Y_1, \dots, Y_N)^T. \end{aligned}$$

Hence, $A\alpha = (\widehat{f}_N(X_1), \dots, \widehat{f}_N(X_N))^T$, and the vector constraint (4) means that

$$\widehat{f}_N(X_i) \geq Y_i\,, \qquad i = 1, \dots, N. \tag{6}$$

In other words, $\widehat{f}_N$ defines the kernel estimate of the support covering all the points and with smallest surface. In practice (see Section 4 for an illustration) the solution of the linear program is sparse in the sense that $n(\alpha) = \text{Card}\{\alpha_i \neq 0\}$ is small (for moderate values of h) and thus the resulting estimate is fast to compute even for large samples.

Let us note that the above described estimator (2)–(5) might be derived as the Maximum Likelihood Estimate related to the approximation family (2). Indeed, the joint probability density function for observations Z_N given parameter function $f(x)$ can be written

$$p(Z_N \mid f) = \prod_{i=1}^N \frac{f(X_i)}{C_f} \cdot \frac{1}{f(X_i)} \mathbf{1}\{0 \leq Y_i \leq f(X_i)\}, \tag{7}$$

where $\mathbf{1}\{.\}$ is the indicator function. Moreover,

$$C_f \Big|_{f=\widehat{f}_N} = \sum_{i=1}^N \alpha_i, \tag{8}$$

and therefore, the Log-Likelihood function is

$$L(\alpha) \triangleq \log p(Z_N \mid \widehat{f}_N) = -N \log \sum_{i=1}^N \alpha_i + \sum_{i=1}^N \log \mathbf{1}\{Y_i \leq \widehat{f}_N(X_i)\}, \tag{9}$$

and its maximization over the set of non-negative parameters α is equivalent to problem (3)–(5).

2.2 Comparison with other methods

Let us remark that other solutions for estimating α in (2) have already been proposed. Girard & Menneteau [19] considered a partition $\{I_r : \ 1 \leq r \leq k\}$ of $[0,1]$, with $k \to \infty$. For all $1 \leq r \leq k$, they introduce $D_r = \{(x,y) : x \in I_r,\ 0 \leq y \leq f(x)\}$, the slice of S built on I_r, $Y_r^* = \max\{Y_i; (X_i, Y_i) \in D_r\}$, and the estimates

$$\widehat{\alpha}_i = \left| \begin{array}{ll} \lambda(I_r)Y_r^* & \text{if } \exists\, r \in \{1,\dots,k\}\,;\, Y_i = Y_r^* \\ 0 & \text{otherwise,} \end{array} \right.$$

where λ is the Lebesgue measure. They propose the following frontier estimate

$$\check{f}_N(x) = \sum_{r=1}^{k} K_h(x - x_r)\lambda(I_r)Y_r^*,$$

where x_r is the center of I_r. This approach suffers from a practical difficulty: the choice of the partition and more precisely the choice of k. In our context, solving the linear problem (3)–(5) direcly yields the support vectors.

In this sense, the estimate proposed in BARRON *et al* [2] is similar to $\widehat{f}_N$. It is defined by the Fourier expansion:

$$\widehat{g}_N(x) = c_0 + \sum_{k=1}^{M} a_k \cos(2\pi kx) + \sum_{k=1}^{M} b_k \sin(2\pi kx), \tag{10}$$

where the vector of parameters $\beta = (c_0, a_1, \dots, a_M, b_1, \dots, b_M)^T$ is solution of the linear programming problem:

$$\min c_0 \quad \left(= \int_0^1 \widehat{g}_N(x)dx \right) \tag{11}$$

under the constraints

$$\widehat{g}_N(X_i) \;\geq\; Y_i, \quad i = 1, \dots, N \tag{12}$$

$$\sum_{k=1}^{M} k(|a_k| + |b_k|) \;\leq\; L/(2\pi). \tag{13}$$

Therefore, $\widehat{g}_N$ defines the Fourier estimate of the support covering all the points (equation (12)), L-Lipschitzian (equation (13)) and with smallest surface (equation (11)). From the theoretical point of view, this estimate benefits from minimax optimality. It is compared to $\widehat{f}_N$ on practical situations in Section 4 for different choices of parameters M, L and h.

3 Main results

The basic assumptions on the unknown boundary function are:

A1. $0 < f_{\min} \leq f(x) < f_{\max} < \infty$, for all $x \in [0,1]$,

A2. $|f(x) - f(y)| \leq L_f|x - y|$, for all $x, y \in [0,1]$, $\quad L_f < \infty$.

The following assumptions on the kernel function are introduced:

B1. $K(t) = K(-t) \geq 0$,

B2. $\displaystyle\int_{\mathbb{R}} K(t)\, dt = 1$,

B3. $|K(s) - K(t)| \leq L_K|s - t|$, $\;L_K < \infty$, B4. $C_0(K) \triangleq \int_{\mathbb{R}} K^2(t)\,dt < \infty$ and $C_2(K) \triangleq \int_{\mathbb{R}} t^2\, K(t)\,dt < \infty$.

We denote $K_{\max} \triangleq \max K(t)$. In the following theorem the consistency of the estimate is established with respect to the L_1 norm on the $[0,1]$ interval.

Theorem 1 *Let $h \to 0$ and $\log N/(Nh^2) \to 0$ as $N \to \infty$. Let the above mentioned assumptions A and B hold true. Then estimator (2)–(5) has the following asymptotic properties:*

$$\limsup_{N\to\infty} \varepsilon_1^{-1}(N)\|\hat{f}_N - f\|_1 \le C(\omega) < \infty \quad \text{a.s.} \tag{14}$$

with

$$\varepsilon_1(N) \triangleq \max\left\{h,\, \sqrt{\log N/(Nh^2)}\right\}. \tag{15}$$

Corollary 1 *The maximum rate of convergence which is guaranteed by Theorem 1*

$$\|\hat{f}_N - f\|_1 = O\left((\log N/N)^{1/4}\right)$$

is attained for

$$h \asymp (\log N/N)^{1/4}. \tag{16}$$

This rate of convergence can be ameliorated at the price of a slight modification of the estimate. In the following, an additional constraint is considered in order to impose to each coefficient α_i to be of order $1/N$. The counterpart of this modification is that the new estimate $\tilde{f}_N$ will usually rely on more support vectors than $\widehat{f}_N$.
Let us modify the estimator (2)–(5) as follows.

$$\tilde{f}_N(x) = \sum_{i=1}^{N} K_h(x - X_i)\alpha_i \tag{17}$$

where vector $\alpha = (\alpha_1, \ldots, \alpha_N)^T$ is defined from the Modified Linear Program

$$J^*_{MP} \triangleq \min_{\alpha} \mathbf{1}^T\alpha \tag{18}$$

subject to

$$A\alpha \ge Y \tag{19}$$

$$0 \le \alpha \le C_\alpha/N \tag{20}$$

with a constant

$$C_\alpha > f_{\max}. \tag{21}$$

Remark. In fact, we need to ensure $C_\alpha > C_f$ which is implied by (21).

The modified estimator (17)–(21) differs from that of (2)–(5) by additionally bounding each α_i from above, see constraints (20). Below we prove that under condition (21) as well as finite support kernel $K(\cdot)$ the Modified Linear Program (18)–(20) has a nonempty set of admissible solutions with the same upper bound as (25) and a better lower bound than (40).

Theorem 2 *Let $h \to 0$ and $\log N/(Nh) \to 0$ as $N \to \infty$. Let kernel function $K(\cdot)$ has a finite support, that is $K(t) = 0 \; \forall \, |t| \geq 1$, and the assumptions A and B hold true. Then estimator (17)–(21) has the following asymptotic properties:*

$$\limsup_{N \to \infty} \varepsilon_2^{-1}(N) \|\tilde{f}_N - f\|_1 \leq C(\omega) < \infty \quad \text{a.s.} \tag{22}$$

with

$$\varepsilon_2(N) \triangleq \max\left\{h, \sqrt{\log N/(Nh)}\right\}. \tag{23}$$

Remark. The support of $K(\cdot)$ is fixed to be the interval $[-1, 1]$ without loss of generality.

Corollary 2 *The maximum rate of convergence which is guaranteed by Theorem 2*

$$\|\tilde{f}_N - f\|_1 = O\left((\log N/N)^{1/3}\right)$$

is attained for

$$h \asymp (\log N/N)^{1/3}. \tag{24}$$

4 Numerical experiments

The simulations presented here illustrate the behaviour of the kernel estimator $\widehat{f}_N$ compared to the estimator based on Fourier expansions $\widehat{g}_N$ proposed in Barron *et al* [2]. Since the Fourier estimator $\widehat{g}_N$ requires the unknown function to be periodic, we choose f such that $f(0) = f(1)$. Besides, to avoid boundary effects on the input domain, we consider functions that are nearly zero when x is close to 0 or 1. In more general situation, boundary corrections should be implemented (see Cowling & Hall [8]). The chosen function

$$\begin{aligned} f(x) = 0.1 \quad &+ \quad 5(x-0.1)\mathbf{1}_{\{x>0.1\}} \\ &- \quad 5(x-0.2)\mathbf{1}_{\{x>0.2\}} \\ &+ \quad 1(x-0.5)\mathbf{1}_{\{x>0.5\}} \\ &- \quad 9(x-0.8)\mathbf{1}_{\{x>0.8\}} \\ &+ \quad 8(x-0.9)\mathbf{1}_{\{x>0.9\}}, \end{aligned}$$

is piecewise linear and locally Lipschitizian with a Lipschitz constant $L_f = 8$. Its graph is given on the right part of Figure 1 (dotted line). Here, $N = 50$ points are uniformly generated on the domain S upper bounded by f (left part of Figure 1). On the right, the kernel estimate (continuous line) is superimposed to the unknown function f. The squares represent the points for which $\widehat{f}_N(X_i) = Y_i$. The triangles represent the support vectors (i.e. the points for which $\alpha_i > 0$). On this example, the estimate $\widehat{g}_N$ is not visually different from $\widehat{f}_N$.

Thus, a more precise comparison should be done. For each estimate, the L_1 error Δ_N as well as the number of effective parameters np (that is n_α and $n_\beta = \text{Card}\{\beta_i \neq 0\}$) are evaluated for $N = 25$ and $N = 100$. The average value and the standard deviation of these quantities are computed on 1000 replications in the first case and on 100 replications in the second one. The estimation is carried out with different values of the parameters,

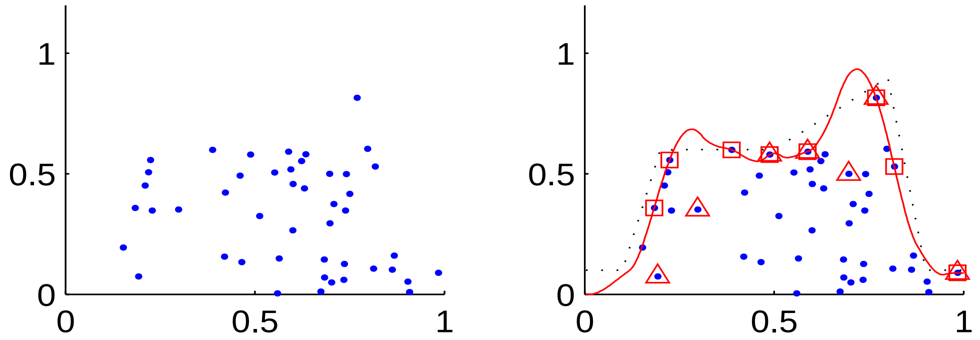

Figure 1: Illustration of the kernel estimation on a simulation.

namely h for the kernel estimate, and L and M for the Fourier estimate. The adaptive choice of these parameters is not implemented in this setting. The results are summarized in Tables 1 and 2. The lowest error is emphasized for each estimate. It can be noted that the mean L_1 error of both estimates are very similar. In fact, the kernel estimate seems to give a slight lower error for small number of points and the Fourier estimate yields better results for large sample size situations, confirming its asymptotic optimality. Let us note that the standard deviation of the L_1 error is in general smaller for the kernel estimate. Regarding the number of parameters, the kernel estimate seems to be more parsimonious than the Fourier estimate.

5 Proofs

The proof of Theorem 1 which is given in subsection 5.3 is based on both upper and lower bounds derived below.

5.1 Upper bound for $\widehat{f}_N$

Lemma 1 *Let $h \to 0$ and $\log N/(Nh) \to 0$ as $N \to \infty$. Let the above mentioned assumptions A and B hold true. Then for almost all $\omega \in \Omega$ there exist finite number $N_0(\omega)$ such that*

$$J_P^* \leq C_f + O(h) + O\left(\sqrt{\log N/(Nh)}\right), \qquad \forall\, N \geq N_0(\omega), \tag{25}$$

with non random both $O(h)$ and $O\left(\sqrt{\log N/(Nh)}\right)$.

Proof of Lemma 1. 1. Since kernel function $K(.)$ is supposed to be even then matrix A is symmetric, and the dual problem associated to (3) – (5) can be written:

$$J_D^* \triangleq \max_{\lambda} Y^T \lambda \tag{26}$$

subject to

$$A\lambda \leq \mathbf{1} \tag{27}$$

$$\lambda \geq 0. \tag{28}$$

Let us replace vector Y in (26) for

$$F \triangleq (f(X_1), \ldots, f(X_N))^T \tag{29}$$

estimate	h	L	M	mean(Δ_N)	st-dev(Δ_N)	mean(np)	st-dev(np)
kernel	0.100			0.123	0.038	5.263	0.970
	0.120			0.116	0.034	4.490	0.841
	0.140			**0.112**	**0.033**	**3.841**	**0.683**
	0.160			0.115	0.031	3.420	0.636
	0.180			0.123	0.027	3.120	0.657
	0.200			0.132	0.023	2.863	0.645
Fourier		3.000	4.000	0.144	0.035	4.567	0.777
		5.000	**4.000**	**0.119**	**0.043**	**5.508**	**0.986**
		7.000	4.000	0.127	0.043	6.572	1.217
		9.000	4.000	0.138	0.044	7.235	1.284
		11.000	4.000	0.147	0.046	7.592	1.249
		13.000	4.000	0.154	0.046	7.815	1.210
Fourier		3.000	8.000	0.144	0.036	4.581	0.800
		5.000	**8.000**	**0.121**	**0.044**	**5.571**	**1.057**
		7.000	8.000	0.129	0.044	6.730	1.379
		9.000	8.000	0.142	0.046	7.632	1.669
		11.000	8.000	0.153	0.047	8.314	1.873
		13.000	8.000	0.163	0.048	8.859	2.050

Table 1: Results for 1000 simulations with $N = 25$ points.

estimate	h	L	M	mean(Δ_N)	st-dev(Δ_N)	mean(np)	st-dev(np)
kernel	0.050			0.073	0.016	13.700	1.560
	0.070			0.060	0.014	9.890	1.246
	0.090			**0.060**	**0.013**	**7.350**	**1.132**
	0.110			0.063	0.012	5.820	0.989
	0.130			0.075	0.012	4.690	0.734
	0.150			0.085	0.013	3.960	0.549
Fourier		3.000	4.000	0.129	0.021	5.120	0.700
		5.000	4.000	0.078	0.020	5.790	0.756
		7.000	**4.000**	**0.061**	**0.012**	**7.630**	**0.960**
		9.000	4.000	0.064	0.013	8.700	0.560
		11.000	4.000	0.069	0.015	8.880	0.409
		13.000	4.000	0.071	0.016	8.950	0.297
Fourier		3.000	8.000	0.129	0.021	5.160	0.762
		5.000	8.000	0.078	0.020	5.920	0.849
		7.000	8.000	0.059	0.013	8.070	1.350
		9.000	**8.000**	**0.059**	**0.015**	**10.470**	**1.630**
		11.000	8.000	0.063	0.015	12.090	1.682
		13.000	8.000	0.069	0.015	13.620	2.068

Table 2: Results for 100 simulations with $N = 100$ points.

and, moreover, change the vector constraint (27) by a scalar one which is directly obtained by just summing all N rows of (27). Thus, we arrive at the modified dual problem

$$J_{MD}^* \triangleq \max_{\lambda} F^T \lambda \tag{30}$$

subject to

$$\mathbf{1}^T A\lambda \leq N \tag{31}$$

$$\lambda \geq 0. \tag{32}$$

Since $F \geq Y$ and according to the well known Duality Theorem (see e.g. Hiriart-Urruty & Lemaréchal [26], chapter 7):

$$J_P^* = J_D^* \leq J_{MD}^*. \tag{33}$$

Now we derive an upper bound on J_{MD}^*.

2. Let us arbitrarily fix a vector λ which meet the constraints (31), (32) and then write inequality (31) in the equivalent form as follows:

$$\frac{1}{N}\sum_{j=1}^{N} \lambda_j \left(K_h(0) + \sum_{i \neq j}^{N} K_h(X_i - X_j) \right) \leq 1, \tag{34}$$

or, equivalently,

$$\frac{1}{N}\sum_{j=1}^{N} \lambda_j \left(\frac{1}{h}K(0) + \sum_{i \neq j}^{N} E\left\{ K_h(X_i - X_j) \mid X_j \right\} + \sum_{i \neq j}^{N} \xi_{ij} \right) \leq 1, \tag{35}$$

with

$$\xi_{ij} \triangleq K_h(X_i - X_j) - E\left\{ K_h(X_i - X_j) \mid X_j \right\}.$$

Now apply upper bound (96), proved in Lemma 5 (see Appendix), to the relation (35) taking into account that $K(0) > 0$ and

$$\begin{aligned} E\left\{ K_h(X_i - X_j) \mid X_j \right\} &= \frac{1}{h}\int_0^1 K\left(\frac{u - X_j}{h}\right) \frac{f(u)}{C_f}\, du & (36) \\ &= \frac{1}{C_f}\int_{\mathbb{R}} K(t) f(X_j + ht)\, dt \\ &= \frac{1}{C_f}\left(f(X_j) + O(h) \right), & (37) \end{aligned}$$

with non random $O(h)$. Hence,

$$\frac{N-1}{C_f N}\sum_{j=1}^{N} \lambda_j \left(f(X_j) + O(h) - C\sqrt{\frac{\log N}{Nh}} \right) \leq 1, \qquad \forall\, N \geq N_2(\omega), \tag{38}$$

with non random constant C. First, inequality (38) implies

$$\sum_{j=1}^{N} \lambda_j \leq \frac{2C_f}{f_{\min}} < \infty, \qquad \forall\, N \geq N_3(\omega), \tag{39}$$

with almost surely finite $N_3(\omega) \geq N_2(\omega)$. Second, (39) and (38) imply upper bound (25) and Lemma 1 is proved. ■

5.2 Lower bound for $\widehat{f}_N$

Lemma 2 *Under the assumptions of Theorem 1, for almost all $\omega \in \Omega$ there exist finite number $N_1(\omega)$ such that for each $x \in (0,1)$*

$$\widehat{f}_N(x) \geq f(x) - O\left(\sqrt{\log N/(Nh^2)}\right), \qquad \forall\, N \geq N_1(\omega), \tag{40}$$

where $O(\cdot)$ do not depend on x.

Proof of Lemma 2. 1. Suppose that for some non-random $\delta_x > 0$ there exists (with probability one) an integer $i_k \in \{1, \dots, N\}$ such that

$$|x - X_{i_k}| \leq \delta_x. \tag{41}$$

Then, the estimation error at a point $x \in (0,1)$ can be expanded as

$$\begin{aligned} f(x) - \widehat{f}_N(x) &= [f(x) - f(X_{i_k})] && (42)\\ &+ \left[f(X_{i_k}) - \widehat{f}_N(X_{i_k})\right] && (43)\\ &+ \left[\widehat{f}_N(X_{i_k}) - \widehat{f}_N(x)\right]. && (44) \end{aligned}$$

The term in the right hand side (42) may be bounded as follows

$$|f(x) - f(X_{i_k})| \leq L_f\,|x - X_{i_k}| \leq L_f \delta_x, \tag{45}$$

as well as the term (44)

$$\left|\widehat{f}_N(X_{i_k}) - \widehat{f}_N(x)\right| \leq L_{\widehat{f}_N}\,|x - X_{i_k}| \leq L_{\widehat{f}_N}\,\delta_x, \tag{46}$$

with a Lipschitz constant $L_{\widehat{f}_N}$ for the function estimate $\widehat{f}_N(x)$, which is bounded below. In order to bound (43) assume that for some non-random $\delta_y > 0$,

$$Y_{i_k} \geq f(X_{i_k}) - \delta_y \text{ a.s.} \tag{47}$$

Remind that $\widehat{f}_N(X_{i_k}) \geq Y_{i_k}$ due to (4) or (6). Thus,

$$f(X_{i_k}) - \widehat{f}_N(X_{i_k}) \leq (Y_{i_k} + \delta_y) - Y_{i_k} = \delta_y. \tag{48}$$

Combining all these bounds we obtain from (42) that for all $N \geq N_0(\omega)$,

$$f(x) - \widehat{f}_N(x) \leq \delta_y + \left(L_f + L_{\widehat{f}_N}\right)\delta_x. \tag{49}$$

2. Note that a straightforward evaluation of the Lipschitz constant for the estimate function yields:

$$\begin{aligned} |\widehat{f}_N(u) - \widehat{f}_N(v)| &\leq \sum_{i=1}^{N} \alpha_i\,|K_h(u - X_i) - K_h(v - X_i)| && (50)\\ &\leq \frac{L_K}{h^2}\left(\sum_{i=1}^{N} \alpha_i\right)|u - v|. && (51) \end{aligned}$$

Hence, due to the upper bound (25), we obtain almost surely

$$L_{\widehat{f}_N} = \frac{L_K}{h^2} C_f (1 + o(1)), \qquad \forall\, N \geq N_0(\omega), \tag{52}$$

with almost surely finite $N_0(\omega)$.
3. Now, we demonstrate that under appropriate definition of δ_x and δ_y as functions of h and N there exist almost surely finite random integer $N_0(\omega)$ such that

$$\forall\, N \geq N_0(\omega), \qquad \exists\, i_k \in \{1, \ldots, N \,:\, (X_{i_k}, Y_{i_k}) \in \Delta(x)\}, \tag{53}$$

with

$$\Delta(x) \triangleq \{(u, v) \,:\, |x - u| \leq \delta_x,\, f(x) - \delta_y \leq v \leq f(u)\}. \tag{54}$$

Indeed, introduce

$$\delta_y \triangleq \left(\frac{\kappa \log N}{N h^2} \right)^{1/2}, \tag{55}$$

and

$$\delta_x = h^2 \delta_y. \tag{56}$$

Then,

$$\begin{aligned} P\{(X_i, Y_i) \notin \Delta(x) \quad \forall\, i = 1, \ldots, N\} &= \left(1 - \frac{1 + o(1)}{C_f} \delta_x \delta_y \right)^N \\ &= \left(1 - \frac{1 + o(1)}{C_f} h^2 \delta_y^2 \right)^N \\ &\leq \exp\left\{ - \frac{1 + o(1)}{C_f} N h^2 \delta_y^2 \right\} \\ &\leq N^{-\kappa/(2C_f)}. \end{aligned} \tag{57}$$

Hence, fixing

$$\kappa > 2C_f \tag{58}$$

implies the convergence of the series

$$\sum_{N=1}^{\infty} P\{(X_i, Y_i) \notin \Delta(x) \quad \forall\, i = 1, \ldots, N\} < \infty, \tag{59}$$

which, due to Borel–Cantelly lemma, implies the existence of almost surely finite $N_0(\omega)$ such that relation (53) holds true.
4. Therefore, substituing relations (52), (55), and (56) to (49) leads to lower bound

$$\begin{aligned} \widehat{f}_N(x) &\geq f(x) - \delta_y - O\left(h^{-2}\right) \delta_x \\ &= f(x) - O\left(\sqrt{\frac{\log N}{N h^2}} \right), \end{aligned} \tag{60}$$

with non-random term $O(\cdot)$ independent of x. ∎

5.3 Proof of Theorem 1

1. Since $|u| = u - 2u\mathbf{1}\{u < 0\}$, the L_1-norm of estimation error can be expanded as

$$
\begin{aligned}
\|\hat{f}_N - f\|_1 &= \int_0^1 \left[\hat{f}_N(x) - f(x)\right] dx & (61)\\
&+2\int_0^1 \left[f(x) - \hat{f}_N(x)\right] \mathbf{1}\left\{\hat{f}_N(x) < f(x)\right\} dx. & (62)
\end{aligned}
$$

2. Applying Lemma 1 to the right hand side (61) yields

$$
\limsup_{N\to\infty} \varepsilon_{UB}^{-1}(N) \left(\int_0^1 \left[\hat{f}_N(x) - f(x)\right] dx \right) \le \text{const} < \infty \quad \text{a.s.} \tag{63}
$$

with

$$
\varepsilon_{UB}(N) \triangleq \max\left\{h, \sqrt{\log N/(Nh)}\right\}. \tag{64}
$$

3. In order to obtain a similar result for the term (62), note that Lemma 2 implies

$$
\zeta_N(x) \triangleq \varepsilon_{LB}^{-1}(N) \left[f(x) - \hat{f}_N(x)\right] \le C(\omega) < \infty \quad \text{a.s.}
$$

uniformly with respect to both x and N, with

$$
\varepsilon_{LB}(N) \triangleq \sqrt{\log N/(Nh^2)}. \tag{65}
$$

Hence, one may apply Fatou lemma, taking into account that $u\mathbf{1}\{u > 0\}$ is a continuous, monotone function:

$$
\begin{aligned}
&\limsup_{N\to\infty} \varepsilon_{LB}^{-1}(N) \int_0^1 \left[f(x) - \hat{f}_N(x)\right] \mathbf{1}\left\{\hat{f}_N(x) < f(x)\right\} dx & (66)\\
\le\ & \int_0^1 \limsup_{N\to\infty} \zeta_N(x)\,\mathbf{1}\{\zeta_N(x) > 0\}\, dx & (67)\\
\le\ & C(\omega) < \infty \quad \text{a.s.} & (68)
\end{aligned}
$$

4. Thus, the obtained relations together with (61) and (62) imply (14), (15) and Theorem 1 is proved. ■

The proof of Theorem 2 which is given in subsection 5.6 is based on the similar ideas as that of Theorem 1, see below.

5.4 Upper bound for $\tilde{f}_N$

Since the admissible set (19), (20) is narrower being compared to that of (4), (5), it is important to demonstrate that the upper bound remains at least the same.

Lemma 3 *Let the assumptions of Theorem 2 hold true. Then for almost all $\omega \in \Omega$ there exist finite number $N_0(\omega)$ such that*

$$
J_{MP}^* \le C_f + O(h) + O\left(\sqrt{\frac{\log N}{Nh}}\right), \qquad \forall\, N \ge N_0(\omega), \tag{69}
$$

with non random both $O(h)$ and $O\left(\sqrt{\log N/(Nh)}\right)$.

Proof of Lemma 3. 1. Since kernel function $K(t)$ is supposed to be even then matrix A is symmetric, and the related to (18)–(20) dual problem looks like

$$J^*_{MD} \triangleq \max_{\lambda,\,\nu} \left(Y^T\lambda - C_\alpha N^{-1}\mathbf{1}^T\nu\right) \tag{70}$$

subject to

$$A\lambda - \nu \le \mathbf{1} \tag{71}$$

$$\lambda \ge 0 \tag{72}$$

$$\nu \ge 0. \tag{73}$$

Let us replace vector Y in (70) for

$$F \triangleq (f(X_1), \ldots, f(X_N))^T, \tag{74}$$

and, moreover, change the vector constraint (71) by a scalar one which is directly obtained by just summing all N rows of (71). Thus we arrive at the modified dual problem

$$J^*_{MMD} \triangleq \max_{\lambda,\,\nu} \left(F^T\lambda - C_\alpha N^{-1}\mathbf{1}^T\nu\right) \tag{75}$$

subject to

$$\mathbf{1}^T A\lambda - \mathbf{1}^T\nu \le N \tag{76}$$

$$\lambda \ge 0 \tag{77}$$

$$\nu \ge 0. \tag{78}$$

Since $F \ge Y$ and according to the well known Duality Theorem

$$J^*_{MP} = J^*_{MD} \le J^*_{MMD}. \tag{79}$$

Now, we derive an upper bound on J^*_{MMD}.

2. Let us arbitrarily fix (λ, ν) which meet the constraints (76)–(78) and then write inequality (76) in the equivalent form as follows:

$$\frac{1}{N}\sum_{j=1}^{N} \lambda_j \left(K_h(0) + \sum_{i \ne j}^{N} K_h(X_i - X_j)\right) \le 1 + \frac{1}{N}\mathbf{1}^T\nu, \tag{80}$$

or, equivalently,

$$\frac{1}{N}\sum_{j=1}^{N} \lambda_j \left(\frac{1}{h}K(0) + \sum_{i \ne j}^{N} E\left\{K_h(X_i - X_j) \mid X_j\right\} + \sum_{i \ne j}^{N} \xi_{ij}\right) \le 1 + \frac{1}{N}\mathbf{1}^T\nu, \tag{81}$$

with

$$\xi_{ij} \triangleq K_h(X_i - X_j) - E\left\{K_h(X_i - X_j) \mid X_j\right\}. \tag{82}$$

Now apply upper bound (96), proved in Lemma 5, to the relation (81) taking into account that $K(0) > 0$ as well as (36)–(37). Hence,

$$\frac{N-1}{C_f N}\sum_{j=1}^{N} \lambda_j \left(f(X_j) + O(h) - C\sqrt{\frac{\log N}{Nh}}\right) \le 1 + \frac{1}{N}\mathbf{1}^T\nu, \qquad \forall\, N \ge N_2(\omega), \tag{83}$$

with non random constant C. First, from inequality (83) it follows that

$$\sum_{j=1}^{N} \lambda_j \leq \frac{C_f}{f_{\min}} \left(2 + \frac{1}{N}\mathbf{1}^T \nu\right), \qquad \forall\, N \geq N_3(\omega), \tag{84}$$

with almost surely finite $N_3(\omega) \geq N_2(\omega)$. Consequently, as it follows from (83), for almost all $\omega \in \Omega$ and sufficiently large N

$$F^T\lambda - \frac{C_\alpha}{N}\mathbf{1}^T\nu \;\leq\; C_f\left(1 + O(h) + O\left(\sqrt{\frac{\log N}{Nh}}\right)\right) \tag{85}$$

$$-(C_\alpha - C_f(1+o(1)))\mathbf{1}^T\nu, \tag{86}$$

with non random $O\left(\sqrt{\log N/(Nh)}\right)$. Thus, (79) and (85) prove the upper bound (69), since (20) implies $C_\alpha > C_f$. ■

5.5 Lower bound for $\tilde{f}_N$

Lemma 4 *Under the assumptions of Theorem 2, for almost all $\omega \in \Omega$ there exist finite number $N_1(\omega)$ such that for each $x \in (0,1)$*

$$\tilde{f}_N(x) \geq f(x) - O\left(\sqrt{\log N/(Nh)}\right), \qquad \forall\, N \geq N_1(\omega), \tag{87}$$

where $O(\cdot)$ do not depend on x.

Proof of Lemma 4 is given in the same manner as that of Lemma 2. The only essential difference is in better Lipschitz constant for $\tilde{f}_N(x)$. Indeed, for any $u, v \in (0,1)$

$$\left|\tilde{f}_N(u) - \tilde{f}_N(v)\right| \;\leq\; \sum_{i=1}^{N} \alpha_i \left|K_h(u - X_i) - K_h(v - X_i)\right| \tag{88}$$

$$\leq\; \frac{L_K}{h^2}\left(\sum_{i\in I(u)} \alpha_i + \sum_{i\in I(v)} \alpha_i\right)|u - v|, \tag{89}$$

with

$$I(\cdot) \triangleq \{i \mid K_h(\cdot - X_i) \neq 0\}. \tag{90}$$

From the Strong Law of Large Numbers,

$$\mathrm{Card}\, I(\cdot) = \frac{f(\cdot)}{C_f} Nh(1 + o(1)) \quad \text{a.s.} \tag{91}$$

and thus,

$$L_{\tilde{f}_N} = \frac{L_K}{h^2}\frac{C_\alpha}{N}\frac{2f_{\max}}{C_f} Nh = O\left(\frac{1}{h}\right) \tag{92}$$

by the upper bound (20) on α. ■

5.6 Proof of Theorem 2

Theorem 2 is proved in the same manner as that of Theorem 1, basing on lemmas 1 and 2. Note, that the lower bound from Lemma 2 is now not worse being compared to the upper bound, which is the result of the estimator modification.

Note: The result (22)–(23) of Theorem 2 may also be proved for differentiable kernel functions with infinite support which meet the condition

$$|K'(t)| \leq \mu K(t), \quad \forall\, t \in \mathbb{R}, \tag{93}$$

with some constant μ. Indeed, (93) implies

$$\left|\tilde{f}_N'(x)\right| \leq \frac{1}{h^2}\sum_{i=1}^{N}\alpha_i \left|K'\left(\frac{x-X_i}{h}\right)\right| \leq \frac{\mu}{h}\tilde{f}_N(x). \tag{94}$$

Consequently, when estimate function $\tilde{f}_N(x)$ is bounded from above, its Lipschitz constant is of order $O\left(h^{-1}\right)$ that is the same as in (92).

6 Appendix

Lemma 5 *Let the assumptions A and B hold true and constant C be sufficiently large. Define the random variables*

$$\xi_{ij} \triangleq K_h(X_i - X_j) - E\left\{K_h(X_i - X_j) \mid X_j\right\}, \quad i \neq j. \tag{95}$$

Then, for almost all $\omega \in \Omega$ there exist finite integer $N_2(\omega)$ such that

$$\max_{j=1,\ldots,N}\left|\frac{1}{N-1}\sum_{i\neq j}^{N}\xi_{ij}\right| \leq C\sqrt{\log N/(Nh)} \qquad \forall\, N \geq N_2(\omega). \tag{96}$$

Proof of Lemma 5. Note that for each $j = 1,\ldots,N$ the unbiased i.i.d. random variables $(\xi_{ij})\,|_{i\neq j}$ have the following properties:

$$|\xi_{ij}| \leq \frac{2}{h}K_{\max} \triangleq a, \tag{97}$$

and

$$\begin{aligned} E\left\{\xi_{ij}^2 \mid X_j\right\} &\leq \frac{1}{h^2 C_f}\int_0^1 K^2\left(\frac{u-X_j}{h}\right) f(u)\,du \\ &\leq \frac{1}{hC_f}\int_{\mathbb{R}} K^2(t)\, f(X_j + ht)\,dt \\ &\leq \frac{C_0(K)}{hC_f} f_{\max} \triangleq \sigma_1^2. \end{aligned} \tag{98}$$

Thus, one may apply the Bernstein inequality (see, e.g., Birgé & Massart [4] or Bosq [6], Theorem 2.6) which leads to

$$P\left\{\left|\frac{1}{N-1}\sum_{i\neq j}^{N}\xi_{ij}\right| > \mu \quad \Big|\, X_j\right\} \;\leq\; 2\exp\left(-\frac{(N-1)\mu^2}{2(\sigma_1^2 + a\mu/3)}\right).$$

Let us put

$$\mu = \sqrt{\frac{\kappa \log N}{Nh}}, \tag{99}$$

with sufficiently large κ which is defined below. Hence, for all $N \geq N_1$, N_1 being sufficiently large non random integer,

$$P\left\{ \left| \frac{1}{N-1} \sum_{i \neq j}^{N} \xi_{ij} \right| > \sqrt{\frac{\kappa \log N}{Nh}} \quad \Big| \, X_j \right\} \;\leq\; 2N^{-\kappa_1},$$

with

$$\kappa_1 \triangleq \frac{\kappa\, C_f f_{\max}}{3C_0(K)}. \tag{100}$$

Therefore,

$$P\left\{ \max_{j=1,N} \left| \frac{1}{N-1} \sum_{i \neq j}^{N} \xi_{ij} \right| > \sqrt{\frac{\kappa \log N}{Nh}} \quad \Big| \, X_j \right\} \tag{101}$$

$$\leq \; \sum_{j=1}^{N} P\left\{ \left| \frac{1}{N-1} \sum_{i \neq j}^{N} \xi_{ij} \right| > \sqrt{\frac{\kappa \log N}{Nh}} \quad \Big| \, X_j \right\} \tag{102}$$

$$\leq \; 2N^{1-\kappa_1}. \tag{103}$$

Consequently, any fixed parameter

$$\kappa > \frac{6C_0(K)}{C_f f_{\max}} \tag{104}$$

ensures $\kappa_1 > 2$ which implies the convergence of series $\sum^{\infty} N^{1-\kappa_1}$ and, due to Borel–Cantelli lemma, the desired result (96). ∎

References

[1] Abbar, H. (1990) Un estimateur spline du contour d'une répartition ponctuelle aléatoire. *Statistique et analyse des données*, **15**(3), 1–19.

[2] Barron, A.R., Birgé, L. and Massart, P. (1999) Risk Bounds for model selection via penalization. *Probab. Theory Relat. Fields*, **113**, 301–413.

[3] Baufays, P. and Rasson, J.P. (1985) A new geometric discriminant rule. *Computational Statistics Quaterly*, **2**, 15–30.

[4] Birgé, L. and Massart, P. (1995) Minimum contrast estimators on sieves. *Preprint Université Paris Sud, France*, **95-42**.

[5] Bonnans, F., Gilbert, J.C., Lemaréchal, C. and Sagastizábal, C. (1997) Optimisation numérique. Aspects théoriques et pratiques. in *Mathématiques & Applications*, **27**, Springer, Paris.

[6] Bosq, D. (2000) Linear processes in function spaces. Theory and applications. in *Lecture Notes in Statistics*, **149**, Springer-Verlag, New York.

[7] Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the inefficiency of decision making units. *European Journal of Operational Research*, **2**, 429–444.

[8] Cowling, A. and Hall, P. (1996) On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society B*, **58**, 551–563.

[9] Cristianini, N. and Shawe-Taylor, J. (2000) *An introduction to support vector machines*, Cambridge University Press.

[10] Deprins, D., Simar, L. and Tulkens, H. (1984) Measuring Labor Efficiency in Post Offices. in *The Performance of Public Enterprises: Concepts and Measurements* by M. Marchand, P. Pestieau and H. Tulkens, North Holland ed, Amsterdam.

[11] Devroye, L.P. and Wise, G.L. (1980) Detection of abnormal behavior via non parametric estimation of the support. *SIAM J. Applied Math.*, **38**, 448–480.

[12] Gardes, L. (2002) Estimating the support of a Poisson process via the Faber-Shauder basis and extreme values. *Publications de l'Institut de Statistique de l'Université de Paris*, **XXXXVI**, 43–72.

[13] Geffroy, J. (1964) Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, **XIII**, 191–200.

[14] Gijbels, I. and Peng, L. (1999). Estimation of a support curve via order statistics. *Discussion Paper* **9905**, Institut de Statistique, Université Catholique de Louvain.

[15] Gijbels, I., Mammen, E., Park, B.U. and Simar, L. (1999). On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association*, **94**, 220–228.

[16] Girard, S. and Jacob, P. (2002a) Extreme values and Haar series estimates of point processes boundaries. *Scandinavian Journal of Statistics*, to appear.

[17] Girard, S. and Jacob, P. (2002b) Projection estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, to appear.

[18] Girard, S. and Jacob, P. (2001) Extreme values and kernel estimates of point processes boundaries. *Technical report ENSAM-INRA-UM2*, **01-02**.

[19] Girard, S. and Menneteau, L. (2002) Limit theorems for extreme values estimates of point processes boundaries. *Technical report INRIA*, **RR-4366**.

[20] Hall, P., Nussbaum, M. and Stern, S.E. (1997) On the estimation of a support curve of indeterminate sharpness. *J. Multivariate Anal.*, **62**, 204–232.

[21] Härdle, W., Hall, P. and Simar, L. (1995) Iterated boostrap with application to frontier models. *J. Productivity Anal.*, **6**, 63–76.

[22] Härdle, W., Park, B. U. and Tsybakov, A. B. (1995) Estimation of a non sharp support boundaries. *J. Multiv. Analysis*, **43**, 205–218.

[23] Härdle, W. (1990) *Applied nonparametric regression*, Cambridge University Press, Cambridge.

[24] Hardy, A. and Rasson, J.P. (1982) Une nouvelle approche des problèmes de classification automatique. *Statistique et Analyse des données*, **7**, 41–56.

[25] Hartigan, J.A. (1975) *Clustering Algorithms,* Wiley, Chichester.

[26] Hiriart-Urruty, J.B., Lemaréchal, C. (1993) Convex analysis and minimization algorithms. Part 1: Fundamentals. in *Grundlehren der Mathematischen Wissenschaften*, **305**, Springer-Verlag, Berlin.

[27] Jacob, P. and Abbar, H. (1989) Estimating the edge of Cox process area. *Cahiers du Centre d'Etudes de Recherche Opérationnelle*, **31**, 215–226.

[28] Jacob, P. and Suquet, P. (1995) Estimating the edge of a Poisson process by orthogonal series. *Journal of Statistical Planning and Inference*, **46**, 215–234.

[29] Korostelev, A., Simar, L. and Tsybakov, A. B. (1995) Efficient estimation of monotone boundaries. *The Annals of Statistics*, **23**, 476–489.

[30] Korostelev, A.P. and Tsybakov, A.B. (1993) Minimax theory of image reconstruction. in *Lecture Notes in Statistics*, **82**, Springer-Verlag, New York.

[31] Schölkopf, B. and Smola, A. (2002) *Learning with kernels,* MIT University Press, Cambridge.

[32] Tarssenko, L., Hayton, P., Cerneaz, N. and Brady, M. (1995) Novelty detection for the identification of masses in mammograms. In *Proceedings fourth IEE International Conference on Artificial Neural Networks*, 442–447, Cambridge.