Statistica Sinica **16**(2006), 1233-1253

ROBUST NONPARAMETRIC FRONTIER ESTIMATORS: QUALITATIVE ROBUSTNESS AND INFLUENCE FUNCTION

Abdelaati Daouia and Anne Ruiz-Gazen

Université Paul Sabatier and Université des Sciences Sociales, Toulouse

Abstract: In the context of productivity analysis, the most popular nonparametric estimators of a monotone boundary are defined as lowest monotone functions covering all sample points and are very non-robust. Two alternatives have been addressed recently: one by Cazals, Florens and Simar (2002) is based on a concept of expected order-m frontiers; the other by Aragon, Daouia and Thomas-Agnan (2005) is based on extreme quantiles of a nonstandard conditional probability density. Unlike usual methods, both alternatives are shown to be qualitatively robust and bias-robust. Moreover, for the quantile approach, the influence function remains bounded even when the quantile order tends to one under the conditions that the conditional density function is not null, and is continuous on its support. When these conditions do not hold, the robust behavior of the quantile approach is shown on two numerical examples. A data set is provided to show the advantage of the robust proposals and the use of gross-error sensitivity as a diagnostic tool to detect anomalous data.

Key words and phrases: Estimation of a monotone boundary, gross-error sensitivity, influence function, maximum bias, outliers detection, productivity analysis, Prohorov distance, qualitative robustness.

1. Introduction

Let Ψ be the support of the joint distribution of a random vector $(X, Y) \in \mathbb{R}^2_+$. Let $\mathcal{X} = \{x \in \mathbb{R}_+ |\ Y(x) \neq \emptyset\}$ where $Y(x) = \{y \in \mathbb{R}_+ |\ (x, y) \in \Psi\}$. The graph of the function $\varphi(x) = \sup Y(x)$ for any $x \in \mathcal{X}$ describes the upper topological boundary of Ψ. Consider the estimation of this upper boundary from a random sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent vectors with the same distribution as (X, Y). We assume that this boundary is monotone nondecreasing in the sense that the frontier function φ is nondecreasing in x.

Denote by F, respectively F_X, the joint distribution function (df) of (X, Y) and the marginal df of X. The monotone boundary of Ψ can be characterized by the function φ which associates to a given level x the upper boundary of the support of Y conditioned by $X \leq x$, i.e., $\varphi(x) = \inf\{y \in \mathbb{R}_+ | F(y|x) = 1\}$, where $F(y|x) = F(x, y)/F_X(x)$. From now on we assume that $x \in \mathbb{R}_+$ is such that $F_X(x) > 0$. The graph of φ is the smallest monotone nondecreasing frontier which is larger than or equal to the boundary of Ψ.

The problem of estimating a monotone boundary appears naturally in the context of productivity analysis. When analyzing the productivity of firms, one may want to compare how the firms transform an input x (e.g., labor, energy or capital) into an output y (e.g., a quantity of produced goods). In this context, Ψ is the attainable production set and φ is the production frontier function, the geometric locus of the optimal production. For a firm operating with input x, $\varphi(x)$ is the maximal level of attainable output. The economic efficiency of a firm is then defined in terms of its ability to operate close to this optimal level $\varphi(x)$; if its production is y, then its efficiency score may be measured via its total value of output relative to the production frontier, that is $y/\varphi(x)$. With this measure, one can detect the most efficient or inefficient firms depending on whether $y/\varphi(x)$ is close to 1 or 0.

It is often reasonable to assume that the monotone frontier φ is a concave function. Then a famous estimator comes from data envelopment analysis (DEA). Farrell (1957) introduced the DEA estimator $\widehat{\Psi}_{DEA}$ of Ψ, the set under the "lowest" concave monotone increasing function covering all the sample points (X_i, Y_i). The DEA estimator of $\varphi(x)$ is then defined by the maximum of y such that (x, y) belongs to $\widehat{\Psi}_{DEA}$. Its consistency was addressed by Banker (1993), and its asymptotic distribution was derived by Gijbels, Mammen, Park and Simar (1999).

The nonparametric estimator of φ with possible non-concavity was first studied by Deprins, Simar and Tulkens (1984) in the context of measuring the efficiency of enterprises. They introduced the free disposal hull (FDH) estimator $\widehat{\varphi}_n(x) = \max_{i|X_i \leq x} Y_i$, to represent the lowest monotone step function covering all the data points (X_i, Y_i). The asymptotic theory of this FDH is now mostly available. See Korostelev, Simar and Tsybakov (1995) for the efficiency in the asymptotic minimax sense, and Park, Simar and Weiner (2000) for the asymptotic distribution.

Besides productivity analysis, DEA and FDH methods have been used in many related fields of application, including analysis of the performance of investment portfolios (Capital Assets Pricing Models, where X measures the volatility or the variance of a portfolio and Y represents its averaged return, see, e.g., Markovitz (1959)), analysis of the performance of public services, judicial activities, urban transit, pharmacies, hospitals, banks, and other institutions (see Seiford (1996) for a survey, and more than 700 references). But in the presence of outliers, these nonparametric envelopment estimators can behave erratically since, by construction, they envelope all the observations. Recently, robust nonparametric frontier estimators have been suggested by Cazals, Florens and Simar (2002) and Aragon, Daouia and Thomas-Agnan (2005).

In place of estimating the full frontier, Cazals, Florens and Simar (2002) propose to estimate a frontier of a discrete order $m \in \mathbb{N}^*$, which increases with

respect to m to achieve the efficient frontier φ when $m \to \infty$. For a given level x, this order-m frontier is defined as the expected value of the maximum of m independent random variables $Y^1, \ldots, Y^m$, drawn from the conditional distribution of Y given $X \leq x$. Formally,

$$\varphi_m(x) = E[\max(Y^1, \ldots, Y^m)] = \int_0^\infty (1 - [F(y|x)]^m) dy,$$

where the integrand is identically zero for $y \geq \varphi(x)$. In the context of productivity analysis, $\varphi_m(x)$ represents the expected maximum achievable level of outputs among m firms drawn from the population of production units using less than x as inputs. A natural estimator is provided by a plug-in argument,

$$\widehat{\varphi}_{m,n}(x) = \widehat{E}[\max(Y^1, \ldots, Y^m)] = \int_0^\infty (1 - [\widehat{F}_n(y|x)]^m) dy, \tag{1.1}$$

where $\widehat{F}_n(y|x) = \widehat{F}_n(x,y)/\widehat{F}_{X,n}(x)$ is the empirical version of $F(y|x)$, with

$$\widehat{F}_n(x,y) = \frac{1}{n}\sum_{i=1}^n 1(X_i \leq x, Y_i \leq y), \quad \widehat{F}_{X,n}(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \leq x),$$

where $1(A)$ stands for the indicator function of the set A. The integrand is identically zero for $y \geq \widehat{\varphi}_n(x)$. This estimator achieves $\sqrt{n}$ consistency and is asymptotically normal. A stronger functional convergence theorem of $\widehat{\varphi}_{m,n}$ to φ_m is provided in the appendix B of Cazals, Florens and Simar (2002). Moreover by choosing m, appropriately as a function of the sample size n, the estimator $\widehat{\varphi}_{m,n}$ as an estimator of the frontier φ recovers the asymptotic properties of the FDH $\widehat{\varphi}_n$. Note also that, due to the expectation in (1.1), the estimator $\widehat{\varphi}_{m,n}$ does not envelope all the data points, even for large m, and so is more robust to extreme values than the FDH $\widehat{\varphi}_n$.

Aragon, Daouia and Thomas-Agnan (2005) introduced a new concept of an order-α frontier, one that increases with respect to the continuous order $\alpha \in [0,1]$, and converges to the efficient frontier $\varphi(x)$ as $\alpha \nearrow 1$. The concept is the conditional α-quantile of the distribution of Y given $X \leq x$, i.e.,

$$q_\alpha(x) := F^{-1}(\alpha|x) = \inf\{y \in \mathbb{R}_+ | F(y|x) \geq \alpha\}.$$

From an economic point of view, this nonstandard conditional quantile has its own interest: it gives the production threshold exceeded by $100(1-\alpha)\%$ of all production units using less than x as inputs. A nonparametric estimator of $q_\alpha(x)$ is easily derived by inverting the empirical version of the conditional df $F(\cdot|x)$, i.e.,

$$\widehat{q}_{\alpha,n}(x) := \widehat{F}_n^{-1}(\alpha|x) = \inf\{y \in \mathbb{R}_+ | \widehat{F}_n(y|x) \geq \alpha\}.$$

As pointed out by Aragon, Daouia and Thomas-Agnan (2005), this estimator is very fast to compute, very easy to interpret, and its order α can be useful in terms of practical efficiency analysis. It has at least the same statistical properties as the nonparametric estimator $\widehat{\varphi}_{m,n}(x)$. In particular, by choosing α as an appropriate function of n, one estimates the true frontier function $\varphi(x)$ and satisfies the asymptotic properties of the FDH estimator, $\widehat{\varphi}_n = \widehat{q}_{1,n}(x)$. Not all observations are enveloped and the estimator is less sensitive to extreme values than the FDH. Moreover, it is shown by numerical analysis with simulated and real data that these order-α frontiers are more resistant to extreme values and to outliers than the order-m frontiers.

In this paper we analyze and compare the reliability of the statistical procedures of Cazals, Florens and Simar (2002) and Aragon, Daouia and Thomas-Agnan (2005) from a robustness theory point of view. Two central concepts are investigated: the qualitative robustness and the influence function.

For statistics $T_n = T(G_n)$ representable as a functional T of an empirical distribution G_n, qualitative robustness at the underlying distribution G is defined as equicontinuity with respect to the Prohorov distance $\pi(\cdot,\cdot)$ (Prohorov (1956)) of the distributions of T_n as n changes (see Hampel (1971, p.1890) for the definition). The distributions G and G_n are probability measures on a measurable space $(\Lambda, \mathcal{A})$ such that Λ is a complete separable metric space and $\mathcal{A}$ denotes the σ-algebra generated by the topology. It is proved in Hampel (1971, Theorem 1a, p.1892) that, if $\{T_n\}$ is continuous at G (see Hampel (1971, p.1891) for the definition) and for every n, T_n is continuous as a point function on Λ^n, except for a set of G^n-measure 0 (where G^n denotes the product measure on Λ^n, determined by G on Λ), then $\{T_n\}$ is qualitatively robust. Qualitative robustness of both nonparametric order-m and order-α frontiers is established in Section 2, but this tells us little about the differences between the two nonparametric procedures.

The richest quantitative robustness information is provided by the influence function $u \mapsto IF(u;T,G)$ of T at G (Hampel (1974)). It is defined as the first Gâteaux derivative of T at G, where the point u plays the role of the coordinate in the infinite-dimensional space of probability distributions. The importance of the influence function lies in its two main uses. First, it describes the effect of an infinitesimal contamination at the point u on the estimate, standardized by the mass of the contamination. Second, it allows one to assess the relative influence of individual observations on the value of the estimate. If this is unbounded, an outlier can cause trouble. The maximum absolute value $\gamma^*(T,G) = \sup_u |IF(u;T,G)|$ defines the gross-error sensitivity of T at G; the supremum being taken over all u where $IF(u;T,G)$ exists. There are other robustness measures derived from the influence function, like the local-shift sensitivity, but the central local robustness measure is γ^*. Thus, besides qualitative

robustness, an important robustness requirement is B-robustness (Rousseeuw (1981)), which corresponds to a finite gross-error sensitivity.

Section 3 shows that γ^* is finite for both sequences of nonparametric estmators $\{\widehat{q}_{\alpha,n}(x)\}$ and $\{\widehat{\varphi}_{m,n}(x)\}$. The difference between the two nonparametric partial frontiers lies in the fact that the influence function is no longer bounded for order-m frontiers when m tends to infinity, while it remains bounded for the conditional quantile frontiers when the quantile order tends to one. This advantage occurs when the nonstandard conditional density function is strictly positive at the upper monotone frontier. In the case where the conditional density function is null at the upper frontier, the robustness of the quantile frontier estimators is illustrated in Section 4 through some simulated examples. A data set is also provided to show the advantage of the robust proposals and the use of the influence function as a diagnostic tool to detect anomalous data. In order to save space, detailed proofs of the theorems are not given in this paper, but can be found in Daouia and Ruiz-Gazen (2004).

2. Continuity and Qualitative Robustness

Let $0 < \alpha < 1$, $m \in \mathbb{N}^*$, $x \in \mathbb{R}_+$, and consider the statistical functionals $S^{m,x}$ and $T^{\alpha,x}$ which associate to a df G on $\mathbb{R}^2$ such that $G(x,\infty) > 0$, the real values

$$S^{m,x}(G) = \int_0^\infty \left(1 - \left[\frac{G(x,y)}{G(x,\infty)}\right]^m\right) dy \text{ and } T^{\alpha,x}(G) = \inf\left\{y \geq 0 \Big| \frac{G(x,y)}{G(x,\infty)} \geq \alpha\right\},$$

where the integrand is identically zero for $y \geq \varphi_G(x) := \inf\{y \mid G(x,y)/G(x,\infty) = 1\}$. For $S^{m,x}(G)$ to be well defined, it suffices for instance that $\varphi_G(x)$ be finite, whereas $T^{\alpha,x}(G)$ exists for any df G such that $G(x,\infty) > 0$ since $\alpha < 1$. It is then clear that

$$\begin{cases} \varphi_m(x) = S^{m,x}(F) \\ \widehat{\varphi}_{m,n}(x) = S^{m,x}(\widehat{F}_n) \end{cases}, \quad \begin{cases} q_\alpha(x) = T^{\alpha,x}(F) \\ \widehat{q}_{\alpha,n}(x) = T^{\alpha,x}(\widehat{F}_n). \end{cases}$$

The following conditions are needed to ensure qualitative robustness of the sequence of estimators $\{\widehat{q}_{\alpha,n}(x)\}$ at F.

H1. F_X is continuous at x with $F_X(x) > 0$;
H2. For any $y \in \mathbb{R}_+$, $u \mapsto F(y|u)$ is nonincreasing on $\{u \geq 0 \mid F_X(u) > 0\}$;
H3. The generalized inverse function $F^{-1}(\cdot|x)$ is continuous at α.

Note that, from an economic point of view, assumption H2 is quite reasonable since the chance of producing less than a value y decreases if a firm uses more

inputs. Note also that the assumption H3 holds if the usual assumption on standard regression quantiles is satisfied, i.e., if

H3′. $F(\cdot|x)$ is differentiable at $q_\alpha(x)$ with strictly positive derivative $f(q_\alpha(x)|x)$.

Theorem 2.0.1. *If conditions H1-H3 hold, then the sequence of estimators* $\{\widehat{q}_{\alpha,n}(x)\}$ *is continuous at* F *with respect to the Prohorov distance, and is qualitatively robust at* F.

In order to prove this theorem, we first show that $T^{\alpha,x}$ is continuous at F with respect to the Prohorov distance $\pi(\cdot,\cdot)$, and then that $T^{\alpha,x}$ is continuous as a function of the observations on $\mathbb{R}^{2n}$ except for a set of F^n-measure 0. Finally, we conclude by applying Hampel's Theorem (1971, Theorem 1a).

As a matter of fact, we obtain a stronger result than the required continuity of the sequence $\{T^{\alpha,x}(\widehat{F}_n)\}$ at F. We prove that the maximal bias

$$b_1(\varepsilon, F) = \sup_{\pi(F,G)<\varepsilon} |T^{\alpha,x}(G) - T^{\alpha,x}(F)|$$

converges to 0 as $\varepsilon \searrow 0$. Note also that, as an immediate consequence of continuity of $T^{\alpha,x}$ at F with respect to π, we obtain from Hampel (1971, Lemma 2) that the sequence of estimators $\{\widehat{q}_{\alpha,n}(x)\}$ is consistent for $q_\alpha(x)$.

The same technique of proof is used to establish the qualitative robustness of $\{\widehat{\varphi}_{m,n}(x)\}$. To prove the continuity of $\{S^{m,x}(\widehat{F}_n)\}$ at F with respect to π, we need to assume that

K1. $F_X(x) > 0$ and $\varphi(x) < \nu$, where ν is a finite positive constant;
K2. F_X is continuous on a neighborhood of x;
K3. F is continuous on $\{x\} \times \mathbb{R}$.

Condition K1 on the upper frontier φ of the support Ψ of (X, Y) is quite reasonable, for instance, in the economic theory underlying efficiency analysis, since Ψ is always bounded in practice.

As it can be seen in the proof of Theorem 2.0.2 below, if F satisfies both K1 and K3, then any df G on $\mathbb{R}^2$ such that $\pi(F, G) < \varepsilon$ satisfies K1 for all ε small enough with the same constant ν as the model distribution F (i.e., $G(x, \infty) > 0$ and $\varphi_G(x) < \nu$), and so

$$S^{m,x}(G) = \int_0^\nu \left(1 - \left[\frac{G(x,y)}{G(x,\infty)}\right]^m\right) dy \tag{2.1}$$

for all $\varepsilon > 0$ sufficiently small. We also show that, under K1, the functional $S^{m,x}(\widehat{F}_n) = S^{m,x}((X_1, Y_1), \ldots, (X_n, Y_n))$ can be defined as a point function on

Ψ^n by

$$S^{m,x}((x_1,y_1),\ldots,(x_n,y_n)) = \int_0^{\varphi(x)} \left(1 - \left[\frac{\sum_{i=1}^n 1(x_i \le x, y_i \le y)}{\sum_{i=1}^n 1(x_i \le x)}\right]^m\right) dy \quad (2.2)$$

if $\sum_{i=1}^n 1(x_i \le x) > 0$. We prove its continuity on Ψ^n instead of $\mathbb{R}^{2n}$, except for a set of F^n-measure 0, without making use of assumptions K2 and K3. It should be clear that the support Ψ endowed with the σ-algebra of its Borel sets defines a measurable space, which is also complete and separable since it is closed in $\mathbb{R}^2$. Thus, Hampel's Theorem can be applied to deduce the qualitative robustness of $\{S^{m,x}(\widehat{F}_n)\}$ at F.

Note that for the quantile framework, we show the continuity of $T^{\alpha,x}(\widehat{F}_n)$ as a point function on the complete and separable metric space $\mathbb{R}^{2n}$ without resorting to any assumptions, even those of Theorem 2.0.1.

Theorem 2.0.2. *If assumptions K1−K3 hold, then the sequence of estimators $\{\widehat{\varphi}_{m,n}(x)\}$ is continuous at F with respect to the Prohorov distance, and is qualitatively robust at F.*

Here also, we show that the maximal bias

$$b_2(\varepsilon, F) = \sup_{\pi(F,G)<\varepsilon} |S^{m,x}(G) - S^{m,x}(F)|$$

converges to 0 as $\varepsilon \searrow 0$, which implies the continuity of $S^{m,x}$ at F with respect to π.

3. Quantitative Robustness: Influence Functions

Let the orders $m \in \mathbb{N}^*$ and $\alpha \in]0,1[$ be fixed, and let $x \in \mathbb{R}_+$ be such that $F_X(x) > 0$. For the conditional quantile framework, it is convenient to consider only pairs (α, x) satisfying assumption H3′. Neither H1-H3 nor K2−K3 are needed in this section.

According to Hampel, Ronchetti, Rousseeuw and Stahel (1986, Definition 1, p.84), the corresponding influence functions of $S^{m,x}$ and $T^{\alpha,x}$ at F can be defined by

$$(x_0,y_0) \in \mathbb{R}_+^2 \mapsto IF((x_0,y_0); S^{m,x}, F) = \frac{\partial}{\partial\lambda} S^{m,x}(F + \lambda(\Delta_{(x_0,y_0)} - F))\Big|_{\lambda=0+},$$
$$(x_0,y_0) \in \mathbb{R}_+^2 \mapsto IF((x_0,y_0); T^{\alpha,x}, F) = \frac{\partial}{\partial\lambda} T^{\alpha,x}(F + \lambda(\Delta_{(x_0,y_0)} - F))\Big|_{\lambda=0+},$$

where $\Delta_{(x_0,y_0)}(u,v) = 1(x_0 \le u, y_0 \le v)$ for any $(u,v) \in \mathbb{R}^2$. Under H3′, it follows from Aragon, Daouia and Thomas-Agnan (2005, see their proof of Theorem 4.1) that

$$IF((x_0,y_0); T^{\alpha,x}, F) = \frac{\alpha 1(x_0 \le x) - 1(x_0 \le x, y_0 \le q_\alpha(x))}{f(q_\alpha(x)|x) F_X(x)}. \quad (3.1)$$

It is also established that

$$\widehat{q}_{\alpha,n}(x) - q_\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} IF((X_i,Y_i);T^{\alpha,x},F) + R_n^{\alpha,x},$$

where $\sqrt{n}R_n^{\alpha,x}$ converges in probability to 0 as $n \to \infty$. Thus $IF((X_i,Y_i);T^{\alpha,x},F)$ represents the approximate contribution, or influence, of the observation (X_i,Y_i) toward the estimation error $\widehat{q}_{\alpha,n}(x) - q_\alpha(x)$.

On the other hand, it can be easily seen that

$$IF((x_0,y_0);S^{m,x},F) = \frac{m}{F_X(x)}1(x_0 \le x)\int_0^\infty F^{m-1}(y|x)\,[F(y|x) - 1(y_0 \le y)]\,dy. \tag{3.2}$$

We also have

$$\sqrt{n}(\widehat{\varphi}_{m,n}(x) - \varphi_m(x)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF((X_i,Y_i);S^{m,x},F) + o_p(1) \quad \text{as} \quad n \to \infty.$$

So here also, the influence function measures the asymptotic bias caused by contamination in the observations (X_i,Y_i), $i = 1,\dots,n$.

Let us start with the robustness properties of the nonparametric estimator of the order-m expected frontier. This frontier is known to be more robust to extreme values than the standard nonparametric envelopment estimators FDH and DEA because it does not envelop all the observed data points. Moreover, the fact that $IF((x_0,y_0);S^{m,x},F)$ is zero when $x_0 > x$ ensures that the estimator $\widehat{\varphi}_{m,n}(x)$ is not influenced by outlying points (X_i,Y_i) with $X_i > x$, for any sample size n. But this estimator is not B-robust since its gross-error sensitivity satisfies

$$\begin{aligned}
\gamma^*(S^{m,x},F) &= \sup_{(x_0,y_0)\in\mathbb{R}_+^2} |IF((x_0,y_0);S^{m,x},F)| \\
&= \frac{m}{F_X(x)} \sup_{y_0\in\mathbb{R}_+} \left| \int_0^\infty F^{m-1}(y|x)\,[F(y|x) - 1(y_0 \le y)]\,dy \right| \\
&\ge \frac{m}{F_X(x)} \sup_{y_0>\varphi(x)} \left| \int_0^\infty F^{m-1}(y|x)\,[F(y|x) - 1(y_0 \le y)]\,dy \right| \\
&= \frac{m}{F_X(x)} \sup_{y_0>\varphi(x)} \int_0^{y_0} F^m(y|x)dy \\
&\ge \frac{m}{F_X(x)} \left[\int_0^{\varphi(x)} F^m(y|x)dy + \sup_{y_0>\varphi(x)} \int_{\varphi(x)}^{y_0} F^m(y|x)dy \right] \\
&= \frac{m}{F_X(x)} \left[\int_0^{\varphi(x)} F^m(y|x)dy + \sup_{y_0>\varphi(x)} (y_0 - \varphi(x)) \right] = \infty.
\end{aligned}$$

This reflects the fact that even a single outlier (X_i, Y_i) with a level of input X_i close to x on the left-hand side, if it is far enough from the cloud of data points in the direction of Y, can attract $\widehat{\varphi}_{m,n}(x)$ nearly to its outlying output Y_i. Besides this deficiency, the local-shift sensitivity defined as the smallest Lipschitz constant the influence function obeys, i.e.,

$$\lambda^*(S^{m,x}, F) = \sup_{s \neq t \in \mathbb{R}^2} |IF(s; S^{m,x}, F) - IF(t; S^{m,x}, F)|/||s - t||,$$

is infinite too since the indicator function $x_0 \mapsto 1(x_0 \leq x)$ has a jump at x. By $|| \cdot ||$, we denote the usual Euclidean norm on $\mathbb{R}^2$. This means that the nonparametric estimator $\widehat{\varphi}_{m,n}(x)$ may also be sensitive to rounding errors.

However, we can show under assumption K1 that $\gamma^*(S^{m,x}, F)$ is finite as follows. Putting $G_\lambda = F + \lambda(\Delta_{(x_0,y_0)} - F)$, the influence function $IF((x_0, y_0); S^{m,x}, F)$ is given by $(\partial/\partial\lambda)S^{m,x}(G_\lambda)|_{\lambda=0+}$. Since $G_\lambda(y|x) := G_\lambda(x, y)/G_\lambda(x, \infty) \to F(y|x)$ as $\lambda \searrow 0$ for every $y \in \mathbb{R}$, we obtain the weak convergence of the conditional distribution functions $G_\lambda(\cdot|x) \rightsquigarrow F(\cdot|x)$, which implies the weak convergence of the quantile functions $G_\lambda^{-1}(\cdot|x) \rightsquigarrow F^{-1}(\cdot|x)$ (i.e., $G_\lambda^{-1}(t|x) \to F^{-1}(t|x)$ as $\lambda \searrow 0$ at every $t \in [0, 1]$, where $F^{-1}(\cdot|x)$ is continuous) in view of a van der Vaart's Lemma (1998, Lemma 21.2, p.305). Hence, since $\lim_{t \nearrow 1} F^{-1}(t|x) = \lim_{t \nearrow 1} q_t(x) = q_1(x) = F^{-1}(1|x)$, we have $\varphi_{G_\lambda}(x) := G_\lambda^{-1}(1|x) \to F^{-1}(1|x) = \varphi(x)$ as $\lambda \searrow 0$. By using the fact that $\varphi(x) < \nu$ in view of K1, we therefore obtain $\varphi_{G_\lambda}(x) < \nu$ for all λ sufficiently small. Thus $S^{m,x}(G_\lambda) = \int_0^{\varphi_{G_\lambda}(x)} (1 - [G_\lambda(y|x)]^m)\, dy$ is well defined for λ small enough, and can be expressed as

$$S^{m,x}(G_\lambda) = \int_0^\nu (1 - [G_\lambda(y|x)]^m)\, dy.$$

Finally, by deriving with respect to λ, we get

$$IF((x_0, y_0); S^{m,x}, F) = \frac{m}{F_X(x)} 1(x_0 \leq x) \int_0^\nu F^{m-1}(y|x)\, [F(y|x) - 1(y_0 \leq y)]\, dy,$$

and so the gross-error sensitivity is finite and such that

$$\begin{aligned} \gamma^*(S^{m,x}, F) &= \frac{m}{F_X(x)} \sup_{y_0 \in \mathbb{R}} \left| \int_0^\nu F^{m-1}(y|x)\, [F(y|x) - 1(y_0 \leq y)]\, dy \right| \\ &\leq \frac{\nu m}{F_X(x)}, \end{aligned}$$

$$\begin{aligned} \gamma^*(S^{m,x}, F) &\geq \frac{m}{F_X(x)} \int_0^\nu F^m(y|x) dy \\ &\geq \frac{m}{F_X(x)} \int_{\varphi(x)}^\nu F^m(y|x) dy = (\nu - \varphi(x))m/F_X(x) > 0. \end{aligned}$$

The lower and upper bounds of $\gamma^*(S^{m,x}, F)$ indicate that the expected order-m frontiers are all the more sensitive to extreme values when the order m is large. Indeed,

$$\lim_{m\nearrow\infty} \gamma^*(S^{m,x}, F) = \infty. \tag{3.3}$$

This means in particular that $\widehat{\varphi}_{m,n}(x)$, when considered as an estimator of the true frontier function $\varphi(x) = \lim_{m\nearrow\infty} \varphi_m(x)$, may be influenced by extreme values or outliers.

Let us now turn to robustness characteristics of the empirical estimator of the nonstandard conditional α-quantile. As with $\widehat{\varphi}_{m,n}$, the frontier $\widehat{q}_{\alpha,n}$ is more robust to extremes than the FDH and DEA estimators in the sense that it does not envelop all the observed data points. The estimator $\widehat{q}_{\alpha,n}(x)$ also rejects outlying observations using more inputs than x since $IF((x_0, y_0); T^{\alpha,x}, F) = 0$ when $x_0 > x$, for any $y_0 \in \mathbb{R}_+$. But, unlike $\widehat{\varphi}_{m,n}(x)$, it possesses a finite gross-error sensitivity without resorting to assumption K1:

$$\gamma^*(T^{\alpha,x}, F) = \sup_{(x_0,y_0)\in\mathbb{R}^2_+} |IF((x_0, y_0); T^{\alpha,x}, F)| = \frac{\max(\alpha, 1-\alpha)}{f(q_\alpha(x)|x)F_X(x)}. \tag{3.4}$$

Moreover, if $F(\cdot|x)$ is continuously differentiable on the support $[0, \varphi(x)]$ with strictly positive derivative $f(\cdot|x)$, then by using the fact that $q_\alpha(x) \nearrow \varphi(x)$ as $\alpha \nearrow 1$, we obtain

$$\lim_{\alpha\nearrow 1} \gamma^*(T^{\alpha,x}, F) = \frac{1}{f(\varphi(x)|x)F_X(x)}. \tag{3.5}$$

This implies that, unlike $\{\widehat{\varphi}_{m,n}(x)\}$, the estimators $\{\widehat{q}_{\alpha,n}(x)\}$ can be resistant to outliers even for large values of α. Nevertheless, this is not necessarily true when the conditional density $f(\cdot|x)$ is zero at the frontier $\varphi(x)$. We will see numerical illustrations of this case in Section 4.

Because of the irregularity of the influence function at $(x, q_\alpha(x))$, the local-shift sensitivity $\lambda^*(T^{\alpha,x}, F)$ is infinite, which means that $\widehat{q}_{\alpha,n}(x)$ may be susceptible to rounding errors. However, this is much less important than the fact that $\gamma^*(T^{\alpha,x}, F)$ is finite. We summarize the above results in the following theorem.

Theorem 3.0.3. *Let $m \in \mathbb{N}^*$ and $\alpha \in]0, 1[$ be fixed orders, and let $x \in \mathbb{R}_+$ be such that $F_X(x) > 0$.*

1. *The sequence $\{\widehat{\varphi}_{m,n}(x)\}_n$ has infinite gross-error sensitivity $\gamma^*(S^{m,x}, F)$ unless K1 holds. In any case, $\lim_{m\nearrow\infty} \gamma^*(S^{m,x}, F) = \infty$.*
2. *Under assumption H3′, $\{\widehat{q}_{\alpha,n}(x)\}_n$ has the finite gross-error sensitivity $\gamma^*(T^{\alpha,x}, F)$ given at (3.4). Furthermore, if $F(\cdot|x)$ is continuously differentiable on the support $[0, \varphi(x)]$ with strictly positive derivative $f(\cdot|x)$, then $\gamma^*(T^{\alpha,x}, F)$ achieves the finite limit (3.5) as $\alpha \nearrow 1$.*

In both cases the local-shift sensitivity λ^* is infinite.

Note that $\{S^{m,x}(\widehat{F}_n)\}$ and $\{T^{\alpha,x}(\widehat{F}_n)\}$ do not in general estimate the same quantity, but in the limiting case where m tends to infinity and α to one, the sequences coincide with $\{\widehat{\varphi}_n(x)\}$ and can be viewed as estimators of the true full frontier $\varphi(x)$. Results (3.3) and (3.5) indicate then that extreme order-α frontiers are more robust than extreme order-m frontiers for estimating $\varphi(x)$. It is also important to note that, if we choose the order α as a function of n such that $\lim_{n\to\infty} n^{3/2}(1-\alpha(n)) = 0$, the functional $T^{\alpha(n),x}(\widehat{F}_n) = \widehat{q}_{\alpha(n),n}(x)$ estimates the upper frontier $\varphi(x)$ itself, as proved in Aragon, Daouia and Thomas-Agnan (2005). Likewise, if $m(n) = O(n\log(n))$ then $S^{m,x}(\widehat{F}_n) = \widehat{\varphi}_{m(n),n}(x)$ estimates $\varphi(x)$ and converges to the same Weibull distribution as $\widehat{\varphi}_n(x)$ and $T^{\alpha(n),x}(\widehat{F}_n)$ (see Cazals, Florens and Simar (2002)). The advantage of quantile-type frontiers is seen by comparing $\lim_{n\to\infty}\gamma^*(\cdot,F)$ of the estimators $\{S^{m(n),x}(\widehat{F}_n)\}$ and $\{T^{\alpha(n),x}(\widehat{F}_n)\}$ of $\varphi(x)$.

Theorem 3.0.4. *Let $x \in \mathbb{R}_+$ be such that $F_X(x) > 0$ and let $\{\alpha(n), n \geq 1\}$ and $\{m(n), n \geq 1\}$ be nondecreasing sequences such that $0 < \alpha(n) < 1$, $\lim_{n\to\infty}\alpha(n) = 1$, $m(n) \geq 1$ and $\lim_{n\to\infty} m(n) = \infty$.*

1. *$\gamma^*(S^{m(n),x}, F)$ is infinite for any n unless $K1$ holds, and $\lim_{n\nearrow\infty}\gamma^*(S^{m(n),x}, F) = \lim_{m\nearrow\infty}\gamma^*(S^{m,x}, F) = \infty$.*
2. *If $F(\cdot|x)$ is differentiable at $q_{\alpha(n)}(x)$ with strictly positive derivative $f(q_{\alpha(n)}(x)|x)$, then $\gamma^*(T^{\alpha(n),x}, F) = \max(\alpha(n), 1-\alpha(n))/f(q_{\alpha(n)}(x)|x)F_X(x)$. If $F(\cdot|x)$ is continuously differentiable on $[0,\varphi(x)]$ with derivative $f(\cdot|x) > 0$, then $\lim_{n\nearrow\infty}\gamma^*(T^{\alpha(n),x}, F) = \lim_{\alpha\nearrow 1}\gamma^*(T^{\alpha,x}, F) < \infty$.*

The explicit values of the orders $m(n)$ and $\alpha(n)$ are not available. In practice, we can choose $\widehat{\varphi}_{m(n),n}$ and $\widehat{q}_{\alpha(n),n}$ to be simply the usual FDH frontier if there are no influential observations in the data. In presence of such observations, we can determine the values $m(n)$ and $\alpha(n)$ by using the simple tool explained below, which also allows one to identify potential outliers.

A methodology for outliers detection using gross-error sensitivity

This is achieved through a sensitivity analysis of extreme order-α frontiers. We propose to choose several large values of α, say, $\alpha = 0.97$, 0.98, 0.99, 0.995, 0.999. Then the basic tool is a plot of $\gamma^*(T^{\alpha,x}, F)$ as a function of x, for the different values of α. By construction, if there are no outliers, the corresponding curve to each value of α should have homogeneous fluctuations (small jumps followed by smooth decreasing slopes). Any strong deviation of at least one of these order-α curves indicates the potential existence of outliers. For instance, if a curve shows a severe jump at a point x_i followed by an immediate fall, the

FDH observation $(x_i, y_i = \widehat{\varphi}_n(x_i))$ is a potential outlier. Indeed, if the suspicious point $(x_i, \widehat{\varphi}_n(x_i))$ is far enough from the cloud of data points in the direction of Y, then the quantile frontiers of extreme orders may be attracted by the outlier $(x_i, \widehat{\varphi}_n(x_i))$, but they come back down immediately, which generates a "free fall" of the empirical version of γ^* after the jump. The next section illustrates the idea with a data set.

As illustrated in Aragon, Daouia and Thomas-Agnan (2005, see Examples 1−3), the extreme order-m frontiers are more sensitive to outliers than the order-α frontiers that seems more appropriate for identifying anomalous data. But they continue to grow after each jump, and this might make the identification of potential outliers more difficult than with extreme quantile frontiers.

Our semi-automatic procedure based on the analysis of order-α (respectively order-m) curves offers also an appealing and useful way to determine $\alpha(n)$ (respectively $m(n)$). The order $\alpha(n)$ (respectively $m(n)$) will correspond to the largest value of α (respectively m) such that the order-α (respectively order-m) curve shows no strong deviation when plotted. For instance, if the curve corresponding to $\alpha = 0.99$ shows no strong deviation whereas that corresponding to $\alpha = 0.995$ is dramatically influenced, then $0.99 \leq \alpha(n) < 0.995$. We can improve the bracket for $\alpha(n)$ with new plots using other values $\alpha \in]0.99; 0.995[$. We obtain $m(n)$ in the same way. Even more strongly we can pick a precise value of $m(n)$ since the order m is discrete.

4. Numerical Illustration

In this section, we present two simulated examples to illustrate the case where there is no mass at the upper boundary, i.e., when $f(\varphi(x)|x) = 0$. These examples are used in Florens and Simar (2005) where the authors show, in particular, that the order-m frontier estimators are more robust to outliers than the OLS-shifted frontier (Greene (1980)). We also show how the gross-error sensitivity can be used as a diagnostic tool to detect anomalous data.

4.1. Example 1

We first consider a case where the monotone boundary of the support of (X, Y) is linear. We choose (X, Y) uniformly distributed over the region $D = \{(x, y)|0 \leq x \leq 1,\ 0 \leq y \leq x\}$. The upper boundary is given by the frontier function $\varphi(x) = x$. Here, the conditional df is $F(y|x) = 2x^{-1}y - x^{-2}y^2$, for $0 < x \leq 1$ and $0 \leq y \leq x$, and the conditional α-quantile is $q_\alpha(x) = x(1 - \sqrt{1-\alpha})$, for $0 < x \leq 1$ and $0 \leq \alpha \leq 1$. The gross-error sensitivity for the sequence of α-quantile frontier estimators is then given by $\gamma^*(T^{\alpha,x}, F) = \max(\alpha, 1-\alpha)/(2x\sqrt{1-\alpha})$. The order-$m$ frontier can be computed as $\varphi_m(x) = x(1 - A_m)$, where $A_m =$

$\sum_{j=0}^{m} m!(-1)^{m-j}2^j/(j!(m-j)!(2m-j+1))$. For $0 < x < 1$ and $(x_0, y_0) \in \mathbb{R}^2$, the influence function is given by

$$IF((x_0,y_0); S^{m,x}, F) = \begin{cases} 0, & \text{if} \quad x_0 > x, \\ \frac{m}{F_X(x)} \mathbb{I}^{m,x}(y_0), & \text{otherwise,} \end{cases}$$

where

$$\mathbb{I}^{m,x}(y_0) = \begin{cases} I_{x,m}(0,1), & \text{if} \quad y_0 \geq 1, \\ I_{x,m}(0,1) - I_{x,m-1}(y_0,1), & \text{if} \quad 0 < y_0 < 1, \\ I_{x,m}(0,1) - I_{x,m-1}(0,1), & \text{otherwise,} \end{cases} \tag{4.1}$$

$$I_{x,m}(a,b) = \int_a^b F^m(y|x)dy = \sum_{j=0}^{m} m!(-1)^{m-j}2^j x^{j-2m} \frac{[b^{2m-j+1} - a^{2m-j+1}]}{j!(m-j)!(2m-j+1)}.$$

Therefore, the order-m gross-error sensitivity can be computed as

$$\gamma^*(S^{m,x}, F) = \frac{m}{F_X(x)} \max\left\{I_{x,m}(0,1); I_{x,m-1}(0,1) - I_{x,m}(0,1)\right\}. \tag{4.2}$$

In this particular example, the order-α and order-m frontiers are both linear in x and can have the same slope. They coincide if and only if $\alpha = 1 - A_m^2$. In this case, the sequences $\{\widehat{q}_{\alpha,n}\}_n$ and $\{\widehat{\varphi}_{m,n}\}_n$ estimate the same frontier ($\varphi_m = q_\alpha$) and their reliability can be analyzed by comparing their γ^* values. In Table 1, we give $\gamma^*(T^{\alpha,x}, F)$ (respectively $\gamma^*(S^{m,x}, F)$) for the corresponding frontier estimators $\{\widehat{q}_{\alpha,n}(x)\}$ (respectively $\{\widehat{\varphi}_{m,n}(x)\}$) when $x = 1/4$, $1/2$ and $3/4$.

As stated by Hampel, Ronchetti, Rousseeuw and Stahel (1986, p.43), the most important robustness requirement, besides qualitative robustness, is a low γ^*. From this basis, it is clear that $\{\widehat{q}_{\alpha,n}(x)\}$ is more robust than $\{\widehat{\varphi}_{m,n}(x)\}$ for estimating the linear partial frontier $\varphi_m(x) \equiv q_\alpha(x)$, since $\gamma^*(T^{\alpha,x}, F) < \gamma^*(S^{m,x}, F)$, as can be seen from Table 1. Another remark of interest lies in the limit case where $\gamma^*(S^{m,x}, F)$ explodes, whereas $\gamma^*(T^{\alpha,x}, F)$ remains small as $\alpha \to 1$ and $m \to \infty$. This indicates that $\{\widehat{q}_{\alpha,n}(x)\}$ is more resistant to extreme values than $\{\widehat{\varphi}_{m,n}(x)\}$ for estimating the full frontier $\varphi(x) = \lim_{\alpha \nearrow 1} q_\alpha(x) = \lim_{m \nearrow \infty} \varphi_m(x)$.

Note that for a fixed value of α or m, the corresponding γ^* decreases with respect to x, which indicates that the estimators $\widehat{q}_{\alpha,n}(x)$ and $\widehat{\varphi}_{m,n}(x)$ are more resistant to extreme values as x increases. This is no surprise due to the conditioning by $X \leq x$, since these nonparametric estimators are not so good at the border where the number of all points in the sample with input value smaller than x is very small.

Note also that for a fixed value of x, γ^* increases with respect to the orders α and m. This is natural since both nonparametric partial frontiers converge to the non-robust FDH frontier as $\alpha \to 1$ and $m \to \infty$.

Table 1. The gross-error sensitivities of the sequences $\{\widehat{\varphi}_{m,n}(x)\}_n$ and $\{\widehat{q}_{\alpha,n}(x)\}_n$ for estimating the linear frontier $\varphi_m(x) \equiv q_\alpha(x)$.

Results for Example 1

m	$\gamma^*(S^{m,1/4}, F)$	$\gamma^*(S^{m,1/2}, F)$	$\gamma^*(S^{m,3/4}, F)$
1	9	1.3333	0.9877
2	68	2.1333	1.6856
3	673	2.7429	2.2675
4	4785	3.2508	2.7799
5	4.5227e+04	3.6941	3.2431
10	1.3557e+09	5.4052	5.1040
15	5.1269e+13	6.6993	6.5178
20	1.4955e+18	27.2388	7.6771
25	5.5683e+22	153290	9.0947
30	1.6209e+27	1.2986e+10	352.793
$\alpha = 1 - A_m^2$	$\gamma^*(T^{\alpha,1/4}, F)$	$\gamma^*(T^{\alpha,1/2}, F)$	$\gamma^*(T^{\alpha,3/4}, F)$
0.5556	1.6667	0.8333	0.5556
0.7156	2.6833	1.3417	0.8944
0.7910	3.4607	1.7304	1.1536
0.8349	4.1092	2.0546	1.3697
0.8635	4.6752	2.3376	1.5584
0.9270	6.8598	3.4299	2.2866
0.9501	8.5102	4.2551	2.8367
0.9622	9.8913	4.9457	3.2971
0.9695	11.1027	5.5514	3.7009
0.9745	12.1947	6.0974	4.0649

4.2. Example 2

Let us now consider a more realistic example from an economic point of view. We choose a non-linear monotone upper boundary given by the Cobb-Douglas model $Y = X^{1/2} \exp(-U)$, where X is uniform on $[0, 1]$ and U, independent of X, is Exponential with parameter $\lambda = 3$. This is a standard example in the literature (see, e.g., Gijbels, Mammen, Park and Simar (1999), Florens and Simar (2005) and Simar (2003)).

Here, the upper boundary of the support of (X, Y) is given by the frontier function $\varphi(x) = x^{1/2}$. For $0 < x \leq 1$ and $0 \leq y \leq \varphi(x)$, the conditional df is $F(y|x) = 3x^{-1}y^2 - 2x^{-3/2}y^3$ and, for $0 \leq \alpha \leq 1$, the conditional α-quantile is given by $q_\alpha(x) = x^{1/2}\mathbb{J}(\alpha)$, where $\mathbb{J}(\alpha) = \cos((\arccos(1-2\alpha) + 4\pi)/3) + 1/2$. The corresponding α-gross-error sensitivity is $\gamma^*(T^{\alpha,x}, F) = \max(\alpha, 1-\alpha)/(6x^{1/2}\mathbb{J}(\alpha)(1-\mathbb{J}(\alpha)))$. The order-$m$ frontier here can be computed as $\varphi_m(x) =$

$x^{1/2}(1 - B_m)$, where $B_m = \sum_{j=0}^{m} m!(-2)^{m-j}3^j/(j!(m-j)!(3m-j+1))$. For $0 < x < 1$ and $(x_0, y_0) \in \mathbb{R}^2$, its influence function $IF((x_0, y_0); S^{m,x}, F)$ is 0 if $x_0 > x$ and $(m|x)\mathbb{I}^{m,x}(y_0)$ otherwise, where $\mathbb{I}^{m,x}(y_0)$ is given by (4.1), with

$$I_{x,m}(a,b) = \sum_{j=0}^{m} m!(-2)^{m-j}3^j x^{\frac{j-3m}{2}} \frac{[b^{3m-j+1} - a^{3m-j+1}]}{j!(m-j)!(3m-j+1)}.$$

The corresponding m-gross-error sensitivity can be computed with (4.2) where $F_X(x) = x$.

In this particular case, the order-α and order-m frontiers are both log-linear in x and coincide if and only if $\alpha = (1 - \cos[3\arccos(1/2 - B_m) - 4\pi])/2$. For such a pair (α, m), the partial concave frontier $q_\alpha \equiv \varphi_m$ can be estimated by $\{\widehat{q}_{\alpha,n}\}$ as well as $\{\widehat{\varphi}_{m,n}\}$. The numerical results are displayed on Table 2. We remark here also that $\gamma^*(S^{m,x}, F)$ is larger than $\gamma^*(T^{\alpha,x}, F)$ and that this latter γ^* remains small for extreme values of α, which is not the case for the order-m gross-error sensitivity.

Table 2. γ^* of the sequences of estimators $\{\widehat{\varphi}_{m,n}(x)\}_n$ and $\{\widehat{q}_{\alpha,n}(x)\}_n$ for estimating the concave frontier $q_\alpha(x) \equiv \varphi_m(x)$, where $\alpha = (1 - \cos[3\arccos(1/2 - B_m) - 4\pi])/2$.

Results for Example 2

m	$\gamma^*(S^{m,1/4}, F)$	$\gamma^*(S^{m,1/2}, F)$	$\gamma^*(S^{m,3/4}, F)$
1	4	1.1716	0.7514
2	10.9714	1.8309	1.1971
3	49.3714	2.3261	1.5603
4	154.6741	2.7359	1.8797
5	807.1289	3.0928	2.1697
10	670000	4.4721	3.3592
15	877380000	5.928	4.293
25	1.7296e+19	8.3577e+09	9.3482e+03
α	$\gamma^*(T^{\alpha,1/4}, F)$	$\gamma^*(T^{\alpha,1/2}, F)$	$\gamma^*(T^{\alpha,3/4}, F)$
0.5	0.6667	0.4714	0.3849
0.6886	0.9832	0.6952	0.5676
0.7749	1.2138	0.8583	0.7008
0.8240	1.4031	0.9921	0.8101
0.8557	1.567	1.108	0.9047
0.9242	2.1957	1.5526	1.2677
0.9486	2.6698	1.8878	1.5414
0.9997	36.6755	25.9335	21.1746

We repeated the same exercise with many other values of x and (m, α) leading to the same advantage of the quantile-type estimators when estimating either the true partial frontier or the full frontier.

4.3. Example 3

We consider a data set concerning the delivery activity of the postal services in France. The data comes from 4,000 post offices observed in 1994, and have been previously analyzed in Cazals, Florens and Simar (2002) and Aragon, Daouia and Thomas-Agnan (2005). For each post office i, the input x_i is the labor cost measured by the quantity of labor, and the output y_i is the volume of delivered mail in number of objects. As can be seen on Figure 1, the data set contains at least the two outlying observations $(965, 7207)$ and $(1051, 11762)$.

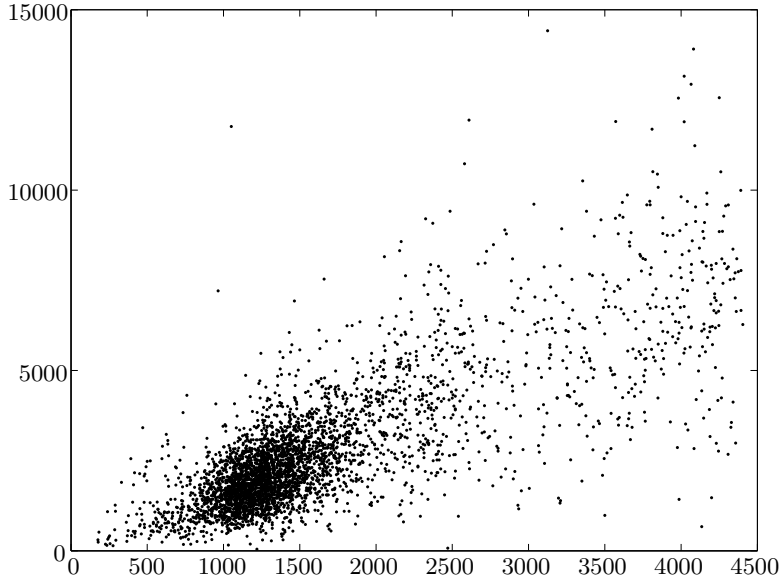

Figure 1. Plot of the volume of delivered mail against the quantity of labor for the 4,000 post offices.

We first propose to make a plot of the influence function of an order-m frontier estimate and of an order-α frontier estimate. Take $x_0 \in [\min_i x_i; \max_i x_i]$ and $y_0 \in [\min_i y_i; \max_i y_i]$. The influence of (x_0, y_0) on an order-m frontier estimate $\widehat{\varphi}_{m,n}(x)$ (respectively on an order-α estimate $\widehat{q}_{\alpha,n}(x)$) can be measured by the value of a sample version of the influence function given by (3.2) (respectively (3.1)). The sample versions are obtained by replacing the unknown quantities

$F_X(x)$, $q_\alpha(x)$, $F(\cdot|x)$ and $f(q_\alpha(x)|x)$ with estimated quantities. $F_X(x)$ and $q_\alpha(x)$ are simply estimated by $\widehat{F}_{X,n}(x)$ and $\widehat{q}_{\alpha,n}(x)$, whereas the estimator of $F(\cdot|x)$ is a triweight kernel estimator (with an empirical choice of the bandwith) from which we also derive an estimator of the density $f(q_\alpha(x)|x)$.

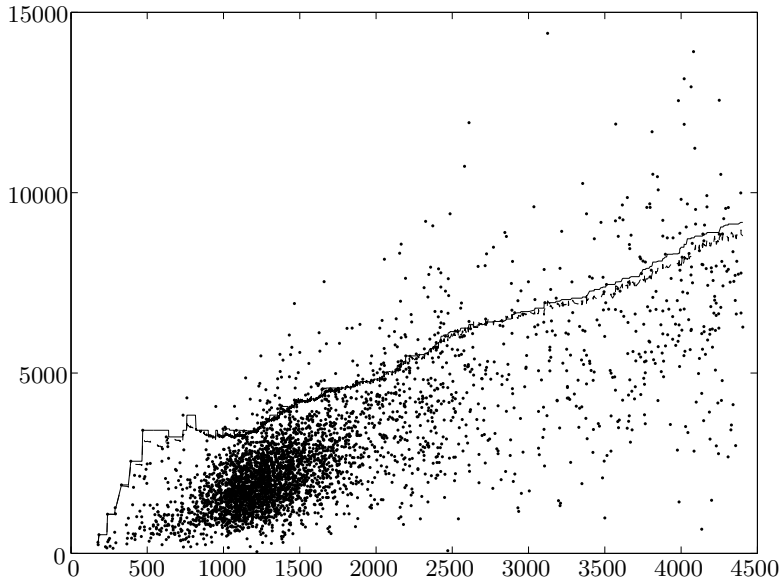

Figure 2. Plot of the quantile frontier estimate for $\alpha = 0.99$ (solid line) and the order-$m = 50$ estimate (dashed line) for the 3998 post offices (2 outliers removed).

In order to compare the estimation methods, we drop the two outliers and consider a quantile frontier estimate with $\alpha = 0.99$ and an order-m estimate with $m = 50$. Figure 2 shows that for this parameters choice, the order-α frontier estimate is very close to the order-m estimate. Considering again the 4,000 observations (the two outliers included), Figure 3 gives the plots of the empirical influence function for the quantile estimate of order $\alpha = 0.99$ and for the order-m ($m = 50$) estimate at the value $x = 1,338$, the median of the x_i's. It is clear that the influence function of the order-50 estimate is much larger than the influence function of the quantile estimate. This indicates that $\widehat{q}_{0.99,n}(x)$ is more resistant to the outlying post offices than $\widehat{\varphi}_{50,n}(x)$. Note also that for given α and x, the influence function of the quantile frontier estimator, as well as its empirical counterpart, takes only three different values according to the values (x_0, y_0).

Recall that we have

$$IF((x_0,y_0);T^{\alpha,x},F)=\begin{cases}\alpha\, c_\alpha(x)>0, & \text{if } x_0\le x,\ y_0>q_\alpha(x),\\ (\alpha-1)\, c_\alpha(x)<0, & \text{if } x_0\le x\ y_0\le q_\alpha(x),\\ 0, & \text{if } x_0>x,\end{cases}$$

with $c_\alpha(x)=1/f(q_\alpha(x)|x)F_X(x)$.

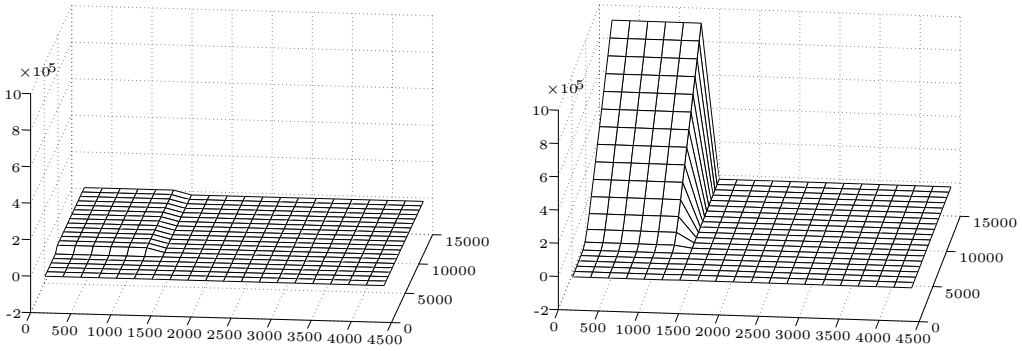

Figure 3. Plots of the empirical influence function for the 0.99 quantile estimate (left plot) and for the 50-order estimate (right plot) at the median of the x_i's for the 4,000 post offices.

Now, we focus on the quantile frontier estimates and illustrate the use of the influence function as a tool for detecting influential observations. In practice, we are interested in quantile frontier estimates for large values of α (at least $\alpha\ge 1/2$), so we have $\alpha\, c_\alpha(x)\ge|\alpha-1|\, c_\alpha(x)$ and the maximum influence (which corresponds to the gross-error sensitivity) is achieved for $x_0<x$ and $y_0>q_\alpha(x)$. Figure 4 shows the function $x\mapsto\alpha\hat{c}_\alpha(x)$, which is a sample version of the gross-error sensitivity, for $\alpha=0.99$ (left plot) and $\alpha=0.999$ (right plot). On both plots, we can see that for small x, the influence function is quite large due to a border problem. Indeed, there are too few observations for estimating the conditional df $F(\cdot|x)$. For $\alpha=0.99$, apart from the smallest values of x, no observation is particularly influential. That's no more true when considering $\alpha=0.999$. In this case, we detect that the two outlying observations previously mentioned heavily influence the estimate. This can be explained as follows: the quantile frontier of order 0.99 does not allow one to identify the two outlying post offices because it is very resistant to these outliers, as shown in Figure 4 (left plot), whereas the frontier of order 0.999 coincides with the non-robust FDH frontier on $[\min_i x_i;1051]$, as seen from Figure 7 of Aragon, Daouia and Thomas-Agnan

(2005). Therefore, the use of very extreme-order α frontiers is necessary to detect anomalous data. Now, for the choice of the order $\alpha(n)$ of the robust quantile estimator of φ, Figure 4 indicates to us the choice of $\alpha(n)$ in the interval $[0.99; 0.999[$.

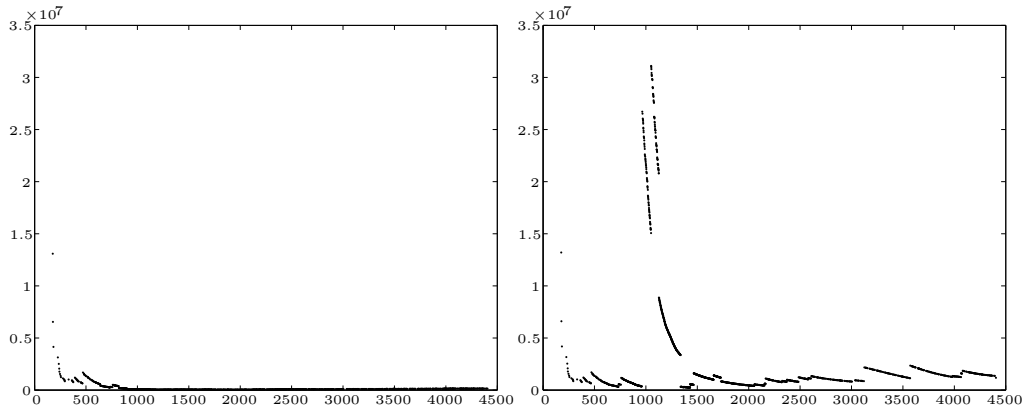

Figure 4. Plots of the empirical gross-error sensitivity of quantile estimates for $\alpha = 0.99$ (left plot) and $\alpha = 0.999$ (right plot) against the value of x for the 4,000 post offices.

5. Concluding remarks

The most popular nonparametric methods of estimating a monotone boundary, such as the DEA or the FDH methods, are highly non-robust. Cazals, Florens and Simar (2002) and Aragon, Daouia and Thomas-Agnan (2005) have proposed alternatives which are shown here to satisfy some interesting robustness properties. Both methods are qualitatively and B-robust, but the gross-error sensitivity of the order-m frontier tends to infinity with m. On the contrary, as soon as we assume that the conditional density function is not null and is continuous on its support, the gross-error sensitivity of the α-quantile estimators tends to a finite limit when α tends to one. In the case where this last assumption is not fulfilled, we consider two simulated examples and propose an empirical comparison of the gross-error sensitivities for different values of α and m. The results clearly favor the quantile approach. We would like to point out that we are concerned with nonparametric estimators and a nonparametric model. The use of robustness concepts such as the qualitative robustness and indicators derived from the influence function is usually devoted to parametric statistics in parametric models (Hampel, Ronchetti, Rousseeuw and Stahel (1986) and Huber (1981)). But, since the estimators we consider depend in a quite simple way on the joint cumulative distribution function of the data, we can consider

robustness properties in a nonparametric context. Note that this is no more possible in a standard quantile regression context because the estimators are not simply defined as functionals of the joint cumulative distribution function. There is another difference between the usual quantile approach and the non-standard one proposed in Aragon, Daouia and Thomas-Agnan (2005). While in the latter case, the influence function is bounded in both coordinates (input and output), in the former case the quantiles (considered in a parametric context) have only a bounded influence function in the output argument.

Acknowledgement

We are grateful to the referees and an associate editor for their helpful comments and suggestions.

References

Aragon, Y., Daouia, A. and Thomas-Agnan, C. (2005). Nonparametric frontier estimation: a conditional quantile-based approach. *Econom. Theory* **21**, 358-389.

Banker, R. D. (1993). Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. *Management Sci.* **39**, 1265-1273.

Cazals, C., Florens, J-P. and Simar, L. (2002). Nonparametric frontier estimation: a robust approach. *J. Econometrics* **106**, 1-25.

Daouia, A. and Ruiz-Gazen, A. (2004), Robust nonparametric frontier estimators: qualitative robustness and influence function. Technical report, GREMAQ et LSP, Université de Toulouse (http://w3.univ-tlse1.fr/GREMAQ/Statistique/ar0903.pdf).

Deprins, D., Simar, L. and Tulkens, H. (1984). Measuring labor efficiency in post offices. In *The Performance of Public Enterprises: Concepts and Measurements* (Edited by M. Marchand, P. Pestieau and H. Tulkens), 243-267, North-Holland, Amsterdam.

Farrell, M. J. (1957). The measurement of productive efficiency. *J. Roy. Statist. Soc. Ser. A* **120**, 253-281.

Florens, J-P. and Simar, L. (2005). Parametric approximations of nonparametric frontiers. *J. Econometrics* **124**, 91-116.

Gijbels, I., Mammen, E., Park, B. U. and Simar, L. (1999). On estimation of monotone and concave frontier functions. *J. Amer. Statist. Assoc.* **94**, 220-228.

Greene, W. H. (1980). Maximum likelihood estimation of econometric frontier functions. *J. Econometrics* **13**, 27-56.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42**, 1887-1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383-393.

Hampel, F. R., Ronchetti, E. M. and Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Korostelev, A., Simar, L. and Tsybakov, A. B. (1995). Efficient estimation of monotone boundaries. *Ann. Statist.* **23**, 476-489.

Markovitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Wiley, New York.

Park, B. U., Simar, L. and Weiner, C. (2000). The FDH estimator for productivity efficiency scores. *Econom. Theory* **16**, 855-877.

Prohorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* **1**, 157-214.

Rousseeuw, P. J. (1981). A new infinitesimal approach to robust estimation. *Z. Wahrsch. Verw. Gebiete* **56**, 127-132.

Seiford, L. M. (1996). Data Envelopment Analysis: The Evolution of the State-of-the-Art (1978-1995). *J. Productiv. Anal.* **7**, 99-138.

Simar, L. (2003). Detecting Outliers in Frontier Models: A Simple Approach. *J. Productiv. Anal.* **20**, 391-424.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Laboratoire de Statistique et Probabilités, Université Paul Sabatier, U.M.R. CNRS 5883, 118 route de Narbonne, 31062 Toulouse Cedex 04, France.

E-mail: daouia@cict.fr

GREMAQ, Université des Sciences Sociales, U.M.R. CNRS 5604
21 allée de Brienne, 31042 Toulouse Cedex, France.

E-mail: ruiz@cict.fr

(Received March 2004; accepted November 2004)