The TZ Robust Nonparametric Frontier Estimator

Hudson Torrent[1] and Flavio A. Ziegelmann[2]

Abstract. In this paper we propose a new general fully nonparametric estimator of deterministic frontier quantiles, which is based on a two-stage approach. The new estimator has many advantages over traditional frontier estimators as DEA and FDH, as well as over others more recently proposed in the literature, such as Martins-Filho and Yao (2007) and Martins-Filho, Torrent and Ziegelmann (2013), specially regarding its robustness to outliers. Our approach may be viewed as a simplification and, at the same time, a generalisation of that proposed by Martins-Filho and Yao (2007), which estimates a frontier model in three stages. We additionally perform simulation studies comparing its performance with other non-robust-to-outliers methods, strongly suggesting that our robust version be adopted. Asymptotic properties are discussed, showing consistency and $\sqrt{nh_n}$ asymptotic normality under standard assumptions.

Keywords and phrases. frontier, quantile, nonparametric; local linear regression.

JEL Classifications. C14, C22

Area: Econometria

[1]Department of Statistics and PPGE, Federal University of Rio Grande do Sul, Porto Alegre - RS. Brazil. email: hudson.torrent@ufrgs.br.

[2]Department of Statistics and PPGE and PPGA, Federal University of Rio Grande do Sul, Porto Alegre - RS. Brazil, email: flavioz@ufrgs.br. The author wishes to thank CNPq (project 305290/2012-6) for financial support.

1 Introduction

Estimation of production frontiers and therefore efficiency (and inefficiency) of production processes has been the subject of a vast and growing literature since Farrell (1957). The problem can be stated as follows. Let $x \in \mathbb{R}^p_+$ be a set of inputs used to produce a set of outputs $y \in \mathbb{R}^q_+$. So, there is a technological or production set defined as $\Psi = \{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \text{ can produce } y\}$. A production frontier associated with Ψ is defined as $\rho(x) = \sup\{y \in \mathbb{R}^q_+ \mid (y, x) \in \Psi\}$ for all $x \in \mathbb{R}^p_+$. Thus, for given $(x_0, y_0) \in \Psi$, efficiency is measured by the distance between y_0 and $\rho(x_0)$. In other terms, $y_0 = \rho(x_0)R$, where $R \in (0, 1]$ is the measure of efficiency or simply efficiency. Since outliers can have a strong influence on frontier estimation, growing interest has emerged in the literature regarding the estimation of frontier quantiles. Hence, in this paper, we are interested in estimating, from a given random sample $\chi = \{(x_i, y_i), i = 1, ..., n\}$, an associated production frontier, that is, $\rho(\cdot)$.

Although more appealing from an econometric perspective, separating inefficiency and random shock in stochastic frontier models requires strong parametric assumptions on the joint density of (X_i, Y_i) (Aigner et al. (1977), Fan et al. (1996), Kumbhakar et al. (2007), Martins-Filho and Yao (2014)). In contrast, deterministic frontier models can be estimated under much milder restrictions on the generating stochastic process. We therefore consider here only the deterministic approach, which relies on the assumption that all sample observations lie in the technological set. That is, one does not consider the problem where there is noise on the data, although our proposed method is somehow armoured against outliers.

Estimation and inference for deterministic frontier models have been largely conducted using DEA (data envelopment analysis) and FDH (free disposal hull) estimators (Charnes et al. (1978), Deprins et al. (1984)). The idea is to estimate a production set from an observed random sample without being necessary to assume any restrictive parametric structure either on the production frontier $\rho(\cdot)$ or on the joint density of (X_i, Y_i). Many works apply these methodologies. Although the asymptotic properties of DEA (Gijbels et al. (1999) and Kneip et al. (2008)) and FDH (Park et al. (2000)) are now well known, these estimators have a few drawbacks: they are not robust to extreme values, are inherently biased downward and generate estimated frontiers that are both non-smooth and discontinuous. To remedy these problems, a number of alternative nonparametric frontier specifications and estimation procedures have been proposed (Aragon et al. (2005), Martins-Filho and Yao (2007, 2008), Daouia and Simar (2007), Daouia et al. (2009, 2010, 2012), Martins-Filho, Torrent and Ziegelmann (2013)).

Martins-Filho and Yao (2007) propose a deterministic production frontier model associated to a nonparametric estimator named NP3S[1]. They derive the asymptotic normality and consistency of both production frontier and efficiency estimators under reasonable assumptions in the nonparametric context. This estimator shares the flexible nonparametric structure of DEA and FDH. Despite not being robust to ouliers (its scale can be affected), is has some extra desirable properties if compared to the two just mentioned above: *i*) NP3S estimator is more robust to extreme values in terms of the frontier shape estimation; *ii*) the frontier estimator is a smooth function of input usage (not discontinuous neither piecewise linear) and *iii*) although the estimator envelops the data, it is not inherently biased as FDH and DEA estimators. The estimation method is fairly simple since it is based on local linear Kernel estimation via a three-step procedure. First step is the estimation of a conditional mean via local linear regression. The second step follows Fan and Yao (1998), i.e., a local linear regression is again used to estimate the conditional variance function, capturing the shape of the frontier. The third and final step, which is based on the assumption that there is at least one efficient firm, estimates the frontier scale, positioning the frontier on the plane.

Martins-Filho, Torrent and Ziegelmann (2013) noticed that an undesirable result might be emerging in the second step of NP3S estimator, since the method allows for a negative estimate of the variance. To overcome this problem, they propose to use the local exponential estimator of Ziegelmann (2002) in the second step, ensuring the nonnegativity of the variance estimate. They derive the asymptotic normality and consistency of production frontier under standard assumptions in the nonparametric context. This estimator consists in using an exponential functional at the minimization problem that characterizes Kernel regressions. We write this frontier estimator as NPE.

Although NP3S and NPE might have advantages in comparison with FDH and DEA estimators, the effects of outliers can be devasting if the outlier is "ellected" to be the efficient firm. Therefore, some improvements are desirable. The former estimators are characterized by an estimation procedure in three steps. The first two steps give the shape of the frontier and the third step is responsible to locate the estimated frontier. Nevertheless, we can eliminate the second step of NP3S and NPE estimators, estimating the frontier in only two steps. Furthermore, our estimator, the NP2S, has as first step exactly the same first step performed by NP3S and NPE estimators, whereas our second step is very similar to their third step, but with a quantile flavour. Therefore, we can in fact eliminate the second step of NP3S and NPE and

[1]In this paper, we name the estimator proposed by Martins-Filho and Yao as NP3S, contrasting with our proposed estimator, for which we write NP2S.

get the frontier in a simpler and more general estimation procedure. Our contribution goes further and is twofold: i) obtaining frontier shape from only one step, and, more importantly, ii) having a fully robust (to outliers) frontier estimator. Besides the simpler and more efficient fashion of our estimator, it maintains the advantages over FDH and DEA listed above.

The rest of this paper is composed as follows. In the second section we present the model originally proposed by Martins-Filho and Yao (2007) - which is slightly rewritten - as well as our new estimation procedure, and compare it with NP3S, NPE, DEA and FDH. Section 3 discusses the asymptotic properties of our estimator. In Section 4 a Monte Carlo study is presented. Finally, in Section 5 conclusions and final comments are stated.

2 Nonparametric Frontier Estimation via Local Kernel Regression

2.1 The Model

In this section we present the model proposed by Martins-Filho and Yao (2007). The problem may be viewed considering a firm that makes only one product from k inputs, that is, $(x, y) \in \mathbb{R}^p_+ \times \mathbb{R}_+$, where x describes p inputs used for production and y describes the output (one-output case) of a production unit. The production set is defined as previously. We then have the following production set:

$$\Psi = \{(x, y) \in \mathbb{R}^{p+1}_+ | x \text{ can produce } y\} \ .$$

The production function or frontier associated with Ψ is

$$\rho(x) = \sup\{y \in \mathbb{R}^q_+ \mid (y, x) \in \Psi\} \text{ for all } x \in \mathbb{R}^p_+.$$

In practice Ψ and its frontier are unknown, so our prior interest is in estimating this frontier from a set of observed firms, i.e., given a random sample of production units $\{(X_i, Y_i)\}_{i=1}^n$ that share a technology Ψ, obtaining estimates of $\rho(\cdot)$. By extension we are interested in constructing efficiency ranks and relative performance of production units. To see this, let $(x_0, y_0) \in \Psi$ characterize the performance of a production unit and define $0 \leq R_0 \equiv \frac{y_0}{\rho(x_0)} \leq 1$ to be this units (inverse) Farrell output efficiency measure.[2] From estimates of ρ we can obtain estimates of R_0.

This frontier regression model consists of a multiplicative regression. We assume that $Z_i \equiv (X_i, R_i)'$ is a $(p+1)$ dimensional random vector with common density g for all $i \in \{1, 2, ...\}$ and $\{Z_i\}$ forms an

[2]Note that if the production level y_0 associated with x_0 lies on the frontier function we have $y_0 = \rho(x_0)$. The production process is efficient and $R_0 = 1$.

independently distributed sequence.

If there are observations on a random variable Y_i, the suitable regression function is defined as

$$Y_i = \rho(X_i)R_i = \frac{m(X_i)}{\mu_R}R_i \tag{1}$$

where R_i is an unobserved random variable, X_i is an observed random vector in $\mathbb{R}^p_+$. In this context Y_i is the output and $\rho(\cdot) = \frac{m(X_i)}{\mu_R}$ is the production frontier, where $m(\cdot) : \mathbb{R}^p_+ \to (0, \infty)$ is a measurable function and μ_R is an unknown parameter. X_i are the inputs and R_i is the efficiency with values in $[0, 1]$. The closer R_i is to 1 the closer are observed output and frontier. For an observed (x_i, y_i), if we have y_i is far from $\rho(x_i)$ it means low efficiency and so a small value for R_i. There is no specification about the R_i density, however two moment restrictions on R_i must be assumed:

$$E(R_i|X_i = x) \equiv \mu_R \tag{2}$$

$$V(R_i|X_i = x) \equiv \sigma^2_R \ , \tag{3}$$

where $R_i \in [0, 1]$ and $0 < \mu_R < 1$ imply by construction that $0 < \sigma^2_R < \mu_R < 1$. The unknown parameter μ_R locates the production frontier. For example, if a random sample of a population is far from the true frontier, efficiency is low hence μ_R and σ_R are small. In this case, DEA or FDH estimators will produce a sub-estimated production frontier. Due to presence of σ_R in NP3S model, the estimated frontier is shifted to a higher level when compared to DEA or FDH. In next subsection, we present the estimation procedure for this model and propose to modify the estimation, eliminating the second step.

2.2 Proposed Estimator

In this section we characterize our estimator, the NP2S. We can rewrite equation (1) as follows:

$$Y_i = \frac{m(X_i)}{\mu_R}R_i = m(X_i) + \frac{\sigma_R}{\mu_R}m(X_i)\frac{(R_i - \mu_R)}{\sigma_R}$$

Hence,

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i \tag{4}$$

where $\epsilon_i = \frac{(R_i - \mu_R)}{\sigma_R}$ and $\sigma(X_i) = \frac{\sigma_R}{\mu_R}m(X_i)$.

Given the conditional moment restrictions (2) and (3) on R_i we have that $E(\varepsilon_i|X_i = x) = 0$ and $V(\varepsilon_i|X_i = x) = 1$. Hence, $E(Y_i|X_i = x) = m(X_i)$ and $V(Y_i|X_i = x) = \sigma^2(x)$. First, we note that

$m(X_i) \equiv \mu_R \rho(X_i)$. Therefore, estimating $m(X_i)$ gives to us $\hat{m}(x) = \mu_R \hat{\rho}(x)$, since μ_R does not depend on X_i. We thus get from $\hat{m}(x)$ an estimation of $\rho(x)$, but in a wrong position. Then, if we have an estimator for μ_R we can propose to estimate the frontier as $\hat{\rho}(X_i) = \frac{\hat{m}(X_i)}{\hat{\mu}_R}$. With this in mind, we propose to estimate $\rho(X_i)$ in two simple steps. The first is simply the local linear Kernel estimator of Fan (1992) with regressand Y_i and regressors X_i. That is, for any $x \in \mathbb{R}^p_+$ we obtain $\hat{m}(x) \equiv \hat{\alpha}$ from the problem below:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \beta(X_i - x))^2 K_{h_n}(X_i - x) \tag{5}$$

where $K(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is a symmetric density function, $K_h(u) = (1/h)K(u/h)$ and $0 < h_n \to 0$ as $n \to \infty$ is a bandwidth. This first step gives us the frontier, but multiplied by μ_R. Then, in the second step, we propose an estimator for μ_R based on a (α)-quantile, that is

$$\hat{\mu}_R = \left[\hat{q}_\alpha \left(\frac{Y_i}{\hat{m}(X_i; h_n)} \right) \right]^{-1}, \tag{6}$$

where $\hat{q}_\alpha(x)$ is the sample (α)-quantile of x. This assumes that there are at least $100 \times (1-\alpha)\%$ of the firms which are efficient, i.e., with R_i identically one.

To understand the idea behind the estimator proposed above one should note that $Y_i = \rho(X_i)R_i = \frac{m(X_i)}{\mu_R} R_i$ and then set the $\alpha-$quantile firm to be efficient. Therefore, after these two steps the proposed estimator for the frontier at $x \in \mathbb{R}^p$ is given by $\hat{\rho}(X_i) = \frac{\hat{m}(X_i, h_n)}{\hat{\mu}_R}$.

2.2.1 Comparing NP2S Estimation Procedure with NP3S and NPE

One of the goals of this paper is to propose an alternative estimation procedure for the model in equation (4). Thus, in this subsection we outline the estimation procedures proposed by Martins-Filho and Yao (2007) (NP3S estimator) and Martins-Filho et. al (2013) (NPE estimator) and compare some features of those estimators with the estimator proposed in this paper (NP2S estimator). NP3S and NPE estimation methods for the model described in equation (4) are composed by three steps. The first two steps are responsible to estimate frontier shape $(\sigma(\cdot))$ while third step gives an estimative of frontiers position (σ_R).

The first step is the same as the first step for NP2S as in equation (5). The second step consists of implementing again a local kernel regression, but now for the conditional volatility function. The idea is using $\hat{m}(\cdot)$ from first step and define $e_i \equiv (Y_i - \hat{m}(X_i))^2$ to obtain $\hat{\sigma}^2(x)$ as follows:

$$(\hat{\alpha}_1, \hat{\beta}_1) = \arg\min_{\alpha_1,\beta_1} \sum_{i=1}^{n} (e_i - \psi(\alpha_1 - \beta_1(X_i - x)))^2 K_{h_n}(X_i - x) \tag{7}$$

For NP3S estimator, we have $\psi(x) \equiv x$ and the estimator for conditional volatility function is given by $\hat{\sigma}_l^2(x) = \hat{\alpha}$. This is the local linear kernel estimator for the variance as defined by Fan and Yao (1998). For NPE estimator, the functional has the form $\psi(x) \equiv \exp(x)$ and the variance estimator is defined as $\hat{\sigma}_e^2(x) = \exp(\hat{\alpha})$, as proposed in Ziegelmann (2002). After that, frontier shape is estimate in NP3S model as $\hat{\sigma}_l(X_i) = \sqrt{\hat{\sigma}_l^2(X_i)}$ and in NPE model as $\hat{\sigma}_e(X_i) = \sqrt{\hat{\sigma}_e^2(X_i)}$.

After obtaining an estimative for frontier shape ($\hat{\sigma}^2(\cdot)$) a third step is proposed to estimate frontier position (σ_R). The proposed estimator is

$$s_R = \left(\max_{1 \leq i \leq n} \frac{Y_i}{\hat{\sigma}(X_i)} \right)^{-1} \tag{8}$$

where $\hat{\sigma}(X_i)$ is the estimative from the respective second step, as described in the previous paragraph. We use s_R^l to represent the location estimator for NP3S, and for NPE we use s_R^e. As pointed out earlier, the intuition behind this estimator is to assume that there exists one observed production unit that is efficient, i.e., there is some R_i identically one. Hence, a production frontier estimator at $x \in \mathbb{R}^p$ is given by $\hat{\rho}_l(\cdot) = \frac{\hat{\sigma}_l(\cdot)}{s_R^l}$ for NP3S case and $\hat{\rho}_e(\cdot) = \frac{\hat{\sigma}_e(\cdot)}{s_R^e}$.

Comparing NP2S, NPE and NP3S, the first step is exactly the same in all cases. Furthermore, the step responsible to locate the frontier - second step in NP2S and third step in NP3s and NPE - is built over the same idea. Note however, that NP2S eliminates one step, and therefore, does not require estimation of a conditional volatility function; and thus NP2S eliminates the need of estimating a regression that has as depended variable residuals of a previous regression. In other words, our secondary contribution, besides robustness to outliers, is to get the frontier shape from only one step, since $m(X_i) = \mu_R \rho(X_i)$. Then we need to correct the frontier position using an estimative for μ_R. For NP3S and NPE cases, frontier shapes are captured after two steps, then a third step is necessary to correct frontier position.

3 Asymptotic Characterization

In this section we discuss the asymptotic properties of the estimator proposed. The following assumptions are assumed:

Assumption A1. 1. $Z_i = (X_i, R_i)'$ for $i = 1, 2, ..., n$ is an independent and identically distributed sequence of random vectors with density f. $f_X(x)$ and $f_R(r)$ denote the common marginal densities of X_i and R_i respectively, and $f_{R|X}(r; X)$ denotes the common density of R_i given X. 2. $0 \leq \underline{B}_{f_X} \leq f_X(x) \leq \bar{B}_{f_X} < \infty$

for all $x \in G$, G a compact subset of $\Theta = \times_{i=1}^{p}(0,\infty)$, which denotes the Cartesian product of the intervals $(0,\infty)$.

Assumption A2. 1. $Y_i = \sigma(X_i)\frac{R_i}{\sigma_R}$. 2. $R_i \in [0,1], X_i \in \Theta$. 3. $E(R_i|X_i) = \mu_R, V(R_i|X_i) = \sigma_R^2$. 4. The regression function $m(x)$ has a bounded and continuous second derivative for all $x \in \Theta$, which will be denoted by $m^{(2)}(x)$. 5. $0 < \underline{B}_\sigma \leq \sigma(x) \leq \bar{B}_\sigma < \infty$ for all $x \in \Theta$.

Assumption A3. $K(x) : S_p \to \mathbb{R}$ is a symmetric density function with bounded support $S_p \to \mathbb{R}^p$ satisfying: 1. $\int xK(x)dx = 0$. 2. $\int x^2K(x)dx = \sigma_p^2$. 3. For all $x \in \text{ı}, |K(x)| < B_p < \infty$. 4. For all $x, x' \in \mathbb{R}_p, |K(x) - K(x')| < m\|x - x'\|$ for some $0 < m < \infty$.

Assumption A4. For all $x, x' \in \Theta, |g_X(x) - g_X(x')| < m_g\|x - x'\|$ for some $0 < m < \infty$.

Assumptions $A1.1$ and $A2$ imply that $\{Y_i, X_i\}_{i=1}^n$ forms an iid sequence of random variables with some joint density $\phi(y,x)$. Comparing with Martins-Filho and Yao (2007) and Martins-Filho et al. (2008), we do not have to assume anything about the second derivative of $\sigma^2(x)$. Furthermore, we do not have to deal with regressands that are themselves residuals from a first stage nonparametric regression, due to elimination of the second step in the estimation procedure. Therefore, asymptotic properties are much easier to obtain. The uniform consistency and asymptotic normality of the frontier estimator are presented in the following Theorems.

Theorem 1 *Suppose that Assumptions A1-A4 are holding. In addition assume that* $E(|\epsilon_i|^{2+\delta}|X_i) < C_1 < \infty$. *Then for every* $x \in G$

$$\sqrt{nh_n}\left(\hat{m}(x,h_n) - m(x) - B_{1n}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{f_X(x)}\int K^2(\phi)d\phi\right),$$

where, $B_{1n} = \frac{h_n^2 m^{(2)}(x)\sigma_k^2}{2} + o_p(h_n^2)$.

Theorem 1 is nowadays standard in nonparametric literature. A proof of this result may be viewed in Fan and Yao (1992). The following theorem establishes the asymptotic normality of the NP2S estimator.

Theorem 2 *Let* L_n *be a non-stochastic sequence such that* $0 < L_n \to 0$ *as* $n \to \infty$ *and suppose that* (i) $\hat{m}(x,g_n) - m(x) = O_p(L_n)$ *uniformly in* G *and* (ii) $1 - \hat{q}_\alpha(R) = O_p(L_n)$. *Then,*

a) $\hat{\mu}_R(g_n) - \mu_R = O_p(L_n)$;

b) Under the assumptions $A1 - A4$, *if* $\frac{ng_n^5}{ln(n)} \to \infty$, $nh_n^5 = o(1)$, *and* $nh_ng_n^4 = O(1)$, *then:*

$$\sqrt{nh_n}\left(\frac{\hat{m}(x,h_n)}{\hat{\mu}_R(g_n)} - \frac{m(x)}{\mu_R} - B_{2n}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{\mu_R^2 f_X(x)}\int K^2(\phi)d\phi\right), \tag{9}$$

where, $B_{2n} = O_p(g_n^2)$.

Informally, assumption (2) in Theorem 2 guarantees that for appropriate values of α the sample $\alpha-$quantile of R is sufficiently close to 1 as the sample size increases. The proof of Theorem 2 is presented in Appendix 2. Its worth to point out that Theorem 2 concerns asymptotic normality of NP2S centered at the true frontier, $\rho(\cdot)$.

3.1 NP2S and NP3S comparison

Now we compare the asymptotic variances of the estimators NP2S and NP3S. Using the results stated in Martins-Filho and Yao (2007) and the results presented in equation (9) we have the ratio between the two aymptotic variances:

$$\frac{Avar_{NP2S}}{Avar_{NP3S}} = \frac{4\sigma_R^2}{\mu_R^2(\mu_4(x)-1)}, \tag{10}$$

where $\mu_4(x) = E(\epsilon_i^4|X_i = x)$ and $\epsilon_i = \frac{R_i=\mu_R}{\sigma_R}$. Therefore, it is clear that the ratio above become smaller as μ_r increases and σ_R decreases. Hence big μ_R combined with small σ_R values tend favorably to NP2S estimator vis-à-vis NP3S estimator. An illustration about this point is made in the next section.

4 Monte Carlo Study

In this section we consider a simulation study to shed some light on the finite sample properties of our estimator, henceforth referred to as $NP2S_{100\alpha}$, where α stands for the quantile order in the second step of the proposed estimator. We consider $\alpha = 0.95, 0.96, 0.97, 0.98, 0.99, 1$. For comparison purposes, we also include in the study the local linear frontier estimator proposed in Martins-Filho and Yao (2007), referred to as NP3S, and the well known FDH estimator. Our simulations are based on model (1), i.e., $Y_i = \frac{\sigma(X_i)R_i}{\sigma_R}$, with $p = 1$. We generate data with the following characteristics. The X_i are pseudorandom variables from a uniform distribution with support given by $[a_l, b_u]$. $R_i = \exp(-Z_i)$, where Z_i are pseudorandom variables from an exponential distribution with parameter $\beta > 0$, therefore R_i has support on $(0, 1]$. We consider two specifications for $\sigma(x)$:

$$\sigma_1(x) = \sqrt{x}, \text{ with } x \in [a_l, b_u] = [10, 100] \text{ and}$$

$$\sigma_2(x) = 3(x - 1.5)^3 + 0.25x + 1.125, \text{ with } x \in [a_l, b_u] = [1, 2],$$

which are associated with convex and non-convex production technologies. Four parameters for the exponential distribution are considered: $\beta_1 = 1$, $\beta_2 = 1/3$. These choices of parameters produce, respectively,

the following values for the parameters of $g_{R|X}$: $(\mu_R, \sigma_R^2) = (0.5, 0.08)$ and $(0.75, 0.04)$. Two sample sizes $n = 200, 400$ were used. Each experiment involved 1000 Monte Carlo replications. We evaluate the frontiers at $x_0 = 55$ for $\sigma_1(x)$ and at $x_0 = 1.5$ for $\sigma_2(x)$. These values of X correspond to the 50th percentile of its support.

The results of our simulations are summarized in figures 1-18. For figures 1-16, the thick horizontal line inside the rectangle in each boxplot corresponds to the median of the distribution, and the rectangle height corresponds to interquartile range. Consequently 50% of data is represented by the rectangle. The two thin horizontal lines below and above the rectangle are the whiskers. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range. Figures 1-8 give boxplots of MSE for the frontier estimator ($\hat{\rho}(\cdot)$). Each boxplot is constructed from 1000 points (repetitions), where each point corresponds to a sample draw and is calculated as the squared Euclidean distance between the estimate and true value of $\rho(\cdot)$, as in equation (11). Figures 9-16 give boxplots of the frontier estimator evaluated at the selected points mentioned above, i.e., $\hat{\rho}(x_0)$. The horizontal red line represents the true value, i.e, $\rho(x_0)$.

$$MSE(\hat{\rho}) = n^{-1} \sum_{i=1}^{n} (\rho(X_i) - \hat{\rho}(X_i))^2. \tag{11}$$

An important aspect in the implementation of our frontier estimator is bandwidth selection. We minimize the following expression, obtained from the asymptotic analysis presented above.

$$A\hat{M}ISE(h) = \frac{1}{nh_n} \int \dot{\sigma}^2(x)dx \int k^2(\phi)d\phi + \left(\hat{q}_\alpha \left(\frac{2\dot{m}(x_i)\dot{R}_i}{2\dot{m}(x_i) + h_n^2 n^{2\gamma} \dot{m}^{(2)}(x_i)\sigma_k^2} \right) \right)^2 \frac{1}{n} \sum_{i=1}^{n} \dot{m}^2(x_i) \tag{12}$$

where γ is set to be zero in all experiments. Other values in the range $(0, 1/6)$ were considered for γ. Since the results were similar we do not present them here. The sequence $\{\dot{m}(X_i)\}_{i=1}^n$ is estimated via local linear regression with a rule-of-thumb bandwidth as in Ruppert et al. (1995). $\{\dot{\sigma}^2(X_i)\}_{i=1}^n$ is estimated via local linear regression of $\{\dot{\epsilon}^2\}_{i=1}^n$ on $\{X_i\}_{i=1}^n$, where $\dot{\epsilon}_i^2 = (Y_i - \dot{m}(X_i))^2$, as in Fan and Yao (1998). Furthermore, $\{\dot{m}(X_i)\}_{i=1}^n$ is used to estimate $\dot{R}_i = \frac{Y_i}{\dot{m}(X_i)} \left(\hat{q}_\alpha \left(\frac{Y_i}{\dot{m}(X_i)} \right) \right)^{-1}$. Finally, the sequence $\{\dot{m}^{(2)}(X_i)\}_{i=1}^n$ is estimated with an ordinary least square quartic regression of $\{Y_i\}_{i=1}^n$ on $\{X_i\}_{i=1}^n$.

4.1 Performance without outliers

As expected from the asymptotic results of section 3, as the sample size n increases, the boxplots in figures 1-8 show that MSE decreases for all estimators and values for μ_R considered. As pointed out in subsection 3.1, the asymptotic variance ratio between NP2S and NP3S depends on the values of μ_R and σ_R. We note that

for NP2S, regarding the frontier estimator, the performance in terms of MSE improves as the value of μ_R increases. This pattern is most likely explained by the fact its variance is inversely proportional to μ_R value, as stated in Theorem 2. In fact, when $\mu_R = 0.5$, $NP2S_{100}$ is similar in overall performance to $NP3S$, but with an appropriate choice of α, $NP2S$ exhibits a much smaller MSE than $NP3S$. In general, FDH presents smaller distance between whiskers than $NP2S$ and $NP3S$, but depending on choice of α, $NP2S$ presents smaller median than FDH and $NP3S$. In figure 17 we illustrate the behavior of the three estimators. We emphasize that $NP2S$ is a smooth estimator with good adjustment to the data.

4.2 Performance with outliers

In Figures 9-16 we consider samples with outliers, as explained in section 4. We see, as expected, that FDH estimator is very sensible to outliers from the point it arises to the point where at least one observed output becomes greater than the referred outlier. $NP3S$ and $NP2S_{100}$ are also very sensible to outliers but in a different fashion. For both estimators an outlier tends to interfere in the positioning step, making the entire estimated frontier to be located above of the true frontier. Nevertheless, choosing α smaller than one seems to be a valuable strategy to overcome this problem. Indeed within this framework the outliers seem to have no interference in the positioning step. In figure 18 we illustrate the behavior of the three estimators in the presence of an outlier. $NP2S$ is not affected by the outlier in that scenario.

5 Conclusion

In this paper we present a production frontier model developed by Martins-Filho and Yao (2007) that uses Kernel regression for estimating production frontier and therefore efficiency for production units with significant advantages when compared to DEA and FDH estimators. However, the estimation process is made in three steps. We then propose a modification on that estimation procedure, eliminating the need of the second step. The result is a simpler estimation procedure that retains all inherent advantages present in the original estimator. Furthermore, we propose a generalization of the positioning step, which results in a robust estimator. A Monte Carlo study was performed comparing three estimators: our estimator, called NP2S; NP3S from Martins-Filho and Yao (2007); and the well known FDH estimator. The results show that depending on choice of α, NP2S outperforms its competitors, specially when outliers are present.

Appendix 1: Tables and Graphics

Figure 1: Frontier I - MSE of Estimators - $n = 200$ - $\mu_R = 0.5$

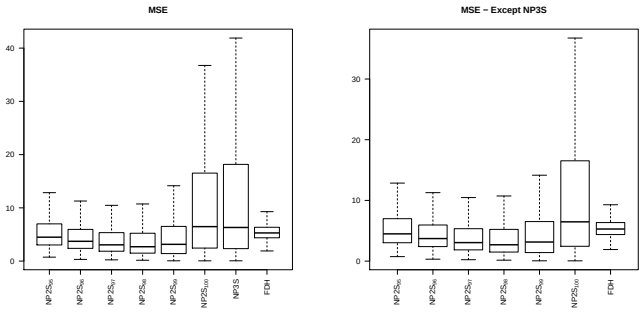


Figure 2: Frontier I - MSE of Estimators - $n = 400$ - $\mu_R = 0.5$

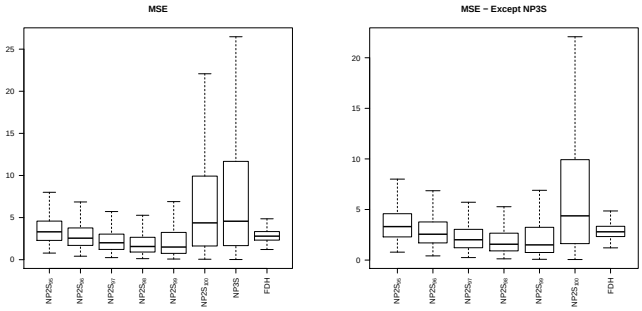


Figure 3: Frontier I - MSE of Estimators - $n = 200$ - $\mu_R = 0.75$

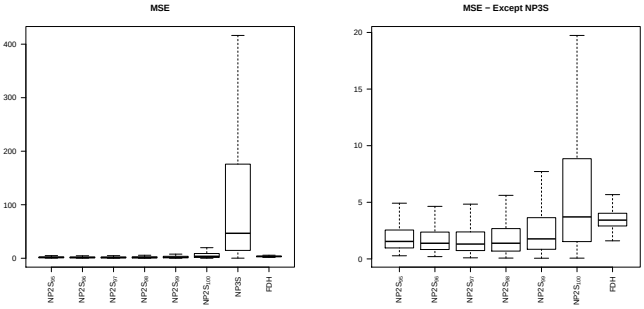


Figure 4: Frontier I - MSE of Estimators - $n = 400$ - $\mu_R = 0.75$

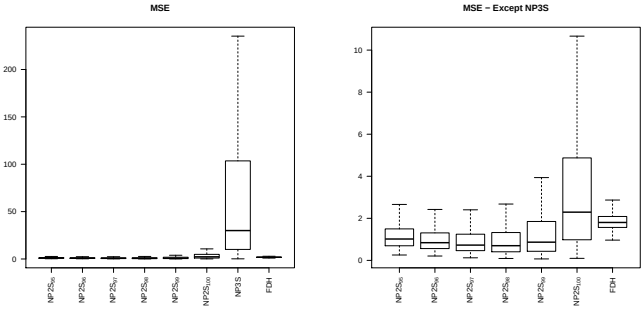


Figure 5: Frontier II - MSE of Estimators - $n = 200$ - $\mu_R = 0.5$

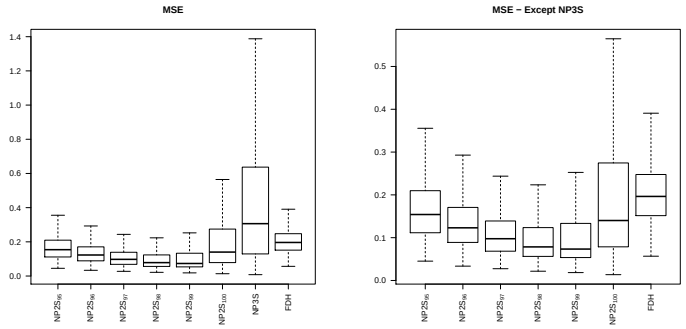


Figure 6: Frontier II - MSE of Estimators - $n = 400$ - $\mu_R = 0.5$

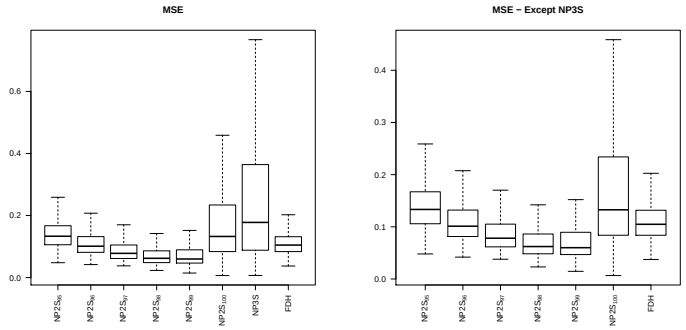


Figure 7: Frontier II - MSE of Estimators - $n = 200$ - $\mu_R = 0.75$

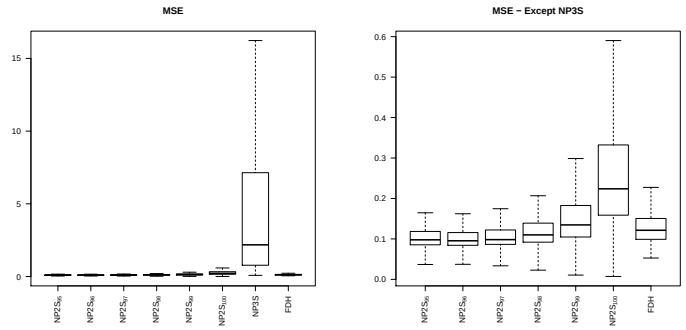


Figure 8: Frontier II - MSE of Estimators - $n = 400$ - $\mu_R = 0.75$

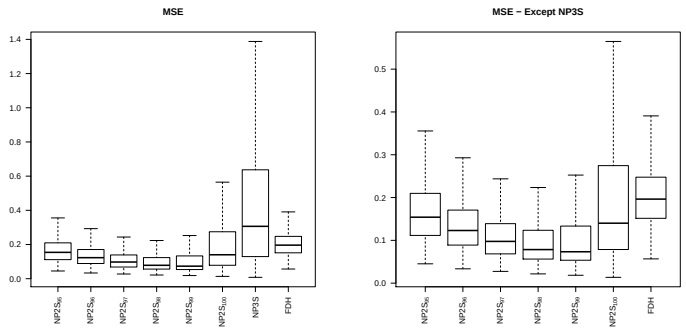


Figure 9: Frontier I - Estimated frontier at selected points - Outlier at $x = 55$ - $n = 200$ - $\mu_R = 0.5$

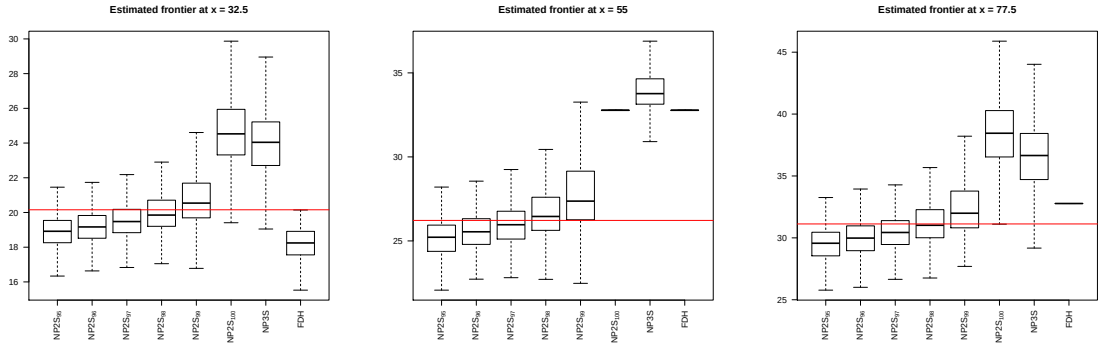


Figure 10: Frontier I - Estimated frontier at selected points - Outlier at $x = 55$ - $n = 400$ - $\mu_R = 0.5$

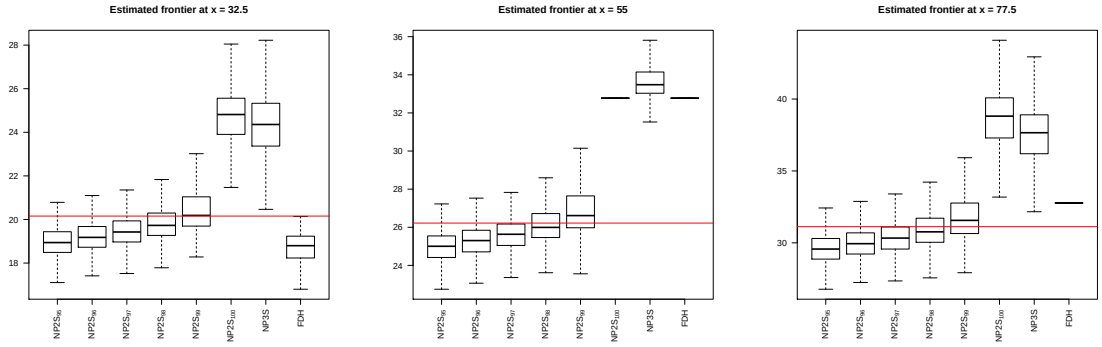


Figure 11: Frontier I - Estimated frontier at selected points - Outlier at $x = 55$ - $n = 200$ - $\mu_R = 0.75$

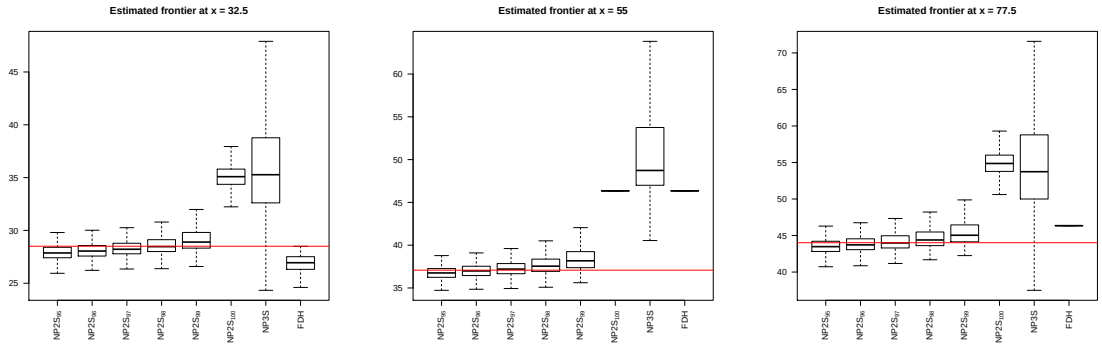


Figure 12: Frontier I - Estimated frontier at selected points - Outlier at $x = 55$ - $n = 400$ - $\mu_R = 0.75$

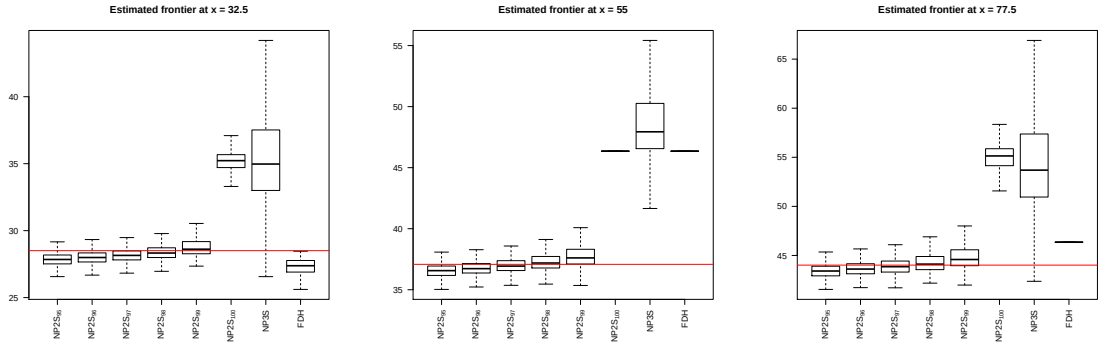


Figure 13: Frontier II - Estimated frontier at selected points - Outlier at $x = 1.5$ - $n = 200$ - $\mu_R = 0.5$

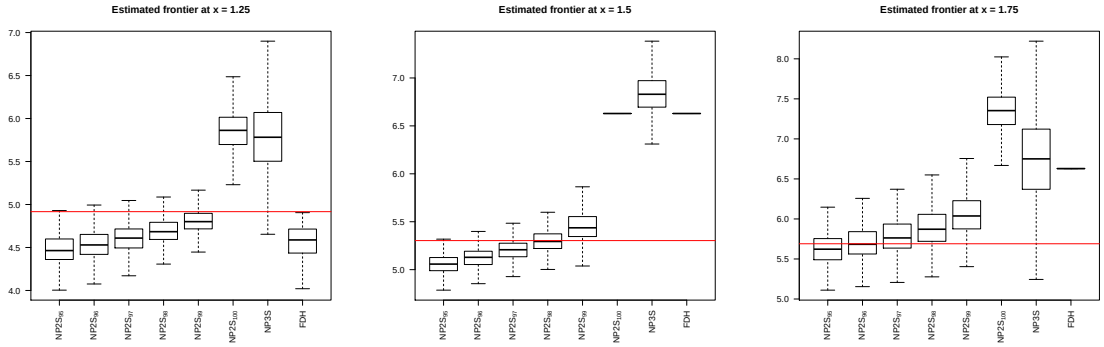


Figure 14: Frontier II - Estimated frontier at selected points - Outlier at $x = 1.5$ - $n = 400$ - $\mu_R = 0.5$

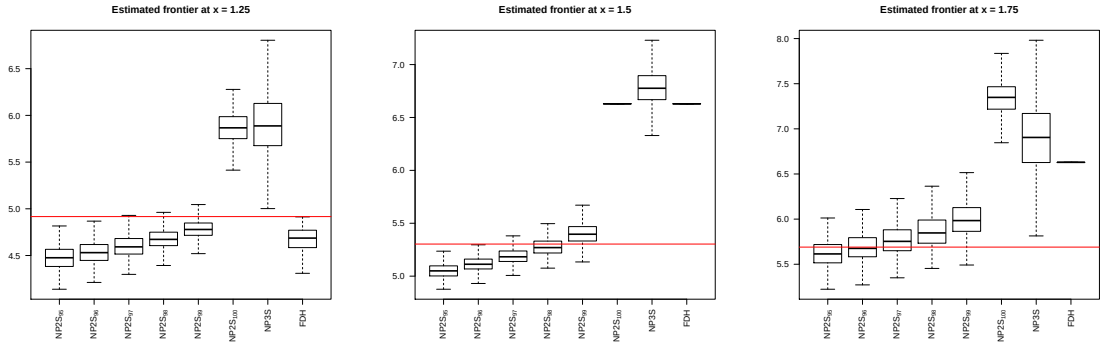


Figure 15: Frontier II - Estimated frontier at selected points - Outlier at $x = 1.5$ - $n = 200$ - $\mu_R = 0.75$

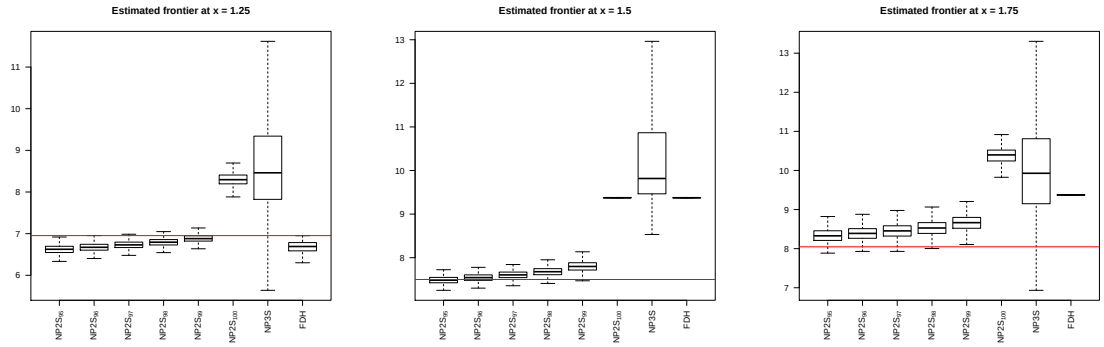


Figure 16: Frontier II - Estimated frontier at selected points - Outlier at $x = 1.5$ - $n = 400$ - $\mu_R = 0.75$

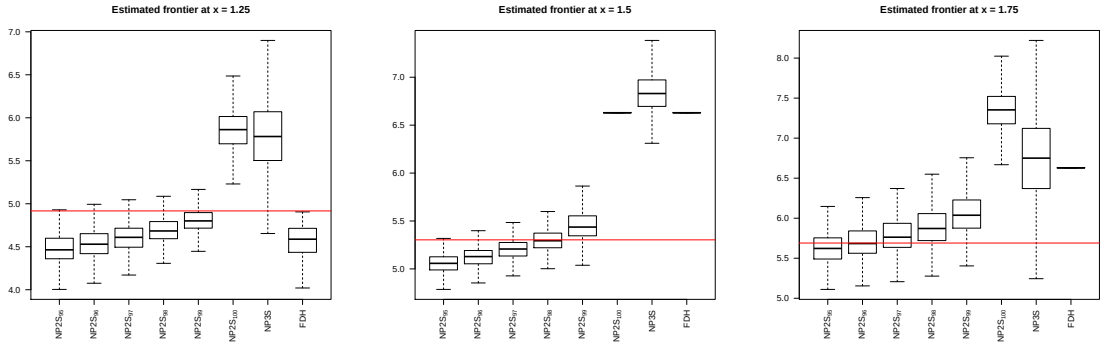
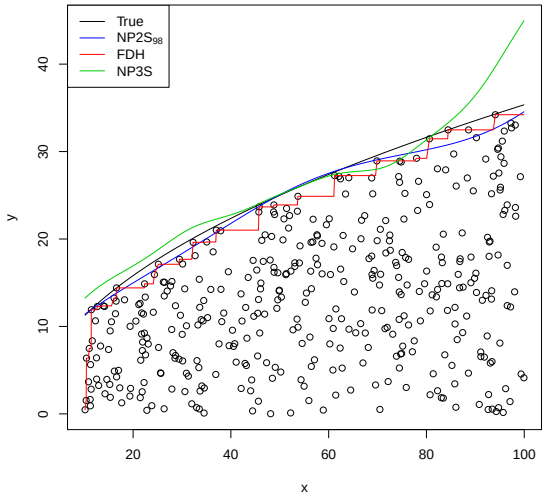
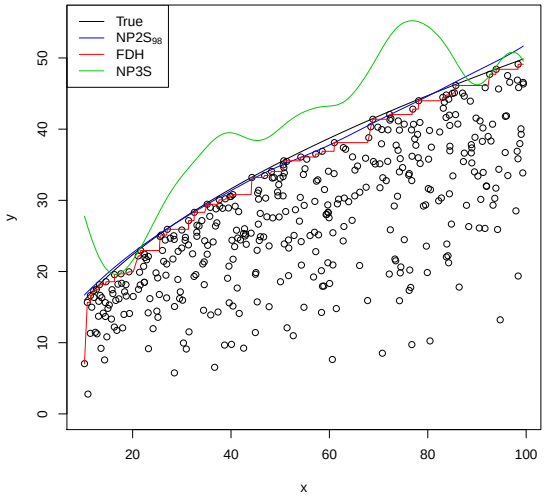


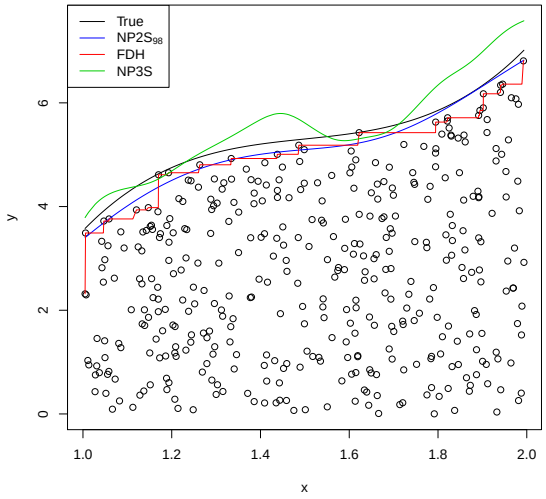
Figure 17: Estimated frontier - Example - $n = 400$



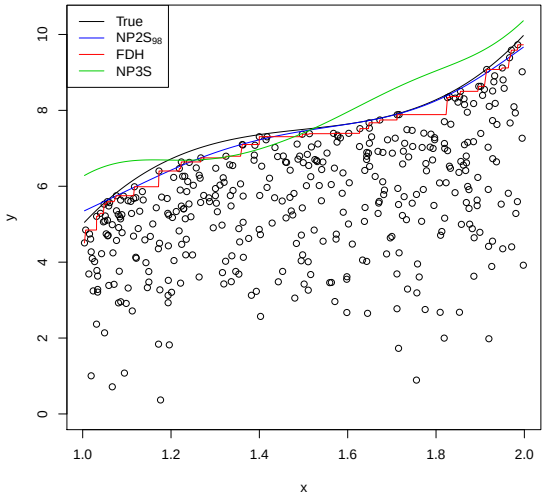
(a) Frontier I - $\mu_R = 0.5$



(b) Frontier I - $\mu_R = 0.75$

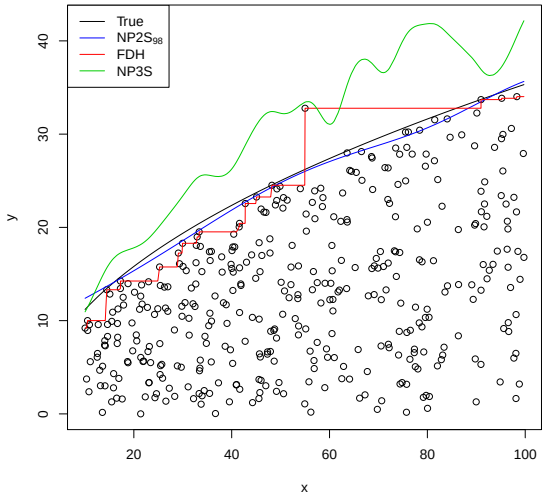


(c) Frontier II - $\mu_R = 0.5$

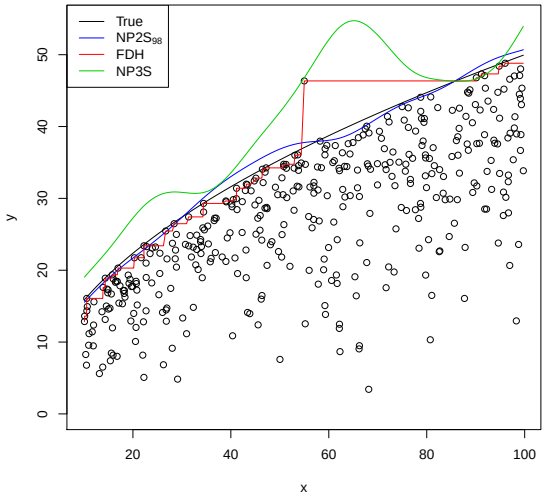


(d) Frontier II - $\mu_R = 0.75$

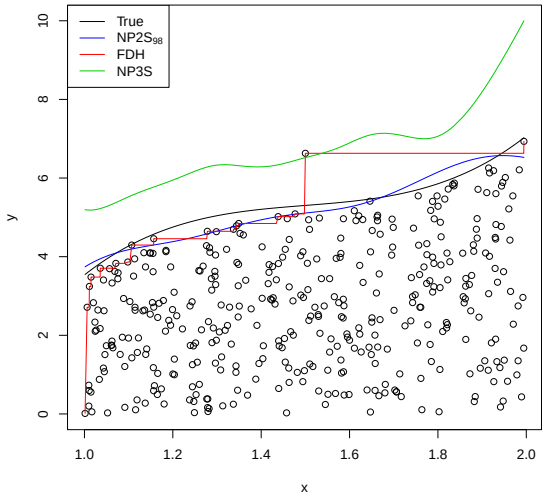
Figure 18: Estimated frontier - Example - $n = 400$ - Outlier



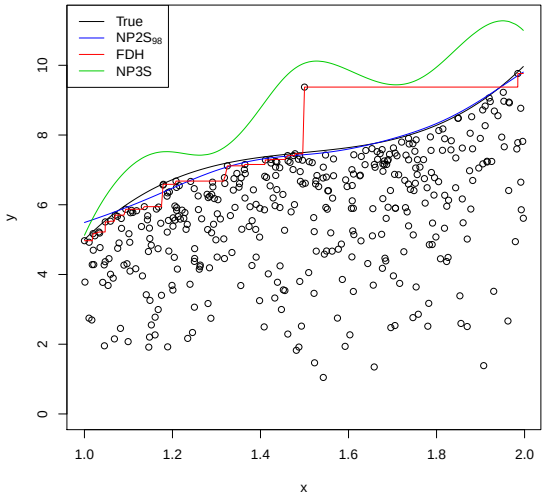
(a) Frontier I - $\mu_R = 0.5$



(b) Frontier I - $\mu_R = 0.75$



(c) Frontier II - $\mu_R = 0.5$



(d) Frontier II - $\mu_R = 0.75$

6 Appendix 2: Proofs

Proof of Theorem 2: To prove Theorem 2 we first note that Martins-Filho and Yao (2007) get after two steps $\hat{\sigma}(X_t, h_n)$ which is in fact $\sigma_R\hat{\rho}(X_t, h_n)$. Then, to obtain asymptotic normality of the estimated frontier, $\hat{\rho}(.)$, they divided $\hat{\sigma}(x, h_n)$ by $s_R(g_n)$ and combine their Theorem 1 and their Theorem 2 part (a) to achieve the desired result. In our case, after one step, we get $\hat{m}(X_i, h_n)$ which is in fact $\mu_R\hat{\rho}(X_i, h_n)$. Therefore, to obtain the result claimed in our Theorem 2 part (b), we just need to combine the results from our Theorem 1 and our Theorem 2 part (a).

To prove Theorem 2 part (a), we use the same argument presented in the proof of Theorem 2 part (a) of Martins-Filho and Yao (2007); but substituting in their proof $\sigma(X_t)$ by $m(X_i)$ as well as $\hat{\sigma}(X_t, g_n)$ by $\hat{m}(X_i, g_n)$, and σ_R by μ_R as well as $s_R(g_n)$ by $\hat{\mu}_R(g_n)$.

For a proof of part (b), we note that

$\sqrt{nh_n}\left(\frac{\hat{m}(x,h_n)}{\mu_R} - \frac{m(x)}{\mu_R} - \frac{B_{1n}}{\mu_R}\right) \equiv \sqrt{nh_n}\left(\frac{\hat{m}(x,h_n)}{\hat{\mu}_R(g_n)} - \frac{m(x)}{\mu_R} - \hat{m}(x,h_n)\left(\frac{1}{\hat{\mu}(g_n)} - \frac{1}{\mu_R}\right) - \frac{B_{1n}}{\mu_R}\right).$

From Theorem 1 we have

$\sqrt{nh_n}\left(\frac{\hat{m}(x,h_n)}{\mu_R} - \frac{m(x)}{\mu_R} - \frac{B_{1n}}{\mu_R}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{\mu_R^2 f_X(x)}\int K^2(\phi)d\phi\right),$

and from Theorem 2 part (a), provided that $\frac{ng_n^5}{ln(n)} \to \infty$ we have that $\hat{\mu}_R(x,h_n)(\hat{\mu}_R(g_n)^{-1} - {\mu_R}^{-1}) = O_p(g_n^2)$. Hence, given that $nh_n^5 \to 0$ and $nh_ng_n^4 = O(1)$

$$\sqrt{nh_n}\left(\frac{\hat{m}(x,h_n)}{\hat{\mu}_R(g_n)} - \frac{m(x)}{\mu_R} - B_{2n}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{\mu_R^2 f_X(x)}\int K^2(\phi)d\phi\right),$$

where, $B_{2n} = O_p(g_n^2)$. □

7 References

Aigner, D., C.A.K. Lovell and P. Schmidt, 1977, Formulation and estimation of stochastic frontiers production function models. Journal of Econometrics 6, 21-37.

Aragon, Y., Daouia A., and Thomas-Agnan, C. 2005, Nonparametric Frontier Estimation: a conditional quantile-based approach. Econometric Theory, 21, 2005, 358-389.

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

Charnes, A.,W. Cooper and E. Rhodes, 1978, Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429-444.

Deprins, D., L. Simar and H. Tulkens, 1984, Measuring labor inefficiency in post offices, in: M. Marchand, P. Pestiau and H. Tulkens, (Eds.), The performance of public enterprises: concepts and measurements. North Holland, Amsterdam.

Fan, J., 1992, Design-adaptive Nonparametric Regression. Journal of the American Statistical Association, Vol. 87, No. 420, 998-1004.

Fan, J. and I. Gijbels, 1995, Data driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. Journal of the Royal Statistical Society, B 57, 371-394.

Fan, J. and Gijbels, I., 1996, Local Polynomial Modelling and Its Applications. London: Chapman and Hall.

Fan, J., and Q. Yao, 1998, Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645-660.

Farrell, M., 1957, The measurement of productive efficiency. Journal of the Royal Statistical Society A 120, 253-290.

Gijbels, I., E. Mammen, B. Park and L. Simar, 1999, On estimation of monotone and concave frontier functions. Journal of the American Statistical Association 94, 220-228.

Korostelev, A. P., L. Simar and A. B. Tsybakov, 1995, Efficient estimation of monotone boundaries. Annals of Statistics 23, 476-489.

Martins-Filho, C. and Yao, F., 2007, Nonparametric frontier estimation via local linear regression. Journal of Econometrics.

Martins-Filho, C., Torrent, H., Ziegelmann, F., 2013, Nonparametric Frontier Estimation: Using Local Exponential Regression for Conditional Variance. Brazilian Review of Econometrics.

Park, B., L. Simar and Ch. Weiner, 2000, The FDH estimator for productivity efficient scores: asymptotic properties. Econometric Theory 16, 855-877.

Seifford, L., 1996, Data envelopment analysis: the evolution of the state of the art (1978-1995). Journal of Productivity Analysis 7, 99-137.

Silverman, B.W., 1986, Density estimation for statistics and data analysis. Chapman and Hall, London.

Ziegelmann, F. A., 2002, Nonparametric estimation of volatility functions: the local exponential estimator. Econometric Theory, 18: 985-992.