Automation and Remote Control, Vol. 65, No. 1, 2004, pp. 58–64. Translated from Avtomatika i Telemekhanika, No. 1, 2004, pp. 66–73.

Nonparametric Frontier Estimation by Linear Programming[1]

G. Bouchard*, S. Girard, A. B. Iouditski**, and A. V. Nazin*****

**INRIA Rhône-Alpes, Grenoble, France*
***Grenoble I University, Grenoble, France*
****Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
Received July 10, 2003

Abstract—A new method for estimating the frontier of a set of points (or a support, in other words) is proposed. The estimates are defined as kernel functions covering all the points and whose associated support is of smallest surface. They are written as linear combinations of kernel functions applied to the points of the sample. The weights 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-norm for the error of estimation is shown to be almost surely converging to zero, and the rate of convergence is provided.

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 frontier or support estimation arises in classification [1], clustering problems [2], discriminant analysis [3], and outliers detection. Applications are found in medical diagnosis [4] as well as in condition monitoring of machines [5]. 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 [6]. We also point out some applications in econometrics [7].

In such cases, the unknown support can be written

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

where $f : [0, 1] \to (0, +\infty)$ is an unknown function. Here, the problem reduces to estimating f, called the production frontier, see for instance [8]. 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. In [9], functions $f(x)$ were supposed 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 piece-wise linear and, up to our knowledge, it is the first frontier estimate computed thanks to a linear programming technique [10]. Its asymptotic distribution is established in [11].

[1] Recommended for publication by Program Committee of the II International Control Conference (Moscow, Institute of Control Sciences, June 17–19, 2003). This work was carried out in framework of IS2 project, INRIA Rhône-Alpes. Financial support for Mr Bouchard from the IAP research network nr P5/24 of the Belgian Government (Federal Office for Scientific, Technical and Cultural Affairs) is gratefully acknowledged.

An early paper was written by Geffroy [12] 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 many directions.

On the one hand, piece-wise 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 in [6, 13] under weak assumptions on the rate of decrease of the density ϕ towards 0. Extreme values methods are then proposed in [14, 15] 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 and Jacob [16] introduced estimates based on kernel regressions and orthogonal series method [17, 18]. In the same spirit, Gardes proposed a Faber-Shauder estimate in [19]. Girard and Menneteau [20] introduced a general framework for studying estimates of this type and generalized them to supports writting

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

where $f : E \to (0, +\infty)$ 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 [21] and Jacob and Suquet [22] 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., [23], chapter 4), its smoothness is directly linked to the smoothness of the chosen kernel and it benefits from interesting theoretical properties like almost surely convergence in L_1-norm with some rate; proofs can be found in [24]. The estimate is defined in Section 2. Its theoretical properties are established in Section 3. The behaviour of the estimate is illustrated in [24] on finite sample situations. Its compared to a similar proposition found in [25].

2. BOUNDARY ESTIMATES

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 $Z_N = (X_i, Y_i)_{i=1,\ldots,N}$. The former represents an independent identically distributed (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 R by introducing $f(x) = 0$ for all $x \notin [0,1]$. Letting

$$C_f \triangleq \int_0^1 f(u)\, du = \int_R f(u)\, du\,,$$

each variable X_i is distributed in [0, 1] with probability density function (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 following family of functions:

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

where $K : R \to [0, +\infty)$ is a kernel function integrating to one, i.e.

$$\int_R K(u)\, du = 1\,,$$

and with bandwidth parameter h. 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 [26] for a review on this topic and to [27], chapter 8, for examples of application of SVM to quantile estimation. The constraint $\alpha_i \geq 0$ for all $i = 1, \ldots, N$ ensures that $\widehat{f}_N(x) \geq 0$ for all $x \in R$ and prevents the estimator from being too irregular. Let us remark that the surface of the estimated support is given by

$$\int_R \widehat{f}_N(x)\, dx = \sum_{i=1}^{N} \alpha_i\,. \tag{3}$$

This suggests to define the vector parameter $\alpha = (\alpha_1, \ldots, \alpha_N)^T$ from the Linear Program as follows

$$J_P^* \triangleq \min_{\alpha \in R^N} \mathbf{1}^T \alpha \tag{4}$$

subject to

$$A\alpha \geq Y, \tag{5}$$

$$\alpha \geq 0\,. \tag{6}$$

Here

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

Hence, $A\alpha = \left(\widehat{f}_N(X_1), \ldots, \widehat{f}_N(X_N)\right)^T$ and the vector constraint (5) means that

$$\widehat{f}_N(X_i) \geq Y_i\,, \qquad i = 1, \ldots, N\,. \tag{7}$$

In other words, $\widehat{f}_N$ defines the kernel estimate of the support covering all the points and with smallest surface. In practice the solution of the linear program is sparse in the sense that the number of nonzero coefficients $n(\alpha) \triangleq \#\{\alpha_i \neq 0\}$ 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)–(6) 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 can be written

$$p(Z_N \mid f) = \prod_{i=1}^{N} \frac{f(X_i)}{C_f} \frac{1}{f(X_i)} \mathbf{1}\{0 \leq Y_i \leq f(X_i)\} = \prod_{i=1}^{N} \frac{1}{C_f} \mathbf{1}\{0 \leq Y_i \leq f(X_i)\}, \tag{8}$$

where $I\{\cdot\}$ is the indicator function. Moreover,

$$C_f\Big|_{f=\widehat{f}_N} = \sum_{i=1}^{N} \alpha_i\,, \tag{9}$$

and therefore, the Log-Likelihood function is

$$L(\alpha) \triangleq \log p(Z_N \mid \widehat{f}_N) = -N \log \sum_{i=1}^{N} \alpha_i + \sum_{i=1}^{N} \log \mathbf{1}\{Y_i \le \widehat{f}_N(X_i)\}\,, \tag{10}$$

and its maximization over the set of non-negative parameters α is equivalent to problem (4)–(6).

Let us remark that other solutions for estimating α have already been proposed. Girard and Menneteau [20] considered a partition of $[0,1]$ by the intervals $\{I_r : 1 \le r \le k\}$, with $k \to \infty$. For each $r = 1, \dots, k$, they introduce $D_r = \{(x,y) : x \in I_r,\ 0 \le y \le f(x)\}$, the slice of S built on I_r, $Y_r^* = \max\{Y_i : (X_i, Y_i) \in D_r\}$ and define the estimates

$$\widehat{\alpha}_i = \begin{cases} \lambda(I_r) Y_r^* (1 + \frac{1}{N_r}), & \text{if } \exists\, r \in \{1, \dots, k\} \,:\, Y_i = Y_r^* \\ 0 & \text{otherwise;} \end{cases}$$

here λ is the Lebesgue measure and N_r is the number of points in the cell D_r (the use of N_r allows to reduce the bias of the estimate). They propose the following frontier estimate

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

where x_r denotes the center of I_r. This approach suffers from a practical difficulty: the choice of the partition of $[0,1]$ and more precisely the choice of k. In our context, solving the linear problem (4)–(6) direcly yields the support vectors.

In this sense, the estimate proposed in [25] is similar to $\widehat{f}_N$. It is defined by the Fourier expansion

$$\widehat{g}_N(x) = c_0 + \sum_{k=1}^{M} a_k \cos(2\pi k x) + \sum_{k=1}^{M} b_k \sin(2\pi k x), \tag{11}$$

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

$$\min\ c_0 \ \left(= \int_0^1 \widehat{g}_N(x)dx\right) \tag{12}$$

under the constraints

$$\widehat{g}_N(X_i) \ \ge \ Y_i, \quad i = 1, \dots, N\,, \tag{13}$$

$$\sum_{k=1}^{M} k\ (|a_k| + |b_k|) \ \le \ \frac{L}{2\pi}\,. \tag{14}$$

Therefore, $\widehat{g}_N$ defines the Fourier estimate of the support covering all the points (13), L-Lipschitzian (14) and with smallest surface (12). From the theoretical point of view, this estimate benefits from minimax optimality.

3. MAIN RESULTS

In this section, we establish that $\widehat{f}_N$ is almost surely (a.s.) convergent for the L_1-norm on $[0,1]$. To this end, the basic assumptions on the unknown boundary function are introduced:

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

(A2) $|f(x) - f(y)| \le L_f |x - y|$ for all $x, y \in [0,1]$; the Lipschitz constant $L_f < \infty$.

The following assumptions on the kernel function are considered:

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

(B2) $\int_R K(t)\,dt = 1$,

(B3) $|K(s) - K(t)| \leq L_K|s - t|$ with $L_K < \infty$,

(B4) $\int_R K^2(t)\,dt < \infty$ and $\int_R t^2\,K(t)\,dt < \infty$.

Below, let symbol "$a_N \asymp b_N$" denotes the "asymptotic equivalence" which means for two sequences of positive real numbers $\{a_N\}$ and $\{b_N\}$ that

$$0 < \liminf_{N\to\infty} a_N/b_N \leq \limsup_{N\to\infty} a_N/b_N < +\infty\,.$$

Theorem 1. *Let $h \to 0$ and $\log N/(Nh^2) \to 0$ as $N \to \infty$. Let the above mentioned assumptions A1, A2 and B1–B4 hold true. Then, estimator* (2)–(6) *has the following asymptotic properties:*

$$\limsup_{N\to\infty} \varepsilon_1^{-1}(N)\,\|\hat{f}_N - f\|_1 \leq C(\omega) < \infty \quad a.s.\,, \tag{15}$$

$$\varepsilon_1(N) \triangleq \max\left\{h, \sqrt{\log N/(Nh^2)}\right\}. \tag{16}$$

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

$$h \asymp (\log N/N)^{1/4}\,. \tag{17}$$

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 $\widetilde{f}_N$ will usually rely on more support vectors than $\widehat{f}_N$.

Let us modify the estimator (4)–(6) as follows.

$$\widetilde{f}_N(x) = \sum_{i=1}^{N} K_h(x - X_i)\,\alpha_i, \tag{18}$$

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

$$J^*_{MP} \triangleq \min_{\alpha\in R^N} \mathbf{1}^T\alpha \tag{19}$$

subject to

$$A\alpha \geq Y, \tag{20}$$

$$0 \leq \alpha \leq C_\alpha/N\,, \tag{21}$$

with a constant

$$C_\alpha > f_{\max}. \tag{22}$$

Remark. In fact, we need to ensure $C_\alpha > C_f$ which is implied by (22).

The modified estimator (18)–(22) differs from that of (2)–(6) by additionally bounding each coefficient α_i from above, see constraints (21).

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 A1, A2 and B1–B4 hold true. Then, estimator* (18)–(22) *has the following asymptotic properties:*

$$\limsup_{N\to\infty} \varepsilon_2^{-1}(N)\, \|\widetilde{f}_N - f\|_1 \leq C(\omega) < \infty \quad a.s.\,, \tag{23}$$

$$\varepsilon_2(N) \triangleq \max\left\{h, \sqrt{\log N/(Nh)}\right\}. \tag{24}$$

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

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

is attained for

$$h \asymp (\log N/N)^{1/3}\,. \tag{25}$$

4. CONCLUSION

The above theoretical results demonstrate that using the kernel estimators (2)–(6) and (18)–(22) with a "correct" bandwidth h one can ensure convergence rate in L_1-norm close in order to $O\left(N^{-1/4}\right)$ and $O\left(N^{-1/3}\right)$ respectively. Calculation of the estimates is reduced to Linear Programming, and the resulting solutions are sparse that is a small portion of support points is taken into account. Thus, some loss of convergence rate being comparing to the optimal rate $O\left(N^{-1/2}\right)$ (for a considered class of Lipschitz frontier functions, see [25]) is explained by a considerable decreasing of computational complexity of resulting estimators.

REFERENCES

1. Hardy, A. and Rasson, J.P., Une nouvelle approche des problèmes de classification automatique, *Statist. Anal. Données*, 1982, vol. 7, pp. 41–56.
2. Hartigan, J.A., *Clustering Algorithms*, Chichester: Wiley, 1975.
3. Baufays, P. and Rasson, J.P., A New Geometric Discriminant Rule, *Comput. Statist. Quaterly*, 1985, vol. 2, pp. 15–30.
4. Tarassenko, L., Hayton, P., Cerneaz, N., and Brady, M., Novelty Detection for the Identification of Masses in Mammograms, *Proc. Fourth IEE Int. Conf. Artificial Neural Networks*, Cambridge, 1995, pp. 442–447.
5. Devroye, L.P. and Wise, G.L., Detection of Abnormal Behavior via non Parametric Estimation of the Support, *SIAM J. Appl. Math.*, 1980, vol. 38, pp. 448–480.
6. Korostelev, A.P. and Tsybakov, A.B., *Minimax Theory of Image Reconstruction*, New York: Springer-Verlag, 1993, vol. 82.
7. Deprins, D., Simar, L., and Tulkens, H., Measuring Labor Efficiency in Post Offices, in *The Performance of Public Enterprises: Concepts and Measurements*, Marchand, M., Pestieau, P., and Tulkens, H., Eds., Amsterdam: North Holland,1984.
8. Härdle, W., Hall, P., and Simar, L., Iterated Bootstrap with Application to Frontier Models, *J. Productiv. Anal.*, 1995, vol. 6, pp. 63–76.

9. Korostelev, A., Simar, L., and Tsybakov, A.B., Efficient Estimation of Monotone Boundaries, *Ann. Statist.*, 1995, vol. 23, pp. 476–489.

10. Charnes, A., Cooper, W.W., and Rhodes, E., Measuring the Inefficiency of Decision Making Units, *Eur. J. Oper. Res.*, 1978, vol. 2, pp. 429–444.

11. Gijbels, I., Mammen, E., Park, B.U., and Simar, L., On Estimation of Monotone and Concave Frontier Functions, *J. Am. Statist. Assoc.*, 1999, vol. 94, pp. 220–228.

12. Geffroy, J., Sur un problème d'estimation géométrique, *Publications de l'Institut de Statist., Université de Paris*, 1964, vol. XIII, pp. 191–200.

13. Härdle, W., Park, B.U., and Tsybakov, A.B., Estimation of a non Sharp Support Boundaries, *J. Multivariate Anal.*, 1995, vol. 43, pp. 205–218.

14. Hall, P., Nussbaum, M., and Stern, S.E., On the Estimation of a Support Curve of Indeterminate Sharpness, *J. Multivariate Anal.*, 1997, vol. 62, pp. 204–232.

15. Gijbels, I. and Peng, L., Estimation of a Support Curve via Order Statistics, *Discussion Paper 9905*, Institut de Statist., Université Catholique de Louvain, 1999.

16. Girard, S. and Jacob, P., Extreme Values and Kernel Estimates of Point Processes Boundaries, *Technical Report ENSAM-INRA-UM2, 01-02*, 2001 (http://www.inrialpes.fr/is2/people/girard/working.html).

17. Girard, S. and Jacob, P., Extreme Values and Haar Series Estimates of Point Processes Boundaries, *Scandinav. J. Statist.*, 2002, vol. 30, no. 2, pp. 369–384.

18. Girard, S. and Jacob, P., Projection Estimates of Point Processes Boundaries, *J. Statist. Planning Inference*, 2003, vol. 116, no. 1, pp. 1–15.

19. Gardes, L., Estimating the Support of a Poisson Process via the Faber–Shauder Basis and Extreme Values, *Publications de l'Institut de Statist., Université de Paris*, 2002, vol. XXXXVI, pp. 43–72.

20. Girard, S. and Menneteau, L., Cental Limit Theorems for Smoothed Extreme Values Estimates of Point Processes Boundaries, *J. Statist. Planning Inference*, 2003.

21. Abbar, H., Un estimateur spline du contour d'une répartition ponctuelle aléatoire, *Statist. Anal. Données*, 1990, vol. 15., no. 3, pp. 1–19.

22. Jacob, P. and Suquet, P., Estimating the Edge of a Poisson Process by Orthogonal Series, *J. Statist. Planning Inference*, 1995, vol. 46, pp. 215–234.

23. Bonnans, F., Gilbert, J.C., Lemaréchal, C., and Sagastizábal, C., Optimisation numérique. Aspects théoriques et pratiques, *Math. Appl.*, Paris: Springer, 1997, vol. 27.

24. Bouchard, G., Girard, S., Iouditski, A., and Nazin, A., Linear Programming Problems for Frontier Estimation, *Technical Report INRIA, RR-4717*, 2003 (http://www.inria.fr/rrrt/rr-4717.html).

25. Barron, A.R., Birgé, L., and Massart, P., Risk Bounds for Model Selection via Penalization, *Prob. Theory Relat. Fields*, 1999, vol. 113, pp. 301–413.

26. Cristianini, N. and Shawe-Taylor, J., *An Introduction to Support Vector Machines*, Cambridge: Cambridge Univ. Press, 2000.

27. Schölkopf, B. and Smola, A., *Learning with Kernels*, Cambridge: MIT Press, 2002.

This paper was recommended for publication by V.A. Lototskii, a member of the Editorial Board