Estimation of a support curve via order statistics

by

I. Gijbels[1,2] and L. Peng[2]

March 1999

Abstract

This paper deals with nonparametric estimation of the boundary curve of the support of a bivariate density function. This estimation problem arises in various contexts, such as for example scatterpoint image analysis and frontier estimation in econometrics. The setup in this paper is a general one, allowing the bivariate density function to be infinite, bounded away from zero or zero at the boundary. Two estimators for the boundary curve are introduced, both based on order statistics. The asymptotic distribution of the estimators and their rate of convergence is established. Via a comparison of the rates of convergence we recommend which estimator to use in a particular situation. Both estimators can be used as an initial estimator in a two-stage procedure, designed for getting a better estimation. Simulations studies illustrate the finite-sample behaviour of the estimators and the proposed two-stage procedure.

AMS 1991 subject classification. Primary 62G07; secondary 62G30.

Key words. Asymptotic distribution, Boundary, Endpoint estimation, Extreme-value index, Frontier function, Maximum, Order Statistics, Pickands' estimator.

[1] Institut de Statistique, Université catholique de Louvain, 20 Voie du Roman Pays, Louvain-la-Neuve, Belgium.
[2] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

1 Introduction

Estimation of the boundary of the support of a density function is a problem arising in various contexts. We give some examples. In scatterpoint image analysis, the boundary is typically the interface of areas of different intensities or different colour tones. See for example Korostelev and Tsybakov (1993) and Mammen and Tsybakov (1995).

Boundary estimation problems form also the core of estimating an efficiency curve in productivity analysis. Here the data typically consist of pairs (X_i, Y_i) where X_i represents the input (for example labor, energy, capital) used to produce an output (for example goods produced) Y_i in a certain firm i. These variables are observed for a number of firms, and the aim is to investigate the economic efficiency of the firms. A possible way to measure this efficiency is by looking at the maximum level of output which is attainable for a given level of input. This problem translates into studying the upper boundary of the set of possible inputs X and outputs Y, the so-called production frontier or frontier function. Of course all observations (X_i, Y_i) lie below this upper boundary curve. So also here the major task is to estimate the support curve of a bivariate density (in case X and Y are univariate). There is a vast econometrics literature on frontier estimation, but the statistical aspects of the estimation procedures have been explored only during the last decade. See Grosskopf (1996), Seiford (1996) and Simar (1996) for recent contributions to statistical aspects of frontier estimation. Econometrics considerations often lead to the assumption that the support curve is concave and monotone. A popular estimation technique is then the Data Envelopment Analysis (DEA) estimator, which is the lowest concave monotone increasing function covering all sample points. See for example Charnes, Cooper, Lewin and Seiford (1995). Statistical studies of this estimation procedure can be found in Banker (1993), Korostelev, Simar and Tsybakov (1995a, b), Kneip, Park and Simar (1998) and Gijbels, Mammen, Park and Simar (1999), among others.

Other related work in boundary estimation includes Härdle, Park and Tsybakov (1995), Hall, Nussbaum and Stern (1997) and Hall, Park and Stern (1998).

The setup in this paper is as follows. Suppose that $(X_1, Y_1), \cdots, (X_n, Y_n)$ are i.i.d. with a density $f(x, y)$ in $I\!R^2$. The support of the density f is assumed to be of the form

$$\{(x, y)|f(x, y) \geq 0\} = \{(x, y)|y \leq g(x)\}, \tag{1.1}$$

So the function $y = g(x)$ is the locus of the curve above which the bivariate density f is zero. Further, we assume that as $y \uparrow g(x)$, the density $f(x,y)$ is an algebraic function of $g(x)-y$

$$f(x,y) = a(x,y)\,\{g(x)-y\}^{\alpha(x)}\left\{1+O\left((g(x)-y)^{\beta(x)}\right)\right\}, \tag{1.2}$$

where $a(x,y)>0$, $\alpha(x)>-1$ and $\beta(x)>0$.

According to the value of $\alpha(x)$ we can distinguish three important situations

$$f(x,g(x)) = \begin{cases} \infty & \text{if} \quad \alpha(x)<0 \\ \in(0,\infty) & \text{if} \quad \alpha(x)=0 \\ 0 & \text{if} \quad \alpha(x)>0, \end{cases}$$

i.e., the cases where the density function is infinite, bounded away from zero, or zero at the boundary. When $\alpha(x)\neq 0$, the quantity $\alpha(x)$ describes the rate at which the density tends to infinity (in case $\alpha(x)<0$) or to 0 (in case $\alpha(x)>0$) at the boundary.

Hall, Nussbaum and Stern (1997) focus on the case $\alpha(x)>1$, where it is known that estimators based on an increasing number of order statistics give optimal convergence rates. Polynomial estimators for the boundary function $g(x)$ have been studied in Hall, Park and Stern (1998).

In this paper we consider the case of general $\alpha(x)$. Our estimators are based on a fixed number of order statistics, in contrast to Hall, Nussbaum and Stern (1997) in which estimation is based on an increasing number of order statistics. When estimating $g(x_0)$, for a given point x_0, our methodology consists of considering a strip around x_0 and focusing then on the Y_i observations falling into this strip. Two estimators for $g(x_0)$ are proposed and studied: the maximum of all these Y_i observations, and an endpoint type of estimator based on three large order statistics of the Y_i's in the strip. These three order statistics are of the intermediate type, involving a sequence of integers $m(n)$ tending to infinity at a slower rate than n. The asymptotic distribution of the estimators and their rate of convergence are obtained. These results reveal that, in case $\beta(x_0)=\infty$, the endpoint estimator has a better rate of convergence than the maximum estimator if and only if $\alpha(x_0)>0$. When $\alpha(x_0)=0$ (the case of a density function bounded away from 0 at the boundary), both estimators have the same rate of convergence.

The final goal is to use one of these estimators in a first step of a two-stage estimation procedure. The idea is to use an initial estimator of $g(x_0)$ to "pull up" all Y_i observations in the small strip around x_0, by replacing all these Y_i's by the initial estimator of $g(x_0)$. Data outside the strip are left unchanged. As a result the transformed data will be more concentrated around the boundary curve $g(x)$ in this small strip. This will then be the starting situation for the second step, which consists of applying a good estimator for $g(x_0)$ to the more concentrated data. For illustrating this two-stage procedure we focus on the case of a concave and monotone increasing function g. It has been shown that the DEA estimator has good properties, and hence we opt for applying the DEA estimation procedure to the transformed data.

In the next section we define the two estimators, and give more details on the setup in this paper. The asymptotic distribution of the estimators and their rate of convergence are provided in Section 3. The two-stage procedure is illustrated in Section 4, and the finite-sample performance is investigated via a simulation study in Section 5.

2 Definition of the estimators

Throughout this paper we will assume that the density $f(x, y)$ of the bivariate random vector (X, Y) is of the form (1.1) and behaves as indicated in (1.2). We further introduce the following conditions on the functions $g(x)$, $a(x, y)$, $\alpha(x)$ and $\beta(x)$.

Condition A.

A1. $g(x) > 0$.

A2. g is a continuous function.

A3. The derivative function $g'(x)$ is Hölder continuous with exponent 1, i.e.

$$|g'(x_1) - g'(x_2)| \leq L|x_1 - x_2|,$$

with L a positive constant.

A4. $a(x, y)$, $\alpha(x)$ and $\beta(x)$ are Hölder continuous with exponent 1, i.e. for positive constants L_α, L_β, L_1 and L_2,

$$\begin{aligned}
&|\alpha(x_1) - \alpha(x_2)| \leq L_\alpha|x_1 - x_2| \\
&|\beta(x_1) - \beta(x_2)| \leq L_\beta|x_1 - x_2| \\
&|a(x_1, y_1) - a(x_2, y_2)| \leq L_1|x_1 - x_2| + L_2|y_1 - y_2|.
\end{aligned}$$

Let $(x_0, g(x_0))$ be a point at the boundary of the set defined in (1.1). To estimate $g(x_0)$ we look at observations Y_i falling in a strip around x_0, and estimate $g(x_0)$ via order statistics of these Y_i's. We consider a strip around x_0 of width $2h$, where $h = h_n > 0$ will play the role of a smoothing parameter. More precisely, let $Z_i = Y_i \mathbf{1}\{|X_i - x_0| \leq h\}$, $i = 1, \cdots, n$, and denote by $Z_{n,1} \leq \cdots \leq Z_{n,n}$ the order statistics of $Z_1, \cdots, Z_n$.

A first estimator of $g(x_0)$ is $Z_{n,n}$, the maximum of the Y_i observations falling in the $2h$-strip around x_0. Note that $Z_{n,n}$ depends on x_0 as well as on h. For simplicity of notation we do not explicitate this dependence.

Remark 1. Note that $g(x_0) > 0$ implies that $Z_{n,n} > 0$ with probability tending to one as n tends to infinity.

The maximum is just one way to estimate the endpoint $g(x_0)$ based on the observations $Z_1, \cdots, Z_n$. Other endpoint estimators have been proposed by Hall (1982), Dekkers and de Haan (1989) and Dekkers, Einmahl and de Haan (1989). Our second estimator follows the approach of Dekkers and de Haan (1989) which is based on Pickands' (1975) estimator of the index of an extreme-value distribution. This endpoint estimator is based on three order statistics.

Let $m = m(n)$ be an intermediate integer sequence such that $m < n/4$, $m \to \infty$ and $m/n \to 0$, as $n \to \infty$. Following Dekkers and de Haan (1989) we define the estimator of $g(x_0)$ as

$$\hat{g}_n(x_0) := \frac{Z_{n,n-m} - Z_{n,n-2m}}{2^{-\hat{\gamma}_n(x_0)} - 1} + Z_{n,n-m}, \tag{2.1}$$

with

$$\hat{\gamma}_n(x_0) := \log \frac{Z_{n,n-m} - Z_{n,n-2m}}{Z_{n,n-2m} - Z_{n,n-4m}} / \log 2. \tag{2.2}$$

Note that $\hat{g}_n(x_0)$ also depends on h.

Remark 2. The condition $g(x) > 0$ is in fact not a crucial one. In case $g(x) < 0$ we work with the random variables $W_i = \dfrac{Y_i}{\mathbf{1}\{|X_i - x_0| \leq h\}}$ instead of the Z_i's.

3 Main results

In this section we derive the asymptotic distribution of the estimators $Z_{n,n}$ and $\hat{g}_n(x_0)$ defined in the previous section.

3.1 Asymptotic distribution of the maximum estimator

The asymptotic distribution of the maximum estimator $Z_{n,n}$ of $g(x_0)$ is established in the next theorem.

Theorem 3.1.
Assume (1.2) and Conditions A1, A2 and A4. If h_n is such that $h_n \to 0$, $nh_n \to \infty$ and $nh_n^{2+\alpha(x_0)} \to 0$, then, for any $z < 0$,

$$\lim_{n\to\infty} P\left\{(nh_n)^{\frac{1}{1+\alpha(x_0)}}\left(Z_{n,n} - g(x_0)\right) \le z\right\} = \exp\left\{-\frac{2a(x_0, g(x_0))}{1+\alpha(x_0)}(-z)^{1+\alpha(x_0)}\right\}. \quad (3.1)$$

Remark 3. Note that the rate of convergence is independent of $\beta(x_0)$. Taking $h_n = cn^{-\delta}$, with c an arbitrary positive constant, the conditions on h_n in the theorem are equivalent to $1/(2+\alpha(x_0)) < \delta < 1$. The best rate of convergence is obtained by taking h_n as large as possible, i.e. by taking $h_n = cn^{-\frac{1}{2+\alpha(x_0)}-\varepsilon}$, with $0 < \varepsilon < 1 - \frac{1}{2+\alpha(x_0)}$ small. This choice of h_n leads to the optimal rate of convergence $n^{-\frac{1}{2+\alpha(x_0)}+\frac{\varepsilon}{1+\alpha(x_0)}}$. Note that in case of a density bounded away from 0 at the boundary (the case $\alpha(x_0) = 0$), this results into choosing $h_n = cn^{-\frac{1}{2}-\varepsilon}$, with $0 < \varepsilon < \frac{1}{2}$, giving the rate of convergence $n^{-\frac{1}{2}+\varepsilon}$.

Proof of Theorem 3.1.
Denote $a_n = (nh_n)^{\frac{1}{1+\alpha(x_0)}}$. Using the behaviour of the densify $f(x,y)$ near the boundary $g(x_0)$, specified in (1.2), and the conditions on $g(\cdot)$ we get, for n sufficiently large,

$$\begin{aligned}
&P\left\{a_n(Z_{n,n} - g(x_0)) \le z\right\} \\
&\quad = \left[P\left\{Z_1 \le g(x_0) + z/a_n\right\}\right]^n \\
&\quad = \left[P\{|X_1 - x_0| > h\} + P\{Y_1 \le g(x_0) + z/a_n, |X_1 - x_0| \le h\}\right]^n \\
&\quad = \left[1 - P\{Y_1 > g(x_0) + z/a_n, |X_1 - x_0| \le h\}\right]^n \\
&\quad = \left[1 - \int_{x_0-h}^{x_0+h} \int_{g(x_0)+z/a_n}^{g(x)} f(x,y)\,dydx\right]^n \\
&\quad = \left[1 - \int_{x_0-h}^{x_0+h} \frac{a(x, g(x_0))}{1+\alpha(x)} \left(g(x) - g(x_0) - \frac{z}{a_n}\right)^{1+\alpha(x)} (1 + O(h))\,dx\right]^n \\
&\quad = \left[1 - \int_{x_0-h}^{x_0+h} \frac{a(x, g(x_0))}{1+\alpha(x)} \left(\frac{-z}{a_n}\right)^{1+\alpha(x)} \left(1 + \frac{g(x) - g(x_0)}{-z/a_n}\right)^{1+\alpha(x)} (1 + O(h))\,dx\right]^n
\end{aligned}$$

$$= \left[1 - \frac{a(x_0, g(x_0))}{1+\alpha(x_0)} 2h \left(\frac{-z}{a_n}\right)^{1+\alpha(x_0)} (1 + O(ha_n))\right]^n .$$

The conclusion of the theorem now follows from the conditions imposed on h_n. □

3.2 Asymptotic distribution of the endpoint estimator

The asymptotic distribution of the endpoint estimator $\hat{g}_n(x_0)$ of $g(x_0)$, defined in (2.1), is given in the next theorem.

Theorem 3.2.
Assume (1.2) and Condition A1, A3 and A4. Let $m = m(n)$ be such that $m < n/4$, $m \to \infty$ and $m/n \to 0$, as $n \to \infty$. Assume further that $h_n \to 0$,

$$\frac{nh_n^{2+\alpha(x_0)}}{m} \to 0, \qquad \sqrt{m}h_n \to 0 \tag{3.2}$$

and

$$\frac{m^{\frac{1+\alpha(x_0)+2\beta(x_0)}{2\beta(x_0)}}}{nh_n} \to 0. \tag{3.3}$$

Then,

$$\sqrt{m}\frac{\hat{g}_n(x_0) - g(x_0)}{Z_{n,n-m} - Z_{n,n-2m}} \xrightarrow{\text{D}} \text{N}(0; \sigma^2(x_0)), \tag{3.4}$$

where

$$\sigma^2(x_0) = \frac{12}{(1+\alpha(x_0))^2 \left(2^{\frac{1}{1+\alpha(x_0)}} - 1\right)^6 4^{\frac{2\alpha(x_0)}{1+\alpha(x_0)}}}.$$

and

$$\frac{Z_{n,n-m} - Z_{n,n-2m}}{\left(\frac{m}{nh_n}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}} \xrightarrow{\text{P}} -1 + 2^{\frac{1}{1+\alpha(x_0)}}. \tag{3.5}$$

Remark 4. From Theorem 3.2 we conclude that the endpoint estimator $\hat{g}(x_0)$ of $g(x_0)$ has rate of convergence $\sqrt{m}(\frac{nh_n}{m})^{\frac{1}{1+\alpha(x_0)}}$. Note that this rate of convergence does not directly depend on the value $\beta(x_0)$ but that this quantity interfers in condition (3.3).

Let's first look at the special case that $\beta(x_0)$ is infinity (meaning that the second term in the algebraic function in (1.2) is infinitely smooth). We take $h_n = cn^{-\delta}$ and $m(n) = dn^{\gamma}$, with c and d some positive constants, and search for the optimal choices of δ and γ, i.e. the choices of δ and γ which give the best rate of convergence. First rewrite the rate of convergence as $m^{\frac{1}{2}-\frac{1}{1+\alpha(x_0)}}(nh_n)^{\frac{1}{1+\alpha(x_0)}}$, and note that the exponent of m is negative, equal to 0, and positive, when $\alpha(x_0) < 1$, equal to 1, and > 1 respectively. So, one needs to study these cases separately. For example, when $\alpha(x_0) < 1$, the best rate of convergence is obtained by taking m as small as possible (i.e. γ as small as possible) and h_n as large as possible (i.e. δ as small as possible) satisfying the conditions of the theorem. A careful detailed study of each of these cases reveals that the optimal choices for h_n and m are

$$h_n = cn^{-\frac{1}{4+\alpha(x_0)}-\varepsilon} \qquad m = dn^{\frac{2}{4+\alpha(x_0)}+\varepsilon},$$

with $0 < \varepsilon < \frac{1+\alpha(x_0)}{2(4+\alpha(x_0))}$. The resulting best possible rate of convergence for the endpoint estimator is $n^{-\frac{2}{4+\alpha(x_0)}-\varepsilon\frac{\alpha(x_0)-3}{2(1+\alpha(x_0))}}$. For the special case $\alpha(x_0) = 0$ and $\beta(x_0) = \infty$ we obtain the optimal choices

$$h_n = cn^{-\frac{1}{4}-\varepsilon} \qquad m = dn^{\frac{1}{2}+\varepsilon},$$

with $0 < \varepsilon < \frac{1}{8}$, leading to the best possible rate of convergence $n^{-\frac{1}{2}+\frac{3}{2}\varepsilon}$ for the endpoint estimator. Comparing the best possible rates of convergence of the two estimators, we find that, when $\beta(x_0) = \infty$, the endpoint estimator $\hat{g}_n(x_0)$ outperforms the maximum estimator $Z_{n,n}$ if and only if $\alpha(x_0) > 0$. When $\alpha(x_0) = 0$ both estimators have rate $n^{-\frac{1}{2}+\varepsilon}$.

We now consider the case that $\alpha(x_0) = 0$ and general $\beta(x_0)$ (i.e. $\beta(x_0) > 0$). Then the rate is $nh_n m^{-1/2}$, and we find that for the optimal choices of h_n and m we should distinguish between the cases $\beta(x_0) \geq 1$ and $0 < \beta(x_0) < 1$. For $\beta(x_0) \geq 1$ the optimal h_n and m are

$$h_n = cn^{-\frac{1}{4}-\varepsilon} \qquad m = dn^{\frac{1}{2}+\varepsilon},$$

with $0 < \varepsilon < \frac{\beta(x_0)-1}{2(1+4\beta(x_0))}$, resulting in the optimal rate $n^{-\frac{1}{2}+\frac{2}{3}\varepsilon}$. In case $0 < \beta(x_0) < 1$ the optimal choices of h_n and m are

$$h_n = cn^{-\frac{1}{2(1+\beta(x_0))}-\varepsilon} \qquad m = dn^{\frac{\beta(x_0)}{1+\beta(x_0)}-\varepsilon},$$

with $0 < \varepsilon < \frac{\beta(x_0)}{1+\beta(x_0)}$, and this leads to the optimal rate $n^{-\frac{1}{2}+\frac{1}{2}\varepsilon}$. Note that the best possible rate of convergence does not depend on $\beta(x_0)$ and is always slightly slower than $n^{-\frac{1}{2}}$.

Proof of Theorem 3.2.

Define

$$G_n(v) := \frac{1}{m}\sum_{i=1}^{n} \mathbf{1}\left(Z_i > g(x_0) - (\frac{m}{nh})^{\frac{1}{1+\alpha(x_0)}} v\right) \qquad \text{for all } v > 0.$$

Then

$$\begin{aligned}
EG_n(v) &= \frac{n}{m}\int_{x_0-h}^{x_0+h}\int_{g(x_0)-(\frac{m}{nh})^{\frac{1}{1+\alpha(x_0)}}v}^{g(x)} f(x,y)\,dydx \\
&= \frac{n}{m}\int_{x_0-h}^{x_0+h} \frac{a(x,g(x_0))}{1+\alpha(x)}\left\{g(x)-g(x_0)+\left(\frac{m}{nh}\right)^{\frac{1}{1+\alpha(x_0)}} v\right\}^{1+\alpha(x)} \\
&\quad \times (1+O(|g(x)-g(x_0)|))\,dx \\
&\quad + \frac{n}{m}\int_{x_0-h}^{x_0+h} O\left(\left\{g(x)-g(x_0)+\left(\frac{m}{nh}\right)^{\frac{1}{1+\alpha(x_0)}} v\right\}^{1+\alpha(x)+\beta(x)}\right)dx \\
&= \frac{n}{m}\int_{x_0-h}^{x_0+h} \frac{a(x,g(x_0))}{1+\alpha(x)}\left(\frac{m}{nh}\right)^{\frac{1+\alpha(x)}{1+\alpha(x_0)}} v^{1+\alpha(x)}\left\{1+\frac{g(x)-g(x_0)}{(\frac{m}{nh})^{\frac{1}{1+\alpha(x_0)}}v}\right\}^{1+\alpha(x)} \\
&\quad \times (1+O(|g(x)-g(x_0)|))\,dx \\
&\quad + O\left(\frac{n}{m}h\left(\frac{m}{nh}\right)^{\frac{1+\alpha(x_0)+\beta(x_0)}{1+\alpha(x_0)}}\right) \\
&= \left\{\frac{n}{m}\int_{x_0-h}^{x_0+h} \frac{a(x,g(x_0))}{1+\alpha(x)}\left(\frac{m}{nh}\right)^{\frac{1+\alpha(x)}{1+\alpha(x_0)}} v^{1+\alpha(x)}\,dx\right\}(1+O(h)) \\
&\quad + \frac{n}{m}\int_{x_0-h}^{x_0+h} \frac{a(x,g(x_0))}{1+\alpha(x)}\left(\frac{m}{nh}\right)^{\frac{1+\alpha(x)}{1+\alpha(x_0)}} v^{1+\alpha(x)}(1+\alpha(x)) \\
&\quad \times \frac{g(x)-g(x_0)}{(\frac{m}{nh})^{\frac{1}{1+\alpha(x_0)}}v}(1+o(1))\,dx \\
&\quad + O\left(\left(\frac{m}{nh}\right)^{\frac{\beta(x_0)}{1+\alpha(x_0)}}\right) \\
&= \frac{n}{m}\frac{a(x_0,g(x_0))}{1+\alpha(x_0)}\frac{m}{nh}v^{1+\alpha(x_0)}2h(1+O(h)) \\
&\quad + \frac{n}{m}a(x_0,g(x_0))\left(\frac{m}{nh}\right)^{\frac{\alpha(x_0)}{1+\alpha(x_0)}} v^{\alpha(x_0)} \\
&\quad \times O\left(\left|\int_{x_0-h}^{x_0+h}[g(x)-g(x_0)-g'(x_0)(x-x_0)]\,dx\right|\right) \\
&\quad + O\left(\left(\frac{m}{nh}\right)^{\frac{\beta(x_0)}{1+\alpha(x_0)}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2a(x_0, g(x_0))}{1+\alpha(x_0)} v^{1+\alpha(x_0)}(1+O(h)) \\
&\quad + a(x_0, g(x_0)) v^{\alpha(x_0)} \left(\frac{n}{m}\right)^{\frac{1}{1+\alpha(x_0)}} h^{-\frac{\alpha(x_0)}{1+\alpha(x_0)}} O(h^3) \\
&\quad + O\left(\left(\frac{m}{nh}\right)^{\frac{\beta(x_0)}{1+\alpha(x_0)}}\right).
\end{aligned}
$$

Note that (3.2) implies that

$$
\sqrt{m}\left(\frac{n}{m}\right)^{\frac{1}{1+\alpha(x_0)}} h^{-\frac{\alpha(x_0)}{1+\alpha(x_0)}} O(h^3) = \sqrt{m} h \frac{h}{\left(\frac{m}{nh}\right)^{\frac{1}{1+\alpha(x_0)}}} O(1) \to 0.
$$

Thus by standard arguments (see de Haan and Resnick (1993) or Mason (1988))

$$
\sqrt{m}\left\{G_n(v) - \frac{2a(x_0, g(x_0))}{1+\alpha(x_0)} v^{1+\alpha(x_0)}\right\} \xrightarrow{\mathrm{D}} B\left(\frac{2a(x_0, g(x_0))}{1+\alpha(x_0)} v^{1+\alpha(x_0)}\right) \tag{3.6}
$$

in $D((0,\infty))$, where B is a standard Brownian motion. Furthermore

$$
\sqrt{m}\left\{G_n\left(\left(\frac{v(1+\alpha(x_0))}{2a(x_0, g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}\right) - v\right\} \xrightarrow{\mathrm{D}} B(v).
$$

By Vervaat's Lemma (Vervaat(1972))

$$
\sqrt{m}\left\{G_n^{-}\left(\left(\frac{v(1+\alpha(x_0))}{2a(x_0, g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}\right) - v\right\} \xrightarrow{\mathrm{D}} -B(v),
$$

i.e.,

$$
\sqrt{m}\left\{\frac{(g(x_0) - Z_{n,n-[mv]})^{1+\alpha(x_0)}}{\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}} - v\right\} \xrightarrow{\mathrm{D}} -B(v).
$$

So

$$
(1+\alpha(x_0)) v^{\frac{\alpha(x_0)}{1+\alpha(x_0)}} \sqrt{m}\left\{\frac{g(x_0) - Z_{n,n-[mv]}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}} - v^{\frac{1}{1+\alpha(x_0)}}\right\} \xrightarrow{\mathrm{D}} -B(v). \tag{3.7}
$$

Note that

$$
\begin{aligned}
2^{\frac{1}{1+\alpha(x_0)}} - 2^{-\hat{\gamma}_n(x_0)} &= 2^{\frac{1}{1+\alpha(x_0)}} (\log 2)\left(\frac{1}{1+\alpha(x_0)} + \hat{\gamma}_n(x_0)\right)(1+o_p(1)) \\
&= 2^{\frac{1}{1+\alpha(x_0)}}\left(\log \frac{2^{\frac{1}{1+\alpha(x_0)}}(Z_{n,n-m} - Z_{n,n-2m})}{Z_{n,n-2m} - Z_{n,n-4m}}\right)(1+o_p(1)) \\
&= 2^{\frac{1}{1+\alpha(x_0)}}\left(\frac{2^{\frac{1}{1+\alpha(x_0)}}(Z_{n,n-m} - Z_{n,n-2m})}{Z_{n,n-2m} - Z_{n,n-4m}} - 1\right)(1+o_p(1))
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2^{\frac{1}{1+\alpha(x_0)}}\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{Z_{n,n-2m}-Z_{n,n-4m}}\Bigg\{-2^{\frac{1}{1+\alpha(x_0)}}\left[\frac{g(x_0)-Z_{n,n-m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-1\right] \\
&\quad + (2^{\frac{1}{1+\alpha(x_0)}}+1)\left[\frac{g(x_0)-Z_{n,n-2m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-2^{\frac{1}{1+\alpha(x_0)}}\right] \\
&\quad - \left[\frac{g(x_0)-Z_{n,n-4m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-4^{\frac{1}{1+\alpha(x_0)}}\right]\Bigg\}(1+o_p(1)) \\
&= \frac{1}{2^{\frac{1}{1+\alpha(x_0)}}-1}\Bigg\{\frac{2^{\frac{1}{1+\alpha(x_0)}}B(1)}{(1+\alpha(x_0))\sqrt{m}}-\frac{(1+2^{\frac{1}{1+\alpha(x_0)}})B(2)}{(1+\alpha(x_0))2^{\frac{\alpha(x_0)}{1+\alpha(x_0)}}\sqrt{m}} \\
&\quad + \frac{B(4)}{(1+\alpha(x_0))4^{\frac{\alpha(x_0)}{1+\alpha(x_0)}}\sqrt{m}}\Bigg\}(1+o_p(1)).
\end{aligned}
$$

We have

$$
\begin{aligned}
&\frac{\hat{g}_n(x_0)-g(x_0)}{Z_{n,n-m}-Z_{n,n-2m}} \\
&= \frac{1}{2^{-\hat{\gamma}_n(x_0)}-1}-\frac{g(x_0)-Z_{n,n-m}}{Z_{n,n-m}-Z_{n,n-2m}} \\
&= \frac{1}{2^{-\hat{\gamma}_n(x_0)}-1}-\frac{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{Z_{n,n-m}-Z_{n,n-2m}} \\
&\quad -\frac{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{Z_{n,n-m}-Z_{n,n-2m}}\left[\frac{g(x_0)-Z_{n,n-m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-1\right] \\
&= \frac{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{(2^{-\hat{\gamma}_n(x_0)}-1)(Z_{n,n-m}-Z_{n,n-2m})}\left[\frac{Z_{n,n-m}-Z_{n,n-2m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-2^{-\hat{\gamma}_n(x_0)}+1\right] \\
&\quad -\frac{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{Z_{n,n-m}-Z_{n,n-2m}}\left[\frac{g(x_0)-Z_{n,n-m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-1\right] \\
&= \frac{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{(2^{-\hat{\gamma}_n(x_0)}-1)(Z_{n,n-m}-Z_{n,n-2m})}\Bigg\{\left[1-\frac{g(x_0)-Z_{n,n-m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}\right] \\
&\quad +\left[\frac{g(x_0)-Z_{n,n-2m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-2^{\frac{1}{1+\alpha(x_0)}}\right]+\left[2^{\frac{1}{1+\alpha(x_0)}}-2^{-\hat{\gamma}_n(x_0)}\right]\Bigg\} \\
&\quad -\frac{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}{Z_{n,n-m}-Z_{n,n-2m}}\left[\frac{g(x_0)-Z_{n,n-m}}{\left(\frac{m}{nh}\frac{1+\alpha(x_0)}{2a(x_0,g(x_0))}\right)^{\frac{1}{1+\alpha(x_0)}}}-1\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \Bigg\{ \frac{1}{(2^{\frac{1}{1+\alpha(x_0)}}-1)(2^{\frac{1}{1+\alpha(x_0)}}-1)} \left[\frac{B(1)}{(1+\alpha(x_0))\sqrt{m}} - \frac{B(2)}{(1+\alpha(x_0))2^{\frac{\alpha(x_0)}{1+\alpha(x_0)}}\sqrt{m}} \right] \\
&\quad + \frac{1}{(2^{\frac{1}{1+\alpha(x_0)}}-1)(2^{\frac{1}{1+\alpha(x_0)}}-1)^2} \left[\frac{2^{\frac{1}{1+\alpha(x_0)}}B(1)}{(1+\alpha(x_0))\sqrt{m}} \right. \\
&\quad \left. - \frac{(1+2^{\frac{1}{1+\alpha(x_0)}})B(2)}{(1+\alpha(x_0))2^{\frac{\alpha(x_0)}{1+\alpha(x_0)}}\sqrt{m}} + \frac{B(4)}{(1+\alpha(x_0))4^{\frac{\alpha(x_0)}{1+\alpha(x_0)}}\sqrt{m}} \right] \\
&\quad + \frac{1}{2^{\frac{1}{1+\alpha(x_0)}}-1} \frac{B(1)}{(1+\alpha(x_0))\sqrt{m}} \Bigg\} (1+o_p(1)) \\
&= \frac{1}{\sqrt{m}(1+\alpha(x_0))(2^{\frac{1}{1+\alpha(x_0)}}-1)} \Bigg\{ \frac{2^{\frac{2}{1+\alpha(x_0)}}}{(2^{\frac{1}{1+\alpha(x_0)}}-1)^2} B(1) \\
&\quad - \frac{2^{\frac{2}{1+\alpha(x_0)}}}{(2^{\frac{1}{1+\alpha(x_0)}}-1)^2} B(2) + (2^{\frac{1}{1+\alpha(x_0)}}-1)^{-2} 4^{-\frac{\alpha(x_0)}{1+\alpha(x_0)}} B(4) \Bigg\} (1+o_p(1)).
\end{aligned}
$$

Hence (3.4) holds with

$$
\begin{aligned}
\sigma^2(x_0) &= \frac{1}{(1+\alpha(x_0))^2(2^{\frac{1}{1+\alpha(x_0)}}-1)^2} \Bigg\{ (2^{\frac{1}{1+\alpha(x_0)}}-1)^{-4} 4^{-\frac{2\alpha(x_0)}{1+\alpha(x_0)}} E\left(B(4)-B(2)\right)^2 \\
&\qquad + \left[(2^{\frac{1}{1+\alpha(x_0)}}-1)^{-2} 4^{-\frac{\alpha(x_0)}{1+\alpha(x_0)}} - \frac{2^{\frac{2}{1+\alpha(x_0)}}}{(2^{\frac{1}{1+\alpha(x_0)}}-1)^2} \right]^2 E\left(B(2)-B(1)\right)^2 \\
&\qquad + (2^{\frac{1}{1+\alpha(x_0)}}-1)^{-4} 4^{-\frac{2\alpha(x_0)}{1+\alpha(x_0)}} EB^2(1) \Bigg\} \\
&= 12(1+\alpha(x_0))^{-2}(2^{\frac{1}{1+\alpha(x_0)}}-1)^{-6} 4^{-\frac{2\alpha(x_0)}{1+\alpha(x_0)}}.
\end{aligned}
$$

This proves the theorem. □

4 Two-stage estimation procedure

In this section we focus on the special case that $\alpha(x) = 0$ and $g(x)$ is a concave and monotone increasing function. A well known estimator in this situation is the so-called Data Envelopment Analysis (DEA) estimator which is the lowest concave monotone increasing function covering all sample points (X_i, Y_i). It has been shown by Kneip, Park and Simar (1998) that the DEA estimator has rate of convergence $n^{-2/3}$ under the assumption that the density $f(x, y)$ is bounded away from zero at the boundary, which corresponds to the case $\alpha(x) = 0$. The DEA estimator is always biased downwards, because it never exceeds the true function $g(x)$. Gijbels, Mammen, Park and Simar (1999) established the asymptotic distribution of the DEA estimator. This gave access to the asymptotic bias, and from this a bias-corrected estimator has been proposed.

The estimators proposed here are different in nature, and have a lower rate of convergence than the DEA estimator when focusing on the case $\alpha(x) = 0$. See Remarks 3 and 4. On the other hand our estimators are obtained under more general conditions; for arbitrary functions $\alpha(x)$ and without making assumptions on the boundary function $g(x)$ (apart from the mild condition of Hölder continuity).

The methodology exposed in Section 2 opens a way to a two-stage estimation procedure as follows. Suppose that the interest is in estimating $g(x_0)$, for x_0 a given point. In case $\alpha(x_0) = 0$, the endpoint and the maximum estimator have the same rate of convergence. We use the simpler maximum estimator here. In a first stage of the estimation procedure we calculate $Z_{n,n}$, the maximum of the Y_i observations in a strip of width $2h$ around x_0. We then replace each observation Y_i in that strip by this maximum, leaving all observations outside the strip unchanged. More precisely, we define

$$\widetilde{Y}_i = \begin{cases} Y_i & \text{if} \quad |X_i - x_0| > h \\ Z_{n,n} & \text{if} \quad |X_i - x_0| \leq h. \end{cases}$$

The effect of this operation is that the data points in a neighbourhood of x_0 will be more concentrated towards the boundary. We then apply the DEA estimator to these transformed data $(X_i, \widetilde{Y}_i)$, giving the estimator $\hat{g}_{n,\text{DEA}}(x_0)$. Since this estimator is based on data points which are more concentrated around the boundary it is natural to expect that it performs better than the DEA estimator of $g(x_0)$ based on the original untransformed data (X_i, Y_i), denoted by $\hat{g}_{\text{DEA}}(x_0)$. The following theorem indeed reveals that the estimator $\hat{g}_{n,\text{DEA}}(x_0)$ is always closer to the true boundary function $g(x_0)$ with probability tending to one.

Theorem 4.1

Assume (1.2) and Condition A1 and A4 with $\alpha(x) = 0$. Suppose $g(x)$ is a concave and monotone increasing function and $h = cn^{-1/2-\epsilon}$ for any $0 < \epsilon < 1/2$. Then

$$g(x_0) \geq \hat{g}_{n,\text{DEA}}(x_0) \geq \hat{g}_{\text{DEA}}(x_0)$$

with probability tending to one.

Remark 5. An immediate consequence of Theorem 4.1 is that the two-stage estimator

$\hat{g}_{n,\mathrm{DEA}}(x_0)$ has a rate of convergence which is at least as good as that of the DEA estimator $\hat{g}_{\mathrm{DEA}}(x_0)$, which is $n^{-2/3}$. To see this, note that Theorem 4.1 implies that

$$0 \leq |\hat{g}_{n,\mathrm{DEA}}(x_0) - g(x_0)| \leq |\hat{g}_{\mathrm{DEA}}(x_0) - g(x_0)|\,,$$

with probability tending to one. Hence if $|\hat{g}_{\mathrm{DEA}}(x_0) - g(x_0)|$ tends to zero at rate say b_n, so does $|\hat{g}_{n,\mathrm{DEA}}(x_0) - g(x_0)|$.

Proof of Theorem 4.1.
Recall that the DEA estimator is the lowest concave monotone increasing function covering all data points. We will show that $P\{Z_{n,n} < g(x_0 - h)\}$ tends to 1 as $n \to \infty$, which implies the statement of the theorem. Using Theorem 3.1, we get

$$\begin{aligned} P\{Z_{n,n} < g(x_0 - h)\} &= P\{Z_{n,n} - g(x_0) < g(x_0 - h) - g(x_0)\} \\ &= P\left\{nh_n\left(Z_{n,n} - g(x_0)\right) < nh_n\left(g(x_0 - h) - g(x_0)\right)\right\} \\ &= \exp\left\{-2a(x_0, g(x_0))nh_n\left(g(x_0) - g(x_0 - h)\right)\right\}\ (1 + o(1)), \end{aligned}$$

which tends to 1 since $nh_n^2 \to 0$. $\square$

5 Simulation study

We now investigate the finite-sample behaviour of the estimators introduced in Section 2 and Section 4. We consider the case of a concave and monotone increasing boundary, and the case of a more general function $g(x)$. In the former case we study the finite-sample performance of the estimator $\hat{g}_{n,\mathrm{DEA}}(x_0)$ and compare it with the finite-sample performance of the DEA estimator $\hat{g}_{\mathrm{DEA}}(x_0)$. In the case of a more general function $g(x)$ we investigate the finite-sample behaviour of the maximum estimator and the endpoint estimator in situations where the density function is bounded away from zero (i.e. $\alpha(x_0) = 0$) or tends to zero at the boundary (i.e. $\alpha(x_0) > 0$).

5.1 Concave and monotone increasing boundary

Throughout this subsection we assume (1.2) holds for $\alpha(x) = 0$ and $g(x)$ is a concave and monotone increasing function. We compare the performance of the DEA estimator $\hat{g}_{\mathrm{DEA}}(x_0)$ and the two-stage estimator $\hat{g}_{n,\mathrm{DEA}}(x_0)$ under the following two models.

Model 1:

$$\begin{aligned} &X \sim \mathrm{U}[0,1], \quad Y = g(X)\exp\{-V\}, \quad g(x) = x^{1/2}, \\ &V \sim \ \mathrm{Exp}(3), \quad \text{independent of } X. \end{aligned}$$

Model 2:

Similar to Model 1 but with $V \sim$ Exp(1).

Gijbels, Mammen, Park and Simar (1999) also used these models in their simulation study when investigating the performance of the DEA estimator and the bias-corrected DEA estimator.

For each model we consider samples of size 100, 250, 500, 750 and 1000, and estimate the function $g(x)$ at three different points $x_0 = 0.25, 0.5, 0.75$. The number of simulations in each realization is $N = 500$. In Tables 1—4 we report the estimated bias

$$\frac{1}{N}\sum_{s=1}^{N} \hat{g}^{(s)}_{n,\mathrm{DEA}}(x_0) - g(x_0),$$

and the corresponding standard error

$$\frac{1}{\sqrt{N}}\sqrt{\frac{1}{N-1}\sum_{s=1}^{N}\left(\hat{g}^{(s)}_{n,\mathrm{DEA}}(x_0) - \frac{1}{N}\sum_{s=1}^{N}\hat{g}^{(s)}_{n,\mathrm{DEA}}(x_0)\right)^2},$$

where $\hat{g}^{(s)}_{n,\mathrm{DEA}}(x_0)$ denotes the estimator for simulation number s. For comparison we also provide the estimated bias and its standard error for the DEA estimator.

The results in Tables 1—4 show that the two-stage estimator $\hat{g}_{n,\mathrm{DEA}}(x_0)$ performs very well, and in fact often outperforms the DEA estimator. Note also that the standard errors for the two-stage estimator are in almost all situations less than these for the DEA estimator.

Table 1: *Simulation results for Model 1. Estimated Bias ($\times 10^{-2}$) with the corresponding standard error ($\times 10^{-4}$) between brackets. With $h = n^{-3/4}$.*

	$x_0 = 0.25$	$x_0 = 0.25$	$x_0 = 0.5$	$x_0 = 0.5$	$x_0 = 0.75$	$x_0 = 0.75$
n	DEA	TWO-STAGE	DEA	TWO-STAGE	DEA	TWO-STAGE
100	-1.4168(3.5139)	0.1861(3.6952)	-1.2964(3.2026)	-0.5142(2.7003)	-1.3420(3.6796)	-0.8798(2.1649)
250	-0.7863(1.9576)	-0.1631(1.6841)	-0.6935(1.7671)	-0.3543(1.2125)	-0.6723(1.6704)	-0.5092(0.8153)
500	-0.4849(1.1811)	-0.1551(0.8715)	-0.4319(1.1054)	-0.2701(0.6138)	-0.4058(0.9947)	-0.3083(0.4661)
750	-0.3671(0.9007)	-0.1538(0.5969)	-0.3254(0.8155)	-0.2344(0.4025)	-0.3140(0.7850)	-0.2500(0.3328)
1000	-0.3075(0.7605)	-0.1485(0.4614)	-0.2762(0.7245)	-0.1998(0.3380)	-0.2580(0.6283)	-0.2163(0.2361)

Table 2: *Simulation results for Model 1. Estimated Bias ($\times 10^{-2}$) with the corresponding standard error ($\times 10^{-4}$) between brackets. With $h = n^{-2/3}$.*

	$x_0 = 0.25$	$x_0 = 0.25$	$x_0 = 0.5$	$x_0 = 0.5$	$x_0 = 0.75$	$x_0 = 0.75$
n	DEA	TWO-STAGE	DEA	TWO-STAGE	DEA	TWO-STAGE
100	-1.4168(3.5139)	1.9835(5.0180)	-1.2964(3.2026)	0.6618(3.7012)	-1.3420(3.6796)	-0.3404(2.8451)
250	-0.7863(1.9576)	0.9669(2.5966)	-0.6935(1.7671)	0.2213(1.9129)	-0.6723(1.6704)	-0.1173(1.3577)
500	-0.4849(1.1811)	0.5594(1.5518)	-0.4319(1.1054)	0.1167(1.1619)	-0.4058(0.9947)	-0.0877(0.7629)
750	-0.3671(0.9007)	0.3960(1.1056)	-0.3254(0.8155)	0.0606(0.8171)	-0.3140(0.7850)	-0.0659(0.6069)
1000	-0.3075(0.7605)	0.2826(0.8901)	-0.2762(0.7245)	0.0144(0.6249)	-0.2580(0.6283)	-0.0687(0.4868)

Table 3: *Simulation results for Model 2. Estimated Bias ($\times 10^{-2}$) with the corresponding standard error ($\times 10^{-4}$) between brackets. With $h = n^{-3/4}$.*

	$x_0 = 0.25$	$x_0 = 0.25$	$x_0 = 0.5$	$x_0 = 0.5$	$x_0 = 0.75$	$x_0 = 0.75$
n	DEA	TWO-STAGE	DEA	TWO-STAGE	DEA	TWO-STAGE
100	-2.8714(6.6605)	-1.8391(4.7621)	-2.6644(6.5500)	-2.1654(3.8374)	-3.2638(9.6096)	-3.0371(3.8374)
250	-1.6160(4.0403)	-1.2171(2.1846)	-1.4395(3.6841)	-1.2200(1.8012)	-1.5142(4.1976)	-1.4245(1.7657)
500	-1.0034(2.4330)	-0.8314(1.1325)	-0.8924(2.2660)	-0.7978(0.9500)	-0.8652(2.1773)	-0.8137(0.8515)
750	-0.7490(1.8408)	-0.6354(0.8154)	-0.6808(1.7298)	-0.6246(0.6756)	-0.6518(1.6264)	-0.6160(0.5962)
1000	-0.6303(1.5163)	-0.5441(0.6317)	-0.5691(1.4388)	-0.5201(0.5396)	-0.5364(1.3237)	-0.5150(0.4435)

Table 4: *Simulation results for Model 2. Estimated Bias ($\times 10^{-2}$) with the corresponding standard error ($\times 10^{-4}$) between brackets. With $h = n^{-2/3}$.*

	$x_0 = 0.25$	$x_0 = 0.25$	$x_0 = 0.5$	$x_0 = 0.5$	$x_0 = 0.75$	$x_0 = 0.75$
n	DEA	TWO-STAGE	DEA	TWO-STAGE	DEA	TWO-STAGE
100	-2.8714(6.6605)	-0.4642(6.4457)	-2.6644(6.5500)	-1.4824(4.7695)	-3.2638(9.6096)	-2.7608(5.2957)
250	-1.6160(4.0403)	-0.5074(3.2339)	-1.4395(3.6841)	-0.8583(2.4063)	-1.5142(4.1976)	-1.2321(2.0631)
500	-1.0034(2.4330)	-0.3771(1.8062)	-0.8924(2.2660)	-0.5506(1.3382)	-0.8652(2.1773)	-0.7096(1.0280)
750	-0.7490(1.8408)	-0.3168(1.2791)	-0.6808(1.7298)	-0.4576(0.9662)	-0.6518(1.6264)	-0.5203(0.7793)
1000	-0.6303(1.5163)	-0.3119(1.0125)	-0.5691(1.4388)	-0.4043(0.7729)	-0.5364(1.3237)	-0.4341(0.6076)

5.2 General Boundary

For a general support curve g we investigate the finite-sample behaviour of the maximum estimator $Z_{n,n}$ and the endpoint estimator $\hat{g}_n(x_0)$ introduced in Section 2.

We consider the following two models.

Model 3:

$$\begin{array}{l} X \sim \mathrm{U}[0,1], \quad Y = g(X)V, \quad g(x) = 2 - 4x + 4x^2, \\ V \sim \ \mathrm{U}[0,1] \quad \text{independent of } X. \end{array}$$

Note that $f(x, g(x)) > 0$ under this model.

Model 4:

Similar to Model 3 but with $V \sim 1 - (1-v)^5$. Note that $f(x,y) = O((g(x)-y)^4)$ as $y \uparrow g(x)$.

Under Model 3 the density function $f(x,y)$ is bounded away from 0 at the boundary, while under Model 4 the density tends to 0 at the boundary at a polynomial rate of

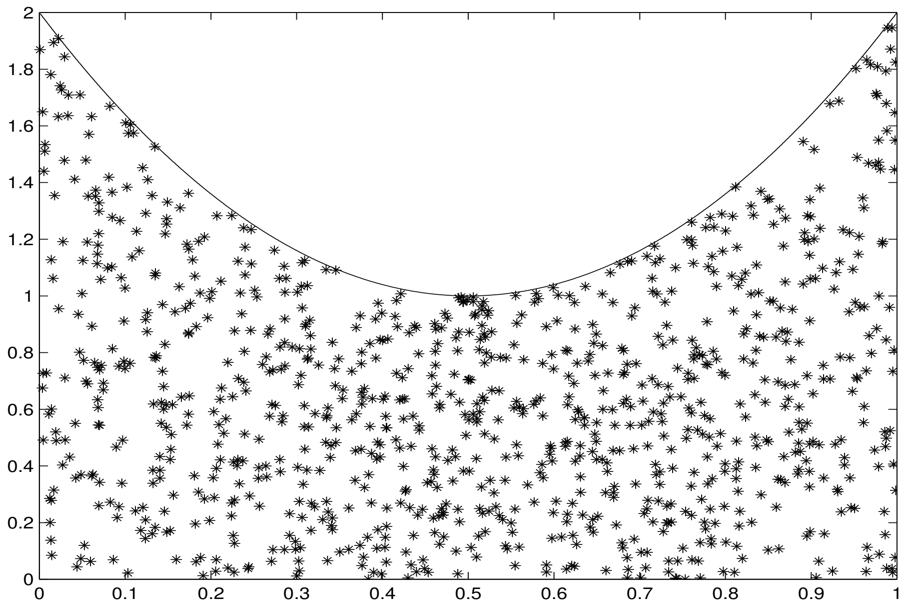

Figure 1: *A typical simulated data set of sample size 1000 from Model 3. The solid line represents the true boundary function g.*

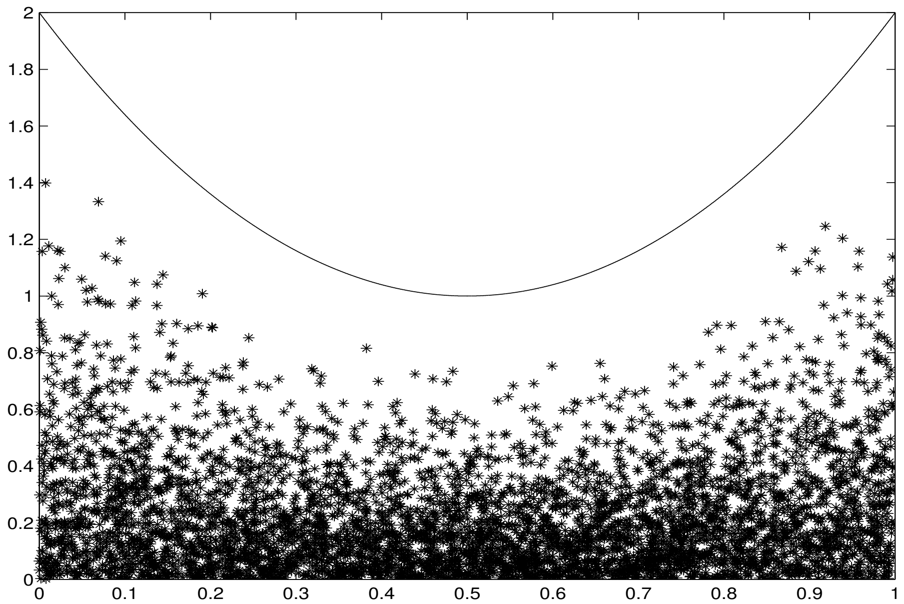

Figure 2: *A typical simulated data set of sample size 5000 from Model 4. The solid line represents the true boundary function g.*

order 4. So, in Model 3, $\alpha(x) = 0$ while in Model 4, $\alpha(x) = 4$. Since the rate of convergence of both estimators, the maximum estimator $Z_{n,n}$ and the endpoint estimator $\hat{g}_n(x_0)$, decreases with $\alpha(x)$ as can be seen from the results in Section 3, the estimation task is far more difficult under Model 4 than under Model 3. Therefore we use a larger

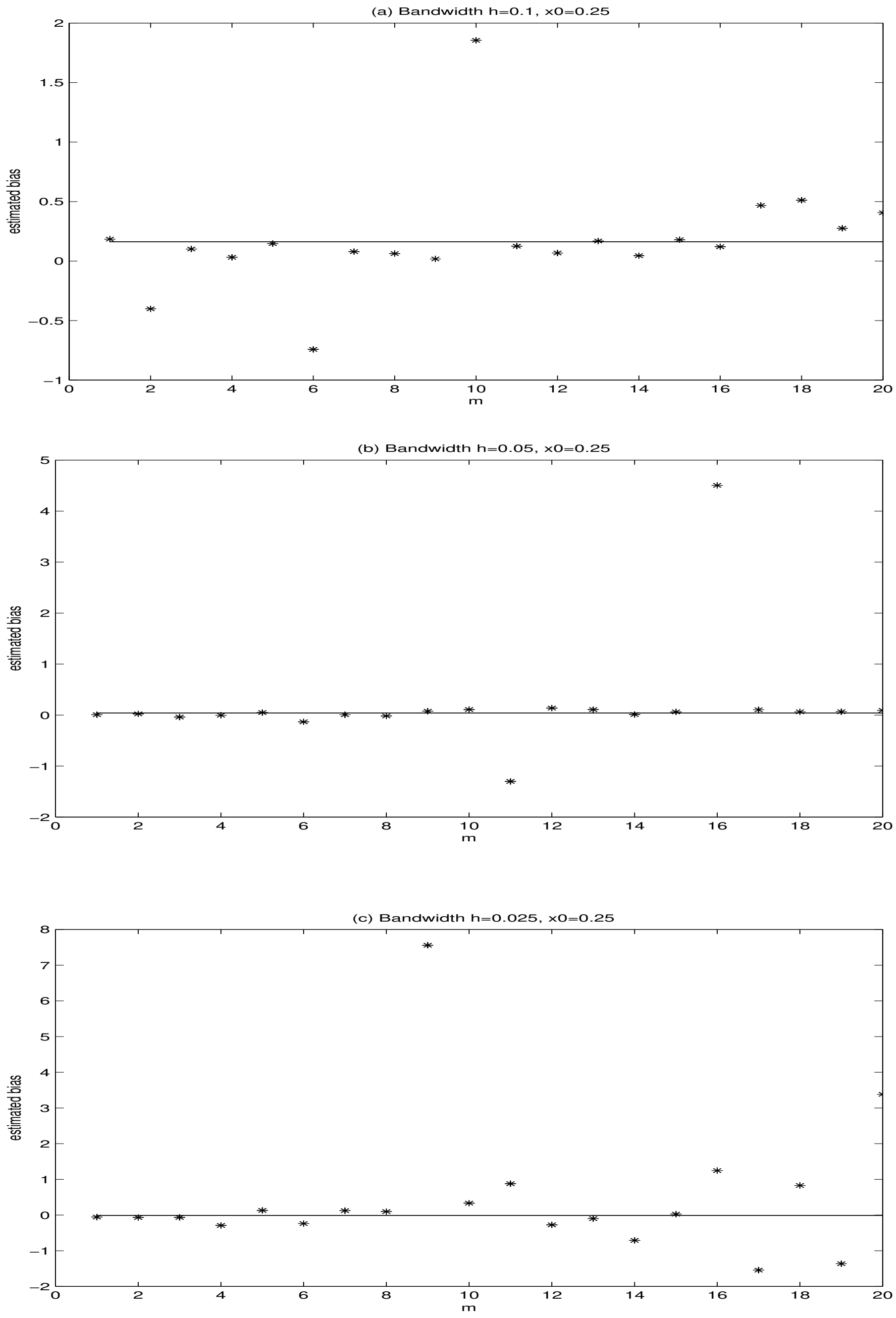

Figure 3: *The character* $'*'$ *indicates the estimated bias of the endpoint estimator* $\hat{g}_n(0.25)$ *and the solid line represents the estimated bias of the maximum estimator* $Z_{n,n}$ *under Model 3. The estimators are calculated with (a):* $h = 0.1$*, (b):* $h = 0.05$ *and (c):* $h = 0.025$*. Note that in (a), (b), (c),* $Z_{n,n} - g(0.25) = 0.1622, 0.0408, -0.0121$ *respectively.*

sample size for Model 4. We simulate samples of size 1000 from Model 3, and samples of size 5000 from Model 4. Typical simulated datasets under Models 3 and 4 are presented in Figures 1 and 2 respectively.

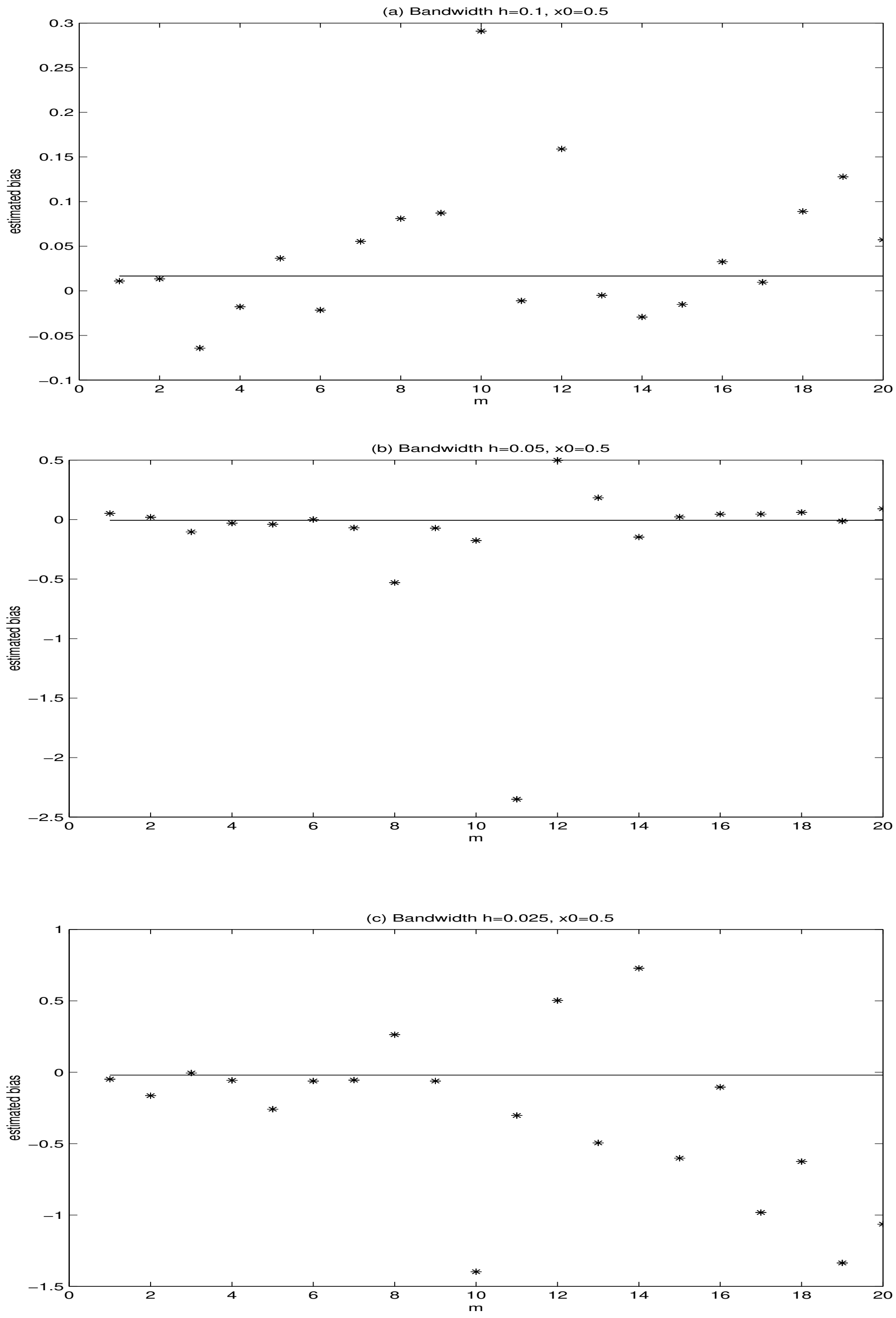

Figure 4: *The character* $'{*}'$ *indicates the estimated bias of the endpoint estimator* $\hat{g}_n(0.50)$ *and the solid line represents the estimated bias of the maximum estimator* $Z_{n,n}$ *under Model 3. The estimators are calculated with (a):* $h = 0.1$*, (b):* $h = 0.05$ *and (c):* $h = 0.025$*. Note that in (a), (b), (c),* $Z_{n,n} - g(0.50) = 0.0165, -0.0063, -0.0193$ *respectively.*

The boundary $g(x_0)$ is estimated at three different points $x_0 = 0.25, 0.5$ and 0.75, using $Z_{n,n}$ and $\hat{g}_n(x_0)$. The number of simulations in each situation is $N = 500$. For each of the chosen smoothing parameters $h = 0.1, 0.05$ or 0.025 we plot the estimated bias against

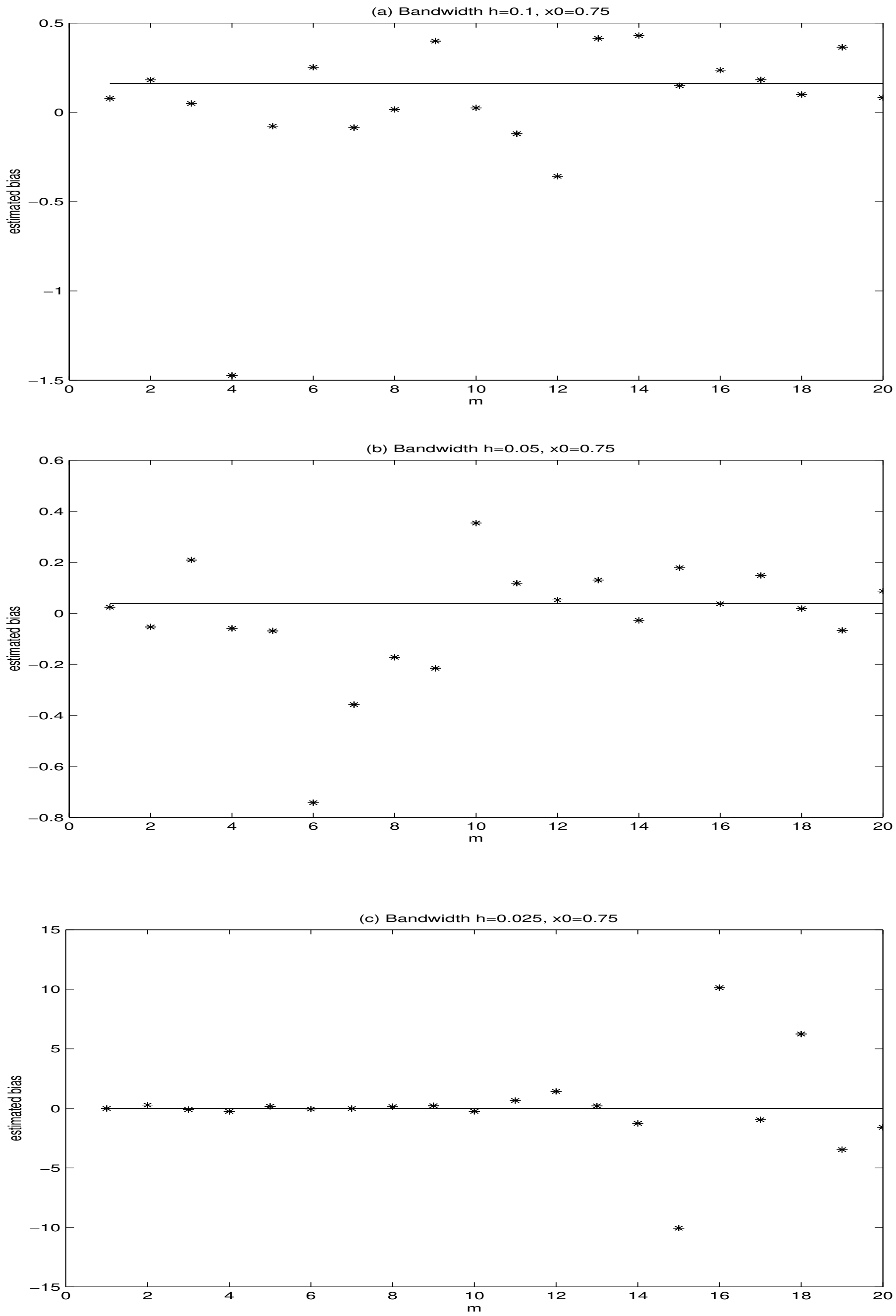

Figure 5: *The character* $'*'$ *indicates the estimated bias of the endpoint estimator* $\hat{g}_n(0.75)$ *and the solid line represents the estimated bias of the maximum estimator* $Z_{n,n}$ *under Model 3. The estimators are calculated with (a):* $h = 0.1$*, (b):* $h = 0.05$ *and (c):* $h = 0.025$*. Note that in (a), (b), (c),* $Z_{n,n} - g(0.75) = 0.1609, 0.0391, -0.0129$ *respectively.*

m for a range of values of m. Given the parameter h, we may modify the approach in Hall and Weissman (1997) or adopt the approach in Ferreira, de Haan and Peng (1999) to choose an optimal m. This will be part of our future work.

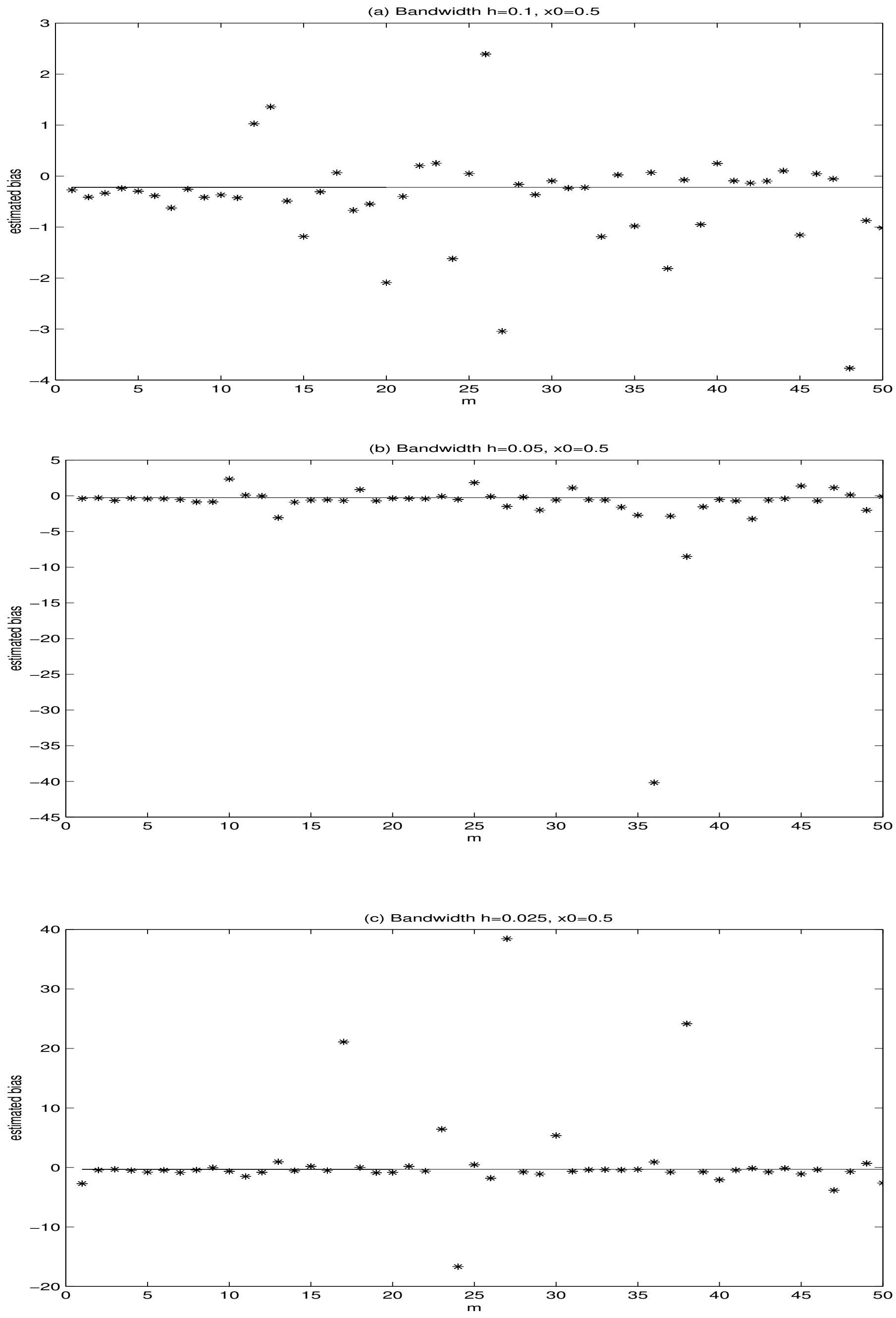

Figure 6: *The character* $'*'$ *indicates the estimated bias of the endpoint estimator* $\hat{g}_n(0.50)$ *and the solid line represents the estimated bias of the maximum estimator* $Z_{n,n}$ *under Model 4. The estimators are calculated with (a):* $h = 0.1$*, (b):* $h = 0.05$ *and (c):* $h = 0.025$*. Note that in (a), (b), (c),* $Z_{n,n} - g(0.50) = -0.2206, -0.2648, -0.3125$ *respectively.*

Figure 3 shows the estimators for $g(0.25)$ under Model 3. In Figures 4 and 5 the estimators for respectively $g(0.50)$ and $g(0.75)$ are presented. For Model 4 we only show the estimators for $g(0.50)$ in Figure 6.

Recall that according to the theoretical results the endpoint estimator has a better rate of convergence than the maximum estimator when $\alpha(x_0) > 0$, as is the case under Model 4. This is not really reflected in the simulation results. This is may be not surprising given the fact that these results are asymptotic results, and that the optimal rates of convergence also assume that the bandwidth for each estimator has been chosen in an optimal way. Note that in our simulations we took the same bandwidth for both estimators.

Acknowledgement

This research has been carried out while the first author was visiting the Centre for Mathematics and its Applications, Australian National University, Canberra. The authors gratefully acknowledge the hospitality and the financial support of the Centre. Gijbels' research was also supported by the 'Projet d'Actions de Recherche Concertées' (No. 98/03–217) .

References

Banker, R.D. (1993). Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. *Management Science*, **10**, 1265–1273.

Charnes, A., Cooper, W.W., Lewin, A.Y. and Seiford, L.M. (Eds) (1995). *Data Envelopment: Theory, Methodology and Applications*. Kluwer Academic, Boston.

Dekkers, A.L.M., Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, **17**, 1833–1855.

Dekkers, A.L.M. and de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.*, **17**, 1795–1832.

de Haan, L. and Resnick, S. (1993). Estimating the limit distribution of multivariate extremes. *Stochastic Models*, **9**, 275–309.

Ferreira, A., de Haan, L. and Peng, L. (1999). Adaptive estimators for the endpoint and high quantiles of a probability distribution. *Manuscript*.

Gijbels, I., Mammen, E., Park, B.U. and Simar, L. (1999). On estimation of monotone and concave frontier functions. *Jour. Amer. Statist. Assoc.*, **94**, to appear.

Grosskopf, S. (1996). Statistical inference and nonparametric efficiency: a selective survey. *Jour. Productivity Anal.*, **7**, 161–176.

Hall, P. (1982). On estimating the endpoint of a distribution. *Ann. Statist.*, **10**, 556–568.

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

Hall, P., Park, B.U. and Stern, S.E. (1998). On polynomial estimators of frontiers and boundaries. *Jour. Multiv. Anal.*, **66**, 71–98.

Hall, P. and Weissman, I. (1997). On the estimation of extreme probabilities. *Ann. Statist.*, **25**, 1311–1326.

Härdle, W., Park, B.U. and Tsybakov, A.B. (1995). Estimation of non sharp support boundaries. *Jour. Multiv. Anal.*, **55**, 205–218.

Kneip, A., Park., B.U. and Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, **14**, 783–793.

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

Korostelev, A.P., Simar, L. and Tsybakov, A.B. (1995b). On estimation of monotone and concave boundaries. *Publications des Instituts de Statistique des Universités de Paris*, XXXIX, 1, 3–18.

Korostelev, A.P. and Tsybakov, A.B. (1993). *Minimax Theory of Image reconstruction.* Lecture Notes in Statistics, **82**. Springer-Verlag, New York/Berlin.

Mammen, E. and Tsybakov, A.B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.*, **23**, 502–524.

Mason, D. (1988). A strong invariance theorem for the tail empirical process. *Ann. Inst. Henri Poincare*, **24**, 491–506.

Pickands III, J. (1975). Statistical inference using extreme-order statistics. *Ann. Statist.*, **3**, 119–131.

Seiford, L.M. (1996). Data envelopment analysis: the evolution of the state of the art (1978–1995). *Jour. Productivity Anal.*, **7**, 99–137.

Simar, L. (1996). Aspects of statistical analysis in DEA-type frontier models. *Jour. Productivity Anal.*, **7**, 177–185.

Vervaat, W. (1972). Functional central limit theorems for processes with positive drift and their inverses. *Z. Wahrsch. verw. Gebiete*, **23**, 249–253.