Robust Nonparametric Frontier Estimation in Two Steps

Yining Chen[†], Hudson S. Torrent[‡] and Flavio A. Ziegelmann[‡]

[†] *Department of Statistics, London School of Economics and Political Science*
[‡] *Department of Statistics, Federal University of Rio Grande do Sul*

ABSTRACT

We propose a robust methodology for estimating production frontiers with multi-dimensional input via a two-step nonparametric regression, in which we estimate the level and shape of the frontier before shifting it to an appropriate position. Our main contribution is to derive a novel frontier estimation method under a variety of flexible models which is robust to the presence of outliers and possesses some inherent advantages over traditional frontier estimators. Our approach may be viewed as a simplification, yet a generalization, of those proposed by Martins-Filho and Yao (2007) and Martins-Filho et al. (2013), who estimate frontier surfaces in three steps. In particular, outliers, as well as commonly seen shape constraints of the frontier surfaces, such as concavity and monotonicity, can be straightforwardly handled by our estimation procedure. We show consistency and asymptotic distributional theory of our resulting estimators under standard assumptions in the multi-dimensional input setting. The competitive finite-sample performances of our estimators are highlighted in both simulation studies and empirical data analysis.

Keywords: *concavity, local polynomial smoothing, monotonicity, outlier detection, shape-constrained regression.*

1. INTRODUCTION

Estimation of production frontiers, and therefore efficiency, has motivated a wide and growing literature during the last decades. Mathematically, the problem can be stated as follows. Let $\mathbf{x} \in \mathbb{R}_+^p$ be some inputs (represented in row vector form) used to produce output $y \in \mathbb{R}_+$. A production set is defined as $\Omega = \{(\mathbf{x}, y) \in \mathbb{R}_+^{p+1} \mid \mathbf{x} \text{ can produce } y\}$, whereas the production frontier associated with Ω is defined as $\rho(\mathbf{x}) = \sup\{y \in \mathbb{R}_+ \mid (\mathbf{x}, y) \in \Omega\}$ for all $\mathbf{x} \in \mathbb{R}_+^p$. For any given $(\mathbf{x}_0, y_0) \in \Omega$, the efficiency is measured by the ratio between y_0 and $\rho(\mathbf{x}_0)$. Our aim is to obtain, from a given random sample $\{(\mathbf{X}_i, Y_i), i = 1, ..., n\}$, nonparametric frontier estimators which can readily deal with multiple inputs and are reasonably robust to the presence of outliers. By referring to the presence of outliers, we mean the potential existence of a few observations which lie outside Ω.

In this manuscript, we restrict ourselves to the deterministic approach for the frontier estimation problem. This approach relies on the assumption that all observations, with perhaps the exception of a few outliers, lie in the production set. Two popular methods in the literature of deterministic frontier estimation are the Free Disposal Hull (FDH) estimator introduced by Deprins et al. (1984) and Data Envelopment Analysis (DEA) represented by Charnes et al. (1978). These methodologies are applied in many subsequent pieces of work, such as Seifford (1996), Daraio and Simar (2007), Simar and Wilson (2013) and Kneip et al. (2015). See also Badunenko et al. (2012). Outlier detection and treatment techniques under these approaches can be found in Simar (2003), Johnson and McGinnis (2008) and Khezrimotlagha et al. (2008), among others. An alternative modelling approach is the so-called *Stochastic Frontier Analysis (SFA)*, introduced by Aigner et al. (1977) and Meeusen and van Den Broeck (1977). For a comprehensive overview, see Parmeter and Kumbhakar (2014), for example. Papadopoulos and Parmeter (2022)

extensively describe quantile methods for SFA robust analyses. See also Parmeter and Racine (2013) where additional constraints are imposed.

Moving away from FDH and DEA in the context of deterministic frontier estimation, Martins-Filho and Yao (2007) proposed a deterministic production frontier model and a nonparametric production frontier estimator called NP3S[1]. They assumed the efficiency score across different observations to have constant mean and variance, and proposed an estimation procedure that consists of three steps. The first step estimates a conditional mean of the output with respect to the inputs using the local linear kernel method. The second step follows Fan and Yao (1998) and again uses the local linear kernel approach to estimate the conditional variance of the output with respect to the inputs. The third and final step gives an original estimator to their proposed production frontier model.

A possible drawback in the second step of NP3S estimator is that it allows for a negative estimate of the variance. To overcome this problem, Martins-Filho et al. (2013) propose to use the local exponential kernel estimator to estimate conditional volatility functions, ensuring its non-negativity. This estimator uses an exponential functional at the minimization problem that characterizes kernel regressions for estimating nonnegative conditional variance (see Ziegelmann (2002)). We call this frontier estimator NPE, standing for **N**on**P**arametric **E**xponential.

Considering NP3S and NPE, we believe some improvements are desirable, as summarized below:

(i) These estimators are characterized by an estimation procedure in three steps. The first two steps capture the shape of the frontier and the third step is responsible for locating the estimated frontier. It is important to emphasize that the second step of NP3S and NPE is based on a regression that has as regressand squared residuals from the first step. This feature is sometimes undesirable, since multiple choices of tuning parameters are needed.

(ii) Two constant conditional moment conditions on the random variable that represents efficiency are required, which might potentially be viewed as restrictive.

(iii) Their methods are solely based on the local linear kernel estimators, making it not straightforward to incorporate some commonly seen or well accepted shape constraints of the frontiers.

(iv) The estimation accuracy of the frontier using these procedures can be severely affected by the existence of a few outliers.

In this work, to address the first two issues, we eliminate the second step of NP3S and NPE estimators, thus estimating the frontier in just two steps. In doing so, we also manage to relax the condition of requiring constant mean and variance for the efficiency scores in Martins-Filho and Yao (2007) to just requiring a constant mean. Not only is this of theoretical interest, we also believe that this relaxation is of practical relevance. For instance, it is plausible that the efficiency scores of larger companies tend to be more concentrated as compared to those of the smaller ones, as in many industries, the spectrum of the larger firms often tends to be more homogeneous than that of the smaller ones.

In addition, to address the third issue, we advocate that our estimation framework go beyond the use of the local linear kernel method of Fan (1992). Constrained regression methods, such as those enforcing concavity, monotonicity and/or additivity (c.f. Groeneboom et al. (2001b), Mammen and Yu (2007), Chen and Samworth (2016), among others) can be used. As most of the frontiers do follow certain shapes, imposing shape constraints is natural and usually helps improving the interpretability of the estimator. Besides, in comparison to kernel-based estimators, shape-constrained approach

[1] In this paper, we call the estimator proposed by Martins-Filho and Yao (2007) as NP3S, standing for **N**on**P**arametric estimation in **3** **S**teps.

does not rely on, or is sensitive to, the careful choice of tuning parameters such as the bandwidth and seems to enjoy better finite-sample performance in many different settings.

Finally, to handle with the fourth issue, we extend this framework to allow the presence of outliers using a quantile-based approach in the second step. These outliers, for one reason or another, should perhaps not have been included in the production set for the analysis. This could arise when there are a few firms behaving significantly different from others in their cohort (so estimating the production set for this analysis based on these observations could be problematic[2]), or very rarely there are perhaps typos in the data recording process. In the asymptotic regime, it also implies that there is a vanishing proportion of outliers, because otherwise identifiability of the frontier could potentially become an issue. Here we propose to invoke the $(100\alpha_n)\%$-quantile to estimate the frontier, where $\alpha_n \in (0,1)$ is a number chosen close to one to guarantee there is no actual effect on extra shifting the estimated frontier curve upwards or downwards. A more-involved iterative procedure, that could further improve its finite-sample performance, is also outlined.

To summarise this manuscript's main contribution to the literature, our proposed robust two-step approach (which will be, for simplicity, called NP2S, standing for Robust **N**on**P**arametric estimation in **2** **S**teps) could easily incorporate nonparametric estimators other than local linear to enforce certain shape-constraints, and allows for the presence of outliers. In addition, we also comprehensively investigate the theoretical properties of our approach beyond the one-dimensional setting (i.e. $p > 1$), which is the main focus of Martins-Filho and Yao (2007) and compare the numerical performance with existing benchmarks such as those by Fang et al. (2022).

We note that similar ideas based on the two-step estimation have also been explored independently by Wang and Yang (2020) in the context of additive models using splines, and by Fang et al. (2022) using quantile regression (their estimator is implemented and used for comparison purposes in our numerical experiments). In comparison to their work, we believe that our work contributes and extends to the literature in the following ways. First, on the theoretical front, we provide both the asymptotic distributional theory and the results regarding the minimaxity of the problem, and make more thorough comparisons with the original NP3S. Second, we investigate models that are not necessarily additive, being more flexible in terms of the shape constraints we incorporate. Third, in terms of methodology, we also establish that robust frontier estimation can be achieved via locating the $(100\alpha_n)\%$-quantile frontier with $\alpha_n \to 1$ as $n \to \infty$. Finally, we also provide theory of our proposed robust procedure in the presence of outliers.

The remaining of this paper is composed as follows. In Section 2, we briefly review the model originally proposed by Martins-Filho and Yao (2007) and present our new estimation process. Section 3 discusses the asymptotic properties of our estimators, as well as the minimax convergence rates. A Monte Carlo study comparing different estimators and an empirical example are presented in Section 4, which is followed by the conclusion and final comments in Section 5. All the proofs are deferred to the supplementary materials.

[2]If one were able to identify relevant environmental variables, which changed the production set depending on their values, then conditional frontier modelling would be a good alternative (see Daraio and Simar (2005), for example.)

2. ROBUST NONPARAMETRIC FRONTIER ESTIMATION IN TWO STEPS

2.1. The model

We start by presenting the model setup developed by Martins-Filho and Yao (2007). Our interest lies in estimating the frontier from a set of observed firms, i.e. given a random sample of production units $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ that share a technology Ω, obtaining estimates of the frontier ρ. By extension we are also interested in constructing efficiency ranks and relative performance of production units. To see this, let $(\mathbf{X}_i, Y_i) \in \Omega$ characterize the performance of a production unit and define $0 \leq R_i \equiv Y_i/\rho(\mathbf{X}_i) \leq 1$ to be this unit's (inverse) Farrell output efficiency measure. R_i can then be estimated based on the estimates of ρ.

Our frontier regression model consists of a multiplicative regression. Primitive assumptions take place on $(\mathbf{X}_i, R_i)$ and the properties of Y_i arise from a suitable regression function. It is typically assumed that $\{(\mathbf{X}, R), (\mathbf{X}_1, R_1), (\mathbf{X}_2, R_2), \ldots\}$ is a sequence of $(p+1)-$dimensional independent and identically distributed random vectors with a common density. Y_i then follows

$$Y_i = \rho(\mathbf{X}_i) R_i \tag{2.1}$$

where R_i is an unobserved random variable and $\mathbf{X}_i$ is an observed random vector in $\mathbb{R}_+^p$. In this context, Y_i is the output, ρ is the production frontier, $\mathbf{X}_i$ are the inputs and R_i is the efficiency with values in $[0, 1]$. The closer R_i is to 1, the closer are the observed output and the frontier. On the contrary, Y_i being far from $\rho(\mathbf{X}_i)$ implies low efficiency and a small value for R_i. There is no specification of the density of R_i, however the following moment restriction is assumed for all $\mathbf{x}$:

$$E(R_i|\mathbf{X}_i = \mathbf{x}) \equiv \mu_R; \text{ where } 0 < \mu_R < 1. \tag{2.2}$$

In addition, we write

$$\mathrm{Var}(R_i|\mathbf{X}_i = \mathbf{x}) \equiv \sigma_R^2(\mathbf{x}).$$

Note that the facts that $R_i \in [0, 1]$ and $0 < \mu_R < 1$ imply by construction that $0 < \sigma_R^2(\mathbf{x}) < \mu_R < 1$, as the variance is majorized by that of a Bernoulli random variable with the probability of success being μ_R.

The unknown quantity μ_R and the function σ_R locate the production frontier. For example, if a random sample of a population is far from the true frontier, efficiency is low hence the corresponding μ_R and σ_R are small. Furthermore, we note that the NP3S estimator requires the additional constant second moment assumption of $\mathrm{Var}(R_i|\mathbf{X}_i = \mathbf{x}) \equiv \sigma_R^2$.

2.2. The estimation procedures

2.2.1. Frontier estimation without outliers To characterize our estimating procedure, we first rewrite equation (2.1) as

$$Y_i = \rho(\mathbf{X}_i) R_i = \mu_R\, \rho(\mathbf{X}_i) + \rho(\mathbf{X}_i)\sigma_R(\mathbf{X}_i)\frac{(R_i - \mu_R)}{\sigma_R(\mathbf{X}_i)}.$$

Hence,

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\epsilon_i, \tag{2.3}$$

where $\epsilon_i = (R_i - \mu_R)/\sigma_R(\mathbf{X}_i)$, $m(\mathbf{x}) = \mu_R\, \rho(\mathbf{x})$ and $\sigma(\mathbf{x}) = \rho(\mathbf{x})\sigma_R(\mathbf{x})$. Given the conditional moment restriction (2.2) on R_i we have that $E(\epsilon_i|\mathbf{X}_i) = 0$. In addition, by definition, $\mathrm{Var}(\epsilon_i|\mathbf{X}_i) = 1$. As such, $E(Y_i|\mathbf{X}_i) = m(\mathbf{X}_i)$ and $\mathrm{Var}(Y_i|\mathbf{X}_i) = \sigma^2(\mathbf{X}_i)$.

The main idea is based on the observation that $m(\mathbf{x}) \equiv \mu_R\, \rho(\mathbf{x})$. Therefore, estimating

m leads to $\hat{m}(\mathbf{x}) = \mu_R \hat{\rho}(\mathbf{x})$, since μ_R does not depend on $\mathbf{x}$. We thus get from $\hat{m}$ an estimator of ρ, but perhaps with a wrong multiplicative constant. Then, if we have an estimator for μ_R, we can propose to estimate the frontier as $\hat{\rho}(\mathbf{x}) = \hat{m}(\mathbf{x})/\hat{\mu}_R$. With this in mind, we propose to estimate ρ in two simple steps.

In the first step, we propose to estimate μ_R by

$$\hat{\mu}_R = \left(\max_{1 \leq i \leq n} \frac{Y_i}{\tilde{m}(\mathbf{X}_i)} \right)^{-1},$$

where $\tilde{m}$ is a pilot estimator of m. We will specify and discuss the choice of $\tilde{m}$ in Section 2.4. Intuitively, this estimator works because it is generally assumed that there exists at least one observed production unit in the domain that is efficient, or at least close enough to being efficient.

The second step involves a multivariate nonparametric estimator with regressand Y_i and regressor vector $\mathbf{X}_i$ for $i = 1, \ldots, n$. Any reasonable nonparametric methods could be used here. We mention some examples in the next subsection, and study their asymptotic properties in Section 3. In particular, apart from the local linear kernel method of Fan (1992), we are also interested in incorporating the commonly seen shape constraints of the frontier surfaces, such as concavity and monotonicity into our estimation procedure. Therefore, recent developed nonparametric shape-constrained methods are also considered in this context. Importantly, these methods are typically free of tuning parameters. Here we denote the resulting estimator by $\hat{m}$.

After these two steps the proposed estimator for the frontier at $\mathbf{x} \in \mathbb{R}^p$ is given by $\hat{\rho}(\mathbf{x}) = \hat{m}(\mathbf{x})/\hat{\mu}_R$. Furthermore, when it becomes evident that the function σ is not constant, then one could consider including an additional step when observations are weighted based on the estimated heteroskedasticity from the previous iteration, which might further improve the efficiency and utilize the additional information from the additive error term.

2.2.2. Robust frontier estimation (in the presence of outliers) In the presence of outliers in the data, we propose modify the previous estimation procedure as follows.

In the first step, we estimate μ_R by the sample α_n-quantile of the ratios $\left\{Y_i/\tilde{m}(\mathbf{x}_i), i = 1, \ldots, n\right\}$ and denote this estimator by $\check{\mu}_R$. Typically we would choose α_n to be close to 1 (theory dictates that $\alpha_n \to 1$ as $n \to \infty$), and shall discuss its choice together with the choice of other tuning parameters in Section 2.4. In addition, the second step remains unchanged and the frontier can be estimated by our robust frontier estimator as $\hat{\rho}_R(\mathbf{x}) = \hat{m}(\mathbf{x})/\check{\mu}_R$, as before. Finally, we remark that an optional re-estimation procedure could be performed where we first eliminate all the observations that lie above the estimated frontier (i.e. those with $Y_i > \hat{\rho}_R(\mathbf{X}_i)$) from the data and then repeat the previous estimation steps again with $\alpha_n = 1$.

Note that this new procedure works even under the scenario where there is no outlier. Some theoretical results regarding this will be presented in Section 3.

2.3. Examples of nonparametric estimators in the second step

2.3.1. Local polynomial kernel (NP2S-LP) We use the multivariate local polynomial kernel to estimate m nonparametrically. Suppose that the polynomial degree is $q \in \mathbb{N}$. Borrowing notation from Masry (1996), for any $\mathbf{x} = (x_1, \ldots, x_p)$ and $\boldsymbol{k} = (k_1, \ldots, k_p) \in \mathbb{N}^p$, let $|\boldsymbol{k}| = \sum_{j=1}^p k_j$ and $\mathbf{x}^{\boldsymbol{k}} = x_1^{k_1} \times \cdots \times x_p^{k_p}$. Write $\mathbf{h}_n \in \mathbb{R}_+^p$ as the bandwidth vector. For notational convenience, we will assume $\mathbf{h}_n = (h_n, \ldots, h_n)$ in the remaining of our manuscript.

Now for any $\mathbf{x} \in \mathbb{R}^p_+$ we obtain $\hat{m}(\mathbf{x}) \equiv \hat{m}(\mathbf{x}; h_n) \equiv \hat{b}_{(0,\ldots,0)}$ with

$$\{\hat{b}_{\boldsymbol{k}}\}_{\boldsymbol{k}:0\leq|\boldsymbol{k}|\leq q} = \arg\min_{\{b_{\boldsymbol{k}}\}_{\boldsymbol{k}:0\leq|\boldsymbol{k}|\leq q}} \sum_{i=1}^{n} \Big(Y_i - \sum_{0\leq|\boldsymbol{k}|\leq q} b_{\boldsymbol{k}}(\mathbf{X}_i - \mathbf{x})^{\boldsymbol{k}}\Big)^2 K_{h_n}(\mathbf{X}_i - \mathbf{x}),$$

where $K : \mathbb{R}^p \to \mathbb{R}$ is a kernel, with $K_h(\mathbf{u}) = (1/h)^p K(\mathbf{u}/h)$ for any $h > 0$.

When $q = 1$, the above definition becomes the local linear kernel, where $\hat{m}(\mathbf{x}) \equiv \hat{m}(\mathbf{x}; h_n) \equiv \hat{\alpha}$ for any $\mathbf{x}$ with

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - (\mathbf{X}_i - \mathbf{x})\beta^\top)^2 K_{h_n}(\mathbf{X}_i - \mathbf{x}).$$

We call the corresponding frontier estimator NP2S-LL.

2.3.2. Concave regression (NP2S-CR) If it is known a priori that the frontier is concave (which is common due to the law of diminishing returns), then we could estimate m using the least squares concave regression (Lim and Glynn (2012); Seijo and Sen (2011)). More specifically, we let

$$\hat{m} \in \arg\min_{f\in\mathcal{F}^{\mathrm{Conc}}} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i))^2,$$

where $\mathcal{F}^{\mathrm{Conc}} = \{f : \mathbb{R}^p \to \mathbb{R} | f \text{ is concave}\}$. Here we use "$\in$" because the minimiser is typically not unique (but its values are uniquely determined over $\{\mathbf{X}_i\}_{i=1}^n$). Alternatively, we could restate the problem as the following Quadratic Program (QP):

$$\begin{aligned} \text{minimize}_{\ a_1,\ldots,a_n,\mathbf{b}_1,\ldots,\mathbf{b}_n} & \sum_{i=1}^{n} (Y_i - a_i)^2 \\ \text{s.t.} \qquad & a_j + \langle \mathbf{b}_j, \mathbf{X}_i - \mathbf{X}_j \rangle \geq a_i; \quad \text{for every } i, j \in \{1, \ldots, n\} \text{ with } i \neq j, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product. We then take

$$\hat{m}(\mathbf{x}) = \min_{j=1,\ldots,n} \{a_j + \langle \mathbf{b}_j, \mathbf{x} - \mathbf{X}_j \rangle\}.$$

2.3.3. (Univariate) S-shaped regression (NP2S-SS) For $p = 1$, we now define the class of S-shaped functions as follows:

$$\mathcal{F}^{\mathrm{S}} = \Big\{ f : \mathbb{R} \to \mathbb{R} \mid f \text{ is increasing; there exists } b \in [-\infty, \infty] \text{ s.t.} \\ f \text{ is convex over } (-\infty, b], f \text{ is concave over } [b, \infty) \Big\}.$$

We remark that S-shape could be useful and preferred for the modelling of production when firms experience increasing returns to scale followed by decreasing returns to scale along their expansion paths. In addition, note that in the above definition we allow b, the inflection point of f, to be taken as $\pm\infty$ so that this class contains and generalises both the class of (increasing) convex and concave functions (i.e., when $b = \infty$ and $b = -\infty$ respectively).

Suppose that $m \in \mathcal{F}^{\mathrm{S}}$ (with unknown location of the inflection point), then we could estimate it using the least squares shape-constrained regression estimator

$$\hat{m} \in \arg\min_{f\in\mathcal{F}^{\mathrm{S}}} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i))^2.$$

See Feng et al. (2022a) for a detailed description of this estimator and its theoretical properties in depth, as well as Feng et al. (2022b) for its computation software.

2.3.4. Additive isotone regression (NP2S-AI) One way to alleviate the curse of dimensionality in the multivariate nonparametric estimation (i.e. where $p > 1$) is to impose additivity on m. See, for instance, Stone (1986). Here a function $f : \mathbb{R}^p \to \mathbb{R}$ is additive if for any $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, one can write

$$f(\mathbf{x}) = c_0 + f_1(x_1) + \cdots + f_p(x_p),$$

where $c_0 \in \mathbb{R}$ is a constant, $f_1, \ldots, f_p$ are univariate functions. For the purpose of identifiability solely, it is also assumed that $\int f_j(x_j)\nu(d\mathbf{x}) = 0$ for $j = 1, \ldots, p$ with some absolutely continuous measure ν. In the context of frontier estimation, typically we could assume that $f_1, \ldots, f_p$ are monotonically increasing. Therefore, we could estimate m via the least squares additive isotone regression estimator Mammen and Yu (2007), defined as

$$\hat{m} \in \arg\min_{f \in \mathcal{F}^{\mathrm{AddIn}}} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i))^2,$$

where

$$\mathcal{F}^{\mathrm{AddIn}} = \Big\{ f : \mathbb{R}^p \to \mathbb{R} | f(\mathbf{x}) = c_0 + f_1(x_1) + \cdots + f_p(x_p) \text{ for all } \mathbf{x} = (x_1, \ldots, x_p); \\ f_1, \ldots, f_p \text{ are increasing} \Big\}.$$

However, as pointed out by Fang et al. (2022), the additivity constraint could be inappropriate if there are interactions between inputs, or if one wants to have $\partial f / \partial x_j$ to depend on all the components of $\mathbf{x}$ rather than a single input. If that is the case, one could also consider applying variable transformation as an attempt to restore the additivity, e.g. log-transform for the output in the Cobb–Douglas model. See also Sun et al. (2011) who address the resulting biases caused by approximating log production.

Finally, we remark that if one has reasons to believe that each of $f_1, \ldots, f_p$ follow different shape constraints (e.g., some of the f_j's are increasing while the others are concave or even S-shaped), we could replace $\mathcal{F}^{\mathrm{AddIn}}$ in the above by the appropriate alternatives, and use the approach of Chen and Samworth (2016) to estimate m.

2.4. Choosing the pilot estimator and the tuning parameters

2.4.1. The pilot estimator We now specify our choice of $\tilde{m}$ for different estimators.

- For NP2S-LL, NP2S-CR and NP2S-SS, we could take $\tilde{m}$ as the local polynomial kernel estimator with $q = 2$ or $q = 3$ and the corresponding bandwidth vector $\mathbf{g}_n = (g_n, \ldots, g_n) \in \mathbb{R}_+^p$. The theoretical requirements for the kernel and the bandwidths are discussed in Section 3.
- For NP2S-AI, to take into account the additive structure, we could take $\tilde{m}$ as the estimator proposed by Horowitz and Mammen (2004).

We remark that our choices above are largely motivated by the convenience of theoretical development (and especially the asymptotic distributional theory) presented in Section 3. In practice, we would recommend taking $\tilde{m} = \hat{m}$ instead as a much more universal choice, which offers similar or sometimes better numerical performance as we experienced in our simulation experiments. We then let

$$\hat{\mu}_R^* = \left(\max_{\{i:\, \mathbf{X}_i \in \mathcal{S}\}} \frac{Y_i}{\hat{m}(\mathbf{X}_i)} \right)^{-1},$$

for the standard version of our estimator and

$$\check{\mu}_R^* = \left(Q_{\alpha_n} \left\{ \frac{Y_i}{\hat{m}(\mathbf{X}_i)} \;\middle|\; i : \mathbf{X}_i \in \mathcal{S} \right\} \right)^{-1},$$

for the robust version, where $Q_{\alpha_n}(z_1, \ldots, z_n)$ returns the α_n-quantile of given samples $\{z_1, \ldots, z_n\}$, and where $\mathcal{S}$ can be taken as any reasonably-sized compact set that is contained in the interior of the support of $\mathbf{X}$ for NP2S-CR and NP2S-AI, or $\mathbb{R}^p$ for NP2S-LP and $\mathbb{R}$ for NP2S-SS. Indeed, in practice, one might simply prefer to use a universally trimmed set $\mathcal{S}$ on which to compute the aforementioned quantities, in order to account for potentially variable or sub-optimal performance of certain nonparametric estimators, especially close to the boundaries.

2.4.2. α_n for robust frontier estimation Here we outline several different strategies for the choice of α_n.

- From a theoretical perspective, we could take some α_n satisfying $1-\alpha_n = o(n^{-4/(p+4)})$. See Section 3 for more details.
- If we have prior knowledge on the number of outliers (i.e. those with $R_i > 1$) in the dataset, denoted by n_o, then we could simply use $\alpha_n = 1 - n_o/n$.
- A more practical approach is to inspect the estimated density function of the (scaled) efficiency scores $\hat{g}_n$ based on $\{Y_i/\tilde{m}(\mathbf{X}_i)\}$ using the kernel density estimator, then find the lowest r_n^* where $\hat{g}_n(r)$ remains small (say, of $O(n^{-1/2})$) for all $r > r_n^*$, either visually or numerically, and finally take $\alpha_n = |\{i : \ Y_i/\tilde{m}(\mathbf{X}_i) \leq r_n^*\}|/n$. Here $|\cdot|$ gives the cardinality of a set.
- The approach of Fang et al. (2022) could also be used, where they identify as outliers those observations with estimated efficiency scores that fall out of the adjusted inter-quartile range, $AIQR$, which is given by $AIQR = [Q_{0.25} - 1.5IQR, Q_{0.75} + 1.5IQR]$, where $IQR = Q_{0.75} - Q_{0.25}$, with $Q_{0.25}$ and $Q_{0.75}$ being, respectively, the first and third quartile.

3. THEORY

3.1. General assumptions

To discuss the asymptotic properties of the proposed estimators, here we list our assumptions in scenarios both with and without outliers.

3.1.1. Without the presence of outliers

Assumption A1

$\{(\mathbf{X}_i, R_i)\}_{i=1}^n$ is an independent and identically distributed (i.i.d.) sequence with each distributed as $(\mathbf{X}, R)$ with density f. $f_{\mathbf{X}}(\mathbf{x})$ and $f_R(r)$ denote the common marginal densities of $\mathbf{X}$ and R respectively, and $f_{R|\mathbf{X}}(r; \mathbf{x})$ denotes the common conditional density of R given $\mathbf{X}$. Finally, denote F_R as the cumulative distribution function corresponding to f_R.

Assumption A2

1 The support of $f_{\mathbf{X}}$, denoted as Θ, is a compact and convex subset of $\times_{j=1}^p (0, \infty)$. Here $\times_{j=1}^p (0, \infty)$ denotes the Cartesian product of the intervals $(0, \infty)$. In addition, $f_{\mathbf{X}}$ is Lipschitz continuous with $\inf_{\mathbf{x}\in\Theta} f_{\mathbf{X}}(\mathbf{x}) > 0$.

2 $0 \leq R \leq 1$. Moreover, there exists some $\delta > 0$ and some $c_R > 0$ such that $\inf_{\mathbf{x}\in\Theta}, f_{R|\mathbf{X}}(r; \mathbf{x})$ is bounded from below by the linear curve $c_R(1 - r)$ over $r \in (1 - \delta, 1]$.

Assumption A3

$Y_i = \rho(\mathbf{X}_i)R_i$, for $i = 1, \ldots, n$.

Assumption A4

1 For every $\mathbf{x} \in \Theta$, $E(R|\mathbf{X} = \mathbf{x}) = \mu_R > 0$.

2 $\sigma_R^2(\mathbf{x}) = \mathrm{Var}(R|\mathbf{X} = \mathbf{x})$, which is continuous with respect to every $\mathbf{x} \in \Theta$. In addition, $\inf_{\mathbf{x}\in\Theta} \sigma_R^2(\mathbf{x}) > 0$.

Assumptions A1, A3 and A4 imply that $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ forms an independent and identically distributed (i.i.d.) sequence of random variables with some joint density. Assumption A2.2 essentially requires that the density of $(R|\mathbf{X})$ does not decay to 0 too slowly at its right boundary. We remark that it is sufficient but not necessary, and is stated in such a way for easy interpretation. This particular assumption implies that $\left|1 - (\max_{i=1,\ldots,n} R_i)^{-1}\right| = O_p(n^{-1/2})$ and $1 - \max_{i=1,\ldots,n} R_i = O_p(n^{-1/2})$, which are the actual equations we use in the proofs, and in general cannot be weaken further without affecting the speed at which information about the frontier can be gathered. This assumption is crucial because it also makes sure that the estimation error in $\hat{\mu}_R$ does not contribute to the asymptotic distribution of the frontier estimator, thus simplifies our analysis.

Comparing with Martins-Filho and Yao (2007) and Martins-Filho et al. (2013), we do not have to assume a constant $\sigma_R^2(\mathbf{x})$. Furthermore, we avoid dealing with regressands that are themselves residuals from a first stage nonparametric regression, due to the elimination of the second step in their estimation procedure. Consequently, asymptotic properties are easier to obtain. In addition, our assumptions here are weaker than requiring $\mathbf{X}$ and R to be independent.

Next, we impose a smoothness condition on the frontier function ρ. We say that a function $\rho : \Theta \to \mathbb{R}$ (with $\Theta \subset \mathbb{R}^p$) has smoothness index s and constant L if ρ is $\lfloor s \rfloor$ times differentiable and all its partial derivatives of order $\boldsymbol{\pi}$ with $|\boldsymbol{\pi}|_1 = \lfloor s \rfloor$ satisfy $|\rho^{(\boldsymbol{\pi})}(\mathbf{x}) - \rho^{(\boldsymbol{\pi})}(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|^{s - \lfloor s \rfloor}$ for all $\mathbf{x}, \mathbf{x}' \in \Theta$.

Assumption A5

The true frontier $\rho : \Theta \to \mathbb{R}$ is a s-smooth function with $s > 2$ and constant $L > 0$. Moreover, $\inf_{\mathbf{x} \in \Theta} \rho(\mathbf{x}) > 0$.

In essence, this assumption requires ρ to be at least twice-differentiable. Although here we focus on $s > 2$ in the development of our theory, many of our results also have analogous versions for other restrictions on s.

For the pilot estimator for NP2S-LP, NP2S-SS and NP2S-CR using local polynomial, we make additional requirements on the kernel and its bandwidth vector.

Assumption A6^a

1 $K(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ is a symmetric density function with bounded support $S_p \subset \mathbb{R}^p$ satisfying:
   (a) $\int \mathbf{x}^\top \mathbf{x} K(\mathbf{x}) d\mathbf{x} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix (N.B. recall that here $\mathbf{x}$ is represented in the row vector form).
   (b) $K(\mathbf{x})$ is bounded and Lipschitz continuous for all $\mathbf{x} \in S_p$.
   (c) $\inf_{\mathbf{x} \in S_p} K(\mathbf{x}) > 0$.

2 The bandwidth $\mathbf{g}_n = (g_n, \ldots, g_n)$ with $g_n \asymp n^{-\gamma}$ and $\gamma \in \Big(\big(2\min(s,3)+p\big)^{-1}, (4+p)^{-1}\Big)$.

Note that symmetry of the kernel also implies that S_p is symmetric, and $\int \mathbf{x} K(\mathbf{x}) d\mathbf{x} = \mathbf{0}$.

For the pilot estimator for NP2S-AI using the method of Horowitz and Mammen (2004), we require the assumptions therein, i.e.

Assumption A6^b

Assumptions A1 – A7 of Horowitz and Mammen (2004) hold.[3]

3.1.2. In the presence of outliers In the presence of outliers, Assumption A3 shall be replaced by the following, where $|\cdot|$ gives the cardinal number (or size) of a set.

[3]More precisely, minor modifications of the assumptions of Horowitz and Mammen (2004) are required, as they took the support of $\mathbf{X}$ to be $[-1, 1]^p$ without loss of generality, which would need to be changed to the smallest rectangle containing Θ. These assumptions are not reproduced here in the interest of space.

Assumption A3*

There exists a constant $\kappa \in (0,1)$ and a set $\mathcal{M}_n \subset \{1,2,\ldots,n\}$ with $|\mathcal{M}_n| \leq n^\kappa$. For non-outliers (i.e. $i \in \{1,2,\ldots,n\}\backslash\mathcal{M}_n$), $Y_i = \rho(\mathbf{X}_i)R_i$, whereas for the outliers, $\liminf_{n\to\infty} \min_{i\in\mathcal{M}_n} Y_i/\rho(\mathbf{X}_i) > 1$ and $\limsup_{n\to\infty} \max_{i\in\mathcal{M}_n} Y_i < \infty$.

Here we assume that the number of outliers can grow with the number of observations, n, at a certain rate to be specified later. This includes the situation where the number of outliers is zero or at a fixed number (which also covers Assumption A3). In addition, we assume that the implied efficiency scores of all the outliers, which could be either random or deterministic, are strictly bounded away from and above 1, with their output values bounded from above. These assumptions could be relaxed but would result in more complex theoretical statements, which we shall not pursue in this manuscript.

3.2. Asymptotic characterization

Now we are in the position to establish the uniform consistency and asymptotic distribution of the frontier estimator and robust frontier estimator using different estimating procedures.

3.2.1. Local polynomial kernel

The case of $q = 1$ First, we study the behaviour of the local linear kernel estimator (i.e. the degree of polynomials, $q = 1$). For the frontier estimation, we have the following result:

THEOREM 3.1. *Suppose that Assumptions A1 – A5 and A6^a hold. Let $q = 1$ and take $h_n = c_h n^{-1/(p+4)}$ for some $c_h > 0$. Then,*

$$\sup_{\mathbf{x}\in\Theta} |\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})| \xrightarrow{p} 0,$$

as $n \to \infty$. In addition, for every $\mathbf{x} \in \text{int}(\Theta)$ (i.e. the interior of Θ),

$$n^{2/(p+4)} \{\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})\} \xrightarrow{d} N\left(c_h^2\{\Delta\rho(\mathbf{x})\}/2,\ \frac{\sigma^2(\mathbf{x})}{c_h^p \mu_R^2 f_{\mathbf{X}}(\mathbf{x})} \int_{S_p} K^2(\mathbf{u}) d\mathbf{u}\right), \tag{3.4}$$

as $n \to \infty$, where $\Delta\rho(\mathbf{x}) = \sum_{j=1}^p \frac{\partial^2 \rho}{\partial x_j^2}(x_j)$ is the Laplacian of ρ at $\mathbf{x} = (x_1,\ldots,x_p)$.

For the robust frontier estimator, a similar result is given as follows:

COROLLARY 3.1. *Under the assumptions and conditions stated in Theorem 3.1 but replacing Assumption A3 by A3*. In addition, let $\kappa \in (0, 2/(p+4))$ and take α_n such that $1-\alpha_n = o(n^{-4/(p+4)})$ which also satisfies $n(1-\alpha_n) \geq n^\kappa$. Then, $\sup_{\mathbf{x}\in\Theta} |\hat{\rho}_R(\mathbf{x}) - \rho(\mathbf{x})| \xrightarrow{p} 0$ as $n \to \infty$, and for every $\mathbf{x} \in \text{int}(\Theta)$*

$$n^{2/(p+4)} \{\hat{\rho}_R(\mathbf{x}) - \rho(\mathbf{x})\} \xrightarrow{d} N\left(c_h^2\{\Delta\rho(\mathbf{x})\}/2,\ \frac{\sigma^2(\mathbf{x})}{c_h^p \mu_R^2 f_{\mathbf{X}}(\mathbf{x})} \int_{S_p} K^2(\mathbf{u}) d\mathbf{u}\right).$$

Minimax lower bounds In this part, we investigate the difficulty of the frontier estimation problem (without outliers) from a minimax perspective.

Let Ψ be the class of models satisfying Assumptions A1–A5 with a common Θ (i.e. the support of $\mathbf{X}$) and universal constants (e.g. ρ's smoothness index s and the corresponding constant L). For any $\psi \in \Psi$, let ρ^ψ be the corresponding frontier. Then we have the following result.

THEOREM 3.2. *Under Assumptions A1–A5, for any* $\mathbf{x} \in \text{int}(\Theta)$ *, we have that*

$$\inf_{\hat{\rho}} \sup_{\psi \in \Psi} E_\psi \left\{ \hat{\rho}(\mathbf{x}) - \rho^\psi(\mathbf{x}) \right\}^2 \geq C(n \log n)^{-2s/(p+2s)}$$

for some $C > 0$ *and for any sufficiently large* n*.*

Since we only restrict ourselves to $s > 2$ in our assumptions, the above rate for the mean-squared loss lower bounds is essentially $n^{-4/(4+p)}$ in the worst case (i.e. by taking s close to 2 and dropping the logarithm factor). As such, our proposed frontier using NP2S-LL can be viewed as rate-optimal estimators from a minimax perspective.

The case of $q = 2, 3$ and beyond Note that there is a bias term in Theorem 3.1 and Corollary 3.1, unless $\Delta\rho(\mathbf{x}) = 0$. This bias term could be eliminated by using local polynomial of degree $q \geq 2$ and choosing a bandwidth of smaller order. In the following, we first focus on the case of $q = 2, 3$, and then explain why using $q \geq 4$ provides no further improvement for $s \in (2, 3)$.

Borrowing additional notation from Masry (1996), we let $N_i = \binom{i+p-1}{p-1}$ be the number of distinct p-tuples $\boldsymbol{k}$ with $|\boldsymbol{k}| = i$. We arrange these p-tuples (i.e. N_i of them in total) as a sequence in a lexicographical order (e.g. with $(i, 0, \ldots, 0)$ being the first item and $(0, \ldots, 0, i)$ being the last item) and let g_i^{-1} denote this one-to-one (i.e. tuple to index) map. In addition, let

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{0,0} & \mathbf{M}_{0,1} & \mathbf{M}_{0,2} \\ \mathbf{M}_{1,0} & \mathbf{M}_{1,1} & \mathbf{M}_{1,2} \\ \mathbf{M}_{2,0} & \mathbf{M}_{2,1} & \mathbf{M}_{2,2} \end{bmatrix} \qquad \mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{0,0} & \mathbf{\Gamma}_{0,1} & \mathbf{\Gamma}_{0,2} \\ \mathbf{\Gamma}_{1,0} & \mathbf{\Gamma}_{1,1} & \mathbf{\Gamma}_{1,2} \\ \mathbf{\Gamma}_{2,0} & \mathbf{\Gamma}_{2,1} & \mathbf{\Gamma}_{2,2} \end{bmatrix}$$

where $\mathbf{M}_{i,j}$ and $\mathbf{\Gamma}_{i,j}$ are $N_i \times N_j$ dimensional matrices whose (l, m) element are $\int_{S_p} \mathbf{u}^{\{g_i(l)+g_j(m)\}} K(\mathbf{u}) d\mathbf{u}$ and $\int_{S_p} \mathbf{u}^{\{g_i(l)+g_j(m)\}} K^2(\mathbf{u}) d\mathbf{u}$, respectively.

THEOREM 3.3. *Suppose that Assumptions A1 – A5 and A6*^{a} *hold. In addition, let* $q = 2$ *or* $q = 3$ *and take* $h_n = c_h n^{-\eta}$ *for some* $c_h > 0$ *and* $\eta \in \left((6+p)^{-1}, (4+p)^{-1}\right)$ *with* $\eta > \gamma$*. Then, for every* $\mathbf{x} \in \text{int}(\Theta)$ *(i.e. the interior of* Θ*), as* $n \to \infty$*,*

$$n^{(1-\eta p)/2} \left\{ \hat{\rho}(\mathbf{x}) - \rho(\mathbf{x}) \right\} \xrightarrow{d} N\left(0, \frac{\sigma^2(\mathbf{x})}{c_h^p \mu_R^2 f_{\mathbf{X}}(\mathbf{x})} \left[\mathbf{M}^{-1} \mathbf{\Gamma} \mathbf{M}^{-1}\right]_{(1,1)}\right),$$

where $[\cdot]_{(i,j)}$ *denotes the* (i, j)*-th entry of a given matrix.*

COROLLARY 3.2. *Under the assumptions and conditions stated in Theorem 3.3 but replacing Assumption A3 by A3***. In addition, let* $\kappa \in (0, (1-\eta p)/2)$ *and take* α_n *such that* $1 - \alpha_n = o(n^{\eta p - 1})$ *and* $n(1-\alpha_n) \geq n^\kappa$*. Then, the conclusion given in Theorem 3.3 also holds for* $\hat{\rho}_R$*.*

As a special case, when $p = 1$, let $\phi_{i,j} = \int u^i K^j(u) du$, then we have that

$$\left[\mathbf{M}^{-1} \mathbf{\Gamma} \mathbf{M}^{-1}\right]_{(1,1)} = \frac{\phi_{4,1}^2 + 2\phi_{2,1}^3 \phi_{4,1} \phi_{2,2} + \phi_{4,2} \phi_{2,1}^4}{(\phi_{4,1} - \phi_{2,1}^2)^2}.$$

Picking higher order polynomials (i.e. $q > 3$) will increase the value of $\left[\mathbf{M}^{-1} \mathbf{\Gamma} \mathbf{M}^{-1}\right]_{(1,1)}$, and thus the mean squared estimation error, if we keep everything else unchanged. See also Fan and Gijbels (1996). Consequently, here we chose not to further pursue the use of higher order polynomials.

Finally, we note that with $s \in (2, 3)$ by picking the bandwidth at an appropriate order, the mean squared error of NP2S-LP (with $q \geq 1$), $E[\{\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})\}^2]$ or

$E[\{\hat{\rho}_R(\mathbf{x}) - \rho(\mathbf{x})\}^2]$, would converge at a rate close to $n^{-s/(2s+p)}$, which appears faster than that of NP2S-LL. However, this would require additional precise knowledge about the smoothness index s, which could be challenging in practice, even with univariate input. As such, we would suggest using NP2S-LL if the number of observations is moderate.

A comparison between NP2S-LL and NP3S Now we compare the asymptotic biases and variances of the estimators NP2S-LL and NP3S for the frontier estimation. Here we focus on the case of $p = 1$ without any outliers, because theoretical results from Martins-Filho and Yao (2007) are stated only for the univariate input with no outliers. Additionally, we assume that $\mathrm{Var}(R_i|X_i = x) \equiv \sigma_R^2$ (as required by NP3S).

First, we compare the ratio between two asymptotic variances (AVARs). Using Theorem 2 of Martins-Filho and Yao (2007) and the results presented in equation (3.4) we have that for the same bandwidth h_n,

$$\frac{\mathrm{AVAR}_{\mathrm{NP2S-LL}}}{\mathrm{AVAR}_{\mathrm{NP3S}}} = \frac{4\sigma_R^2}{\mu_R^2(\mu_4(x) - 1)},$$

where $\mu_4(x) = E(\epsilon_i^4|X_i = x)$ with $\epsilon_i = (R_i - \mu_R)/\sigma_R$ being the standardised noise. It is clear that the ratio above becomes smaller as μ_R increases, σ_R decreases, or the kurtosis of R_i increases. As an example, suppose that $(R_i|X_i)$ follows a symmetric $\mathrm{Beta}(\alpha, \alpha)$ distribution for some $\alpha > 0$. Then after some calculation we could derive that

$$\frac{\mathrm{AVAR}_{\mathrm{NP2S-LL}}}{\mathrm{AVAR}_{\mathrm{NP3S}}} = \frac{2\alpha + 3}{\alpha(2\alpha + 1)},$$

i.e. this ratio is smaller than 1 if $\alpha > 3/2$.

Second, we take into account the bias term and compare the ratio between the asymptotic mean squared errors (AMSEs) of two estimators at x. Here different oracle bandwidths apply to NP2S-LL and NP3S. If these bandwidths are chosen in the optimal manner (i.e. minimizing the corresponding AMSE), then it can be shown that

$$\frac{\mathrm{AMSE}_{\mathrm{NP2S-LL}}}{\mathrm{AMSE}_{\mathrm{NP3S}}} = \left\{\frac{4\sigma_R^2}{\mu_R^2(\mu_4(x) - 1)}\right\}^{4/5} \left|1 + \frac{\{\rho^{(1)}(x)\}^2}{\rho(x)\rho^{(2)}(x)}\right|^{-2/5}. \tag{3.5}$$

Hence a combination of larger μ_R, smaller σ_R and positive curvature of the frontier ρ at x tend to lead to better performance of NP2S-LL as compared to that of NP3S. Once again, we note that picking the oracle bandwidths could be quite challenging in practice, especially for NP3S, which has a more complicated estimation procedure as compared with NP2S-LL. Interestingly, our experience from numerical experiments suggests that NP2S-LL could offer better finite-sample performance than NP3S even in the settings where the ratio displayed in (3.5) would have indicated otherwise.

3.3. Other nonparametric estimators (with shape constraints)

3.3.1. Concave regression We now investigate the asymptotic properties of frontier estimation using NP2S-CR. Here we focus on the case of $p = 1$, in which we derive the asymptotic distributional theory largely based on Groeneboom et al. (2001b) and Ghosal and Sen (2017) . We then study the more general case of $p > 1$ and briefly discuss the estimation consistency.

First, we introduce the "invelope" process studied in Groeneboom et al. (2001a) and Groeneboom et al. (2001b), which is closely related to the integrated Brownian motion. It will later appear in the pointwise limiting distribution of $\hat{\rho}$. Let $X(t) = W(t) + 4t^3$, where $W(t)$ is a standard two-sided Brownian motion starting from 0 (i.e. $W(0) = 0$), and let Y be the integral of X satisfying $Y(0) = 0$. Then according to Groeneboom

et al. (2001a), there exists an almost surely uniquely defined random continuous function $H : \mathbb{R} \to \mathbb{R}$ satisfying the following:

(I) $H(t) \geq Y(t)$ for every $t \in \mathbb{R}$;
(II) H has a convex second derivative, and with probability 1, is three times differentiable at $t = 0$;
(III) $\int_R \{H(t) - Y(t)\} dH^{(3)}(t) = 0$.

Here $H^{(j)}$ represents the j-th derivative of H.

For this part only, to simplify our analysis (mainly in the characterization of asymptotic distribution), we shall use the following random design (in additional to Assumption A1), where we assume that the input random variable X is drawn from a uniform distribution over $[a, b]$.

Assumption A1$^+$.

$p = 1$ and X is drawn from a uniform distribution over $[a, b]$.

THEOREM 3.4. *Suppose Assumptions A1$^+$, A1 - A5 and A6^a hold. In addition, assume that ρ is concave with $\sup_{x\in[a,b]} \rho^{(2)}(x) < 0$. Then, for any $x \in (a, b)$, we have that*

$$n^{2/5}\Big(\hat{\rho}(x) - \rho(x)\Big) \xrightarrow{d} \left\{(b-a)^{2/5}\left(\frac{\rho^{(2)}(x)\sigma^4(x)}{24\mu_R^4}\right)^{1/5}\right\} H^{(2)}(0),$$

as $n \to \infty$, where H is the invelope process defined as before.

COROLLARY 3.3. *Under the assumptions and conditions stated in Theorem 3.4 but replacing Assumption A3 by A3*. In addition, let $\kappa \in (0, 2/5)$ and take α_n such that $1 - \alpha_n = o(n^{-4/5})$ with $n(1 - \alpha_n) \geq n^\kappa$. Then, the conclusion given in Theorem 3.4 also holds for $\hat{\rho}_R$.*

If $p > 1$, we could still show that $\hat{\rho}$ and $\hat{\rho}_R$ are consistent under their respective settings.

THEOREM 3.5. *Suppose that Assumptions A1–A5, A6^a hold and ρ is concave. Then for any $\mathbf{x} \in \text{int}(\Theta)$, $\hat{\rho}(\mathbf{x}) \xrightarrow{p} \rho(\mathbf{x})$ as $n \to \infty$. Similarly, when Assumption A3 is replaced by A3*, for any $\mathbf{x} \in \text{int}(\Theta)$, $\hat{\rho}_R(\mathbf{x}) \xrightarrow{p} \rho(\mathbf{x})$ as $n \to \infty$.*

In fact, for the purpose of establishing consistency, we are able to relax Assumption A5 to only requiring $\inf_{\mathbf{x}\in\Theta} \rho(x) > 0$ (i.e. dropping the smoothness condition on ρ). On the other hand, deriving the convergence rate and pointwise asymptotic distribution for $p > 1$ is beyond the scope of this manuscript. To our best knowledge, it is still an actively-researched open problem in the regression setting (even with i.i.d. Gaussian noise). However, we would also like to point the readers to Han and Wellner (2016) who show that a variant of the least square convex regression estimator is sub-optimal in terms of the global convergence rate when $p \geq 4$.

3.3.2. (Univariate) S-shaped regression (NP2S-SS) The asymptotic results based on the (univariate) S-shaped regression could also be derived, which turn out to be very similar to those presented in Theorem 3.4 and Corollary 3.3 (except for the frontier at precisely the inflection point). For completeness, we state the results below.

THEOREM 3.6. *Suppose Assumptions A1$^+$, A1 - A5 and A6^a hold. In addition, assume that ρ is S-shaped with a unique inflection point at $c^* \in (a, b)$ and $\rho^{(2)}(x) \neq 0$ for*

$x \in [a,b]\backslash\{c^*\}$. Then, for any $x \in (a,b)\backslash\{c^*\}$, we have that

$$n^{2/5}\Big(\hat{\rho}(x) - \rho(x)\Big) \xrightarrow{d} \left\{(b-a)^{2/5}\left(\frac{\rho^{(2)}(x)\sigma^4(x)}{24\mu_R^4}\right)^{1/5}\right\} H^{(2)}(0),$$

as $n \to \infty$, where H is the invelope process defined as before.

COROLLARY 3.4. *Under the assumptions and conditions stated in Theorem 3.6 but replacing Assumption A3 by A3*. In addition, let $\kappa \in (0, 2/5)$ and take α_n such that $1-\alpha_n = o(n^{-4/5})$ with $n(1-\alpha_n) \geq n^{\kappa}$. Then, the conclusion given in Theorem 3.6 also holds for $\hat{\rho}_R$.*

As a final remark for this part, we note that to our knowledge the asymptotic distribution at the inflection point remains an open problem in the literature, even in the fixed design regression setting.

3.3.3. Additive isotone regression We now give the asymptotic properties of frontier and robust frontier estimators using NP2S-AI.

Some additional notation is required to handle the case of $p > 1$. Let $\mathbf{X} = (X^1, \ldots, X^p)$. We denote f_{X^j} as the marginal density function of X^j for $j = 1, \ldots, p$. In addition, the conditional variance of the unstandardized error (i.e. in Equation (2.3)) given $X^j = u$, $\mathrm{Var}\{\rho(\mathbf{X})R|X^j = u\}$, is denoted by $\sigma_j^2(u)$.

Assumption A7

1 ρ is a differentiable and additive increasing function over Θ. Moreover, write $\mathbf{x} = (x_1, \ldots, x_p)$,

$$\inf_{\mathbf{x}\in\Theta} \min_{j=1,\ldots,p} \frac{\partial \rho}{\partial x_j}(\mathbf{x}) > 0,$$

2 For every $j = 1, \ldots, p$, $\sigma_j^2(x_j)$ is continuous for any x_j in the interior of the support of f_{X^j}.

THEOREM 3.7. *Suppose that Assumptions A1–A5, A7 and A6^b hold. Then, for any $\mathbf{x} \in \mathrm{int}(\Theta)$,*

$$n^{1/3}\Big(\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})\Big) \xrightarrow{d} \frac{1}{\mu_R^{2/3}} \sum_{j=1}^{p} \left\{\left(\frac{\sigma_j^2(x_j)\frac{\partial \rho}{\partial x_j}(\mathbf{x})}{2f_{X^j}(x_j)}\right)^{1/3} G_j\right\},$$

as $n \to \infty$. Here $G_1, \ldots, G_p \overset{\text{i.i.d.}}{\sim} G$, where G is the distribution of the slope of the greatest convex minorant of $W(t) + t^2$ at $t = 0$, with W being a two-sided standard Brownian motion.

COROLLARY 3.5. *Under the assumptions and conditions stated in Theorem 3.7 but replacing Assumption A3 by A3*. In addition, let $\kappa \in (0, 1/3)$ and use α_n such that $1-\alpha_n = o(n^{-2/3})$ and $n(1-\alpha_n) \geq n^{\kappa}$. Then, the conclusion given in Theorem 3.7 also holds for $\hat{\rho}_R$.*

Here as a special case, when $p = 1$, the result from Theorem 3.7 can be simplified to

$$n^{1/3}\Big(\hat{\rho}(x) - \rho(x)\Big) \xrightarrow{d} \left(\frac{\sigma^2(x)\rho^{(1)}(x)}{2\mu_R^2 f_X(x)}\right)^{1/3} G,$$

as $n \to \infty$, where σ and ρ are defined as previously, and where G is still the distribution of the slope of the greatest convex minorant of $W(t) + t^2$ at $t = 0$, with W being a two-sided standard Brownian motion.

Note that in Theorem 3.7, Assumption 5 can be weakened to only requiring the the true frontier to have smoothness index of $s = 1$. Consequently, when $p = 1$, the convergence rate of $n^{-1/3}$ stated here is slower than $n^{-2/5}$ given in Theorem 3.1, as less restrictive smoothness assumption is required using tools based on isotone regression. In fact, the canonical (additive) isotone regression estimator would appear piecewise constant, and thus discontinuous; in practice, if additional smoothness assumption is warranted, one could apply post-smoothing methods, such as Mammen (1991), to further improve its finite-sample performance.

It is also worth noting that even with multiple inputs, both $|\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})|$ and $|\hat{\rho}_R(\mathbf{x}) - \rho(\mathbf{x})|$ are of $O_p(n^{-1/3})$, which is free of p, the dimension of $\mathbf{X}$. In other words, the assumption of additivity allows us to effectively circumvent the curse of dimensionality. See also Stone (1986).

4. NUMERICAL EXPERIMENTS

4.1. Simulation Study

4.1.1. Settings We conduct a simulation study to evaluate the finite-sample properties of the estimators discussed in Sections 2 and 3. We consider the following data generating processes (DGPs) with either univariate or bivariate inputs:

(i) $p = 1$. Random samples are generated from $Y_i = \rho(X_i)R_i$ for $i = 1, \ldots, n$ with
$$\rho(x) = \sqrt{x}.$$
$\{(X_i, R_i)\}_{i=1}^n$ are i.i.d. pairs with $X_i \sim U_{[0,1]}$ and independent of $R_i \sim Beta(a, b)$. Note that here ρ is concave and increasing.

(ii) $p = 2$. Random samples are generated from $Y_i = \rho(X_{i1}, X_{i2})R_i$ for $i = 1, \ldots, n$ with
$$\rho(x_1, x_2) = \sqrt{x_1} + \sqrt{x_2}.$$
$\{(\mathbf{X}_i, R_i)\}_{i=1}^n$ are i.i.d. pairs with $\mathbf{X}_i = (X_{i1}, X_{i2}) \sim U_{[0,1]\times[0,1]}$ and independent of $R_i \sim Beta(a, b)$. Here ρ is additive with concave and increasing individual components.

(iii) $p = 2$. As in the second DGP, random samples are generated from $Y_i = \rho(X_{i1}, X_{i2})R_i$ for $i = 1, \ldots, n$. Nevertheless, here the frontier is non-additive, assuming the following Cobb–Douglas-type frontier function
$$\rho(x_1, x_2) = \sqrt{x_1} \times \sqrt{x_2}.$$
$\{(\mathbf{X}_i, R_i)\}_{i=1}^n$ are again i.i.d. pairs with $\mathbf{X}_i = (X_{i1}, X_{i2}) \sim U_{[0,1]\times[0,1]}$ and independent of $R_i \sim Beta(a, b)$.

Two sample sizes $n = 200, 400$ are used as well as two different choices of the pair (a, b) for the Beta distribution: $(2, 2)$ and $(4, 2)$.

To examine the performance of our robust estimators, we also consider the same settings with outliers for DGPs (i) and (ii), where we randomly select 1% of the previously generated observations and increase their corresponding efficiency scores to 1.5.

Each experiment involves 500 Monte Carlo replicates. Here the variance of R is taken as constant (by assuming R and $\mathbf{X}$ to be independent) in order to facilitate fair comparison between different estimators.

4.1.2. Estimators We compare our proposal to `NP3S` (Nonparametric in 3 Steps by Martins-Filho and Yao (2007)) and `NPE` (Nonparametric Exponential by Martins-Filho et al. (2013)) estimators, as well as to the three estimators by Fang et al. (2022): `UQS` (Unconstrained Quantile Spline), `MCQS` (Monotone Constrained Quantile Spline) and `MCCQS` (Monotone and Concave Constrained Quantile Spline). Hereafter we use the following abbreviations for the proposed variations of our 2-step estimators:

- `LL` for NP2S with local linear estimator;
- `AI` for NP2S with additive increasing regression;
- `ACR` for NP2S with additive concave regression;
- `CNA` for NP2S with concave non-additive regression with multi-dimensional input.

When $p = 1$, in NP2S-LL, we use the bandwidth with value $h_n = c_h n^{-1/5}$ for $c_h = 1.5\sqrt{\widehat{\text{Var}}(\{X_i\}_{i=1}^n)}$. We have also experimented other selection rules such as those in Ruppert et al. (1995) and Fan and Gijbels (1995), but do not see much improvement. In addition, as advocated in Section 2.4, we bypass the pilot estimator in our simulation and take $\hat{\mu}_R = \left(\max_{\{i:\, X_i \in \mathcal{S}\}} Y_i/\hat{m}(X_i)\right)^{-1}$ with $\mathcal{S} = \mathbb{R}_+$ for LL, and $\mathcal{S} = \left[Q_{0.1}(\{X_i\}_{i=1}^n), Q_{0.9}(\{X_i\}_{i=1}^n)\right]$ for all the shape-constraint based approaches, where recalling that $Q_\alpha(\{z_i\}_{i=1}^n)$ is the sample α-quantile based on $\{z_1, \ldots, z_n\}$. When $p = 2$, similar implementation for the bandwidth of LL is used, and we take $\mathcal{S} = \mathbb{R}_+^2$ for LL and $\mathcal{S} = \left[Q_{0.1}(\{X_{i1}\}_{i=1}^n), Q_{0.9}(\{X_{i1}\}_{i=1}^n)\right] \times \left[Q_{0.1}(\{X_{i2}\}_{i=1}^n), Q_{0.9}(\{X_{i2}\}_{i=1}^n)\right]$ for the other approaches.

Finally, in terms of the robust version of our procedure, for sake of comparison, in our numerical experiments, we pick α_n using the outlier detection scheme proposed by Fang et al. (2022), where they identify as outliers those observations of $Y_i/\hat{m}(X_i)$ that fall out of the adjusted inter-quartile range, $AIQR$, which is given by $AIQR = [Q_{0.25} - 1.5IQR, Q_{0.75} + 1.5IQR]$. Here $IQR = Q_{0.75} - Q_{0.25}$ and $Q_{0.25}$ and $Q_{0.75}$ are, respectively, the first and third quartile of the estimated efficiency scores.

4.1.3. Results For performance comparison, we report the distributions of the rooted mean squared errors (RMSE) of all the estimators, given as $n^{-1}\sum_{i=1}^n\{\tilde{\rho}(\mathbf{X}_i) - \rho(\mathbf{X}_i)\}^2$ (where $\tilde{\rho}$ represents the respective estimators) via the boxplots in Figures 1 to 4 for DGPs (i), (ii) and (iii) with $Beta(2,2)$ and $Beta(4,2)$ distributions.

In Figure 1 we illustrate that, regardless the sample size (200 or 400) and the dimension (univariate or bivariate), the errors keep the same relative structures among the different estimators. Hence, for brevity, in the next figure we only report results based on the sample size 200. We can also see that, for these concave and additive concave frontier shapes, our proposal, under its several variants, produces better results when compared to the five benchmark approaches. The `ACR` has the best overall performance.

Figure 2 reports results when outliers are intentionally included in the data, either for the univariate or bivariate case (DGPs (i) and (ii)). We use this plot to show the effect that outliers can cause on the frontier estimators. Going from the left panel plots, where non-robust estimators are used, to the right panel plots, where the robust versions are implemented, we see a very sharp reduction in the errors, making it clear how relevant the robust versions are. Furthermore, we can noticeably see that all our estimators present homogeneously superior or at least similar performance compared to the benchmarks.

On Figure 3, we explore what the effect of outliers and the effect of increasing the number of outliers (along with the sample size) are on the frontier estimators. For the univariate case, in the top two rows of Figure 3, our proposals are superior, especially `LL` and `ACR`. Panels in the bottom two rows of Figure 3 show a bit more balanced situation, although our proposals still perform slightly better than, or at least comparable to the benchmarks.

Finally, Figure 4 shows the results when a Cobb–Douglas function is used as the DGP, that is, our DGP (iii). We only illustrate the non-outliers case. Here we see that, in terms of the median of the errors, `CNA` performs best, followed by `NP3S`, `NPE` and `LL`, which present similar overall results. These findings are in line with our expectation, as in this case, the true frontier is concave but not additive, so the model is misspecified for `ACR` and those by Fang et al. (2022). The better performance from `CNA` confirms the

usefulness of shape-constrained approach over kernel-based methods when the model is correctly or nearly-correctly specified.

4.2. An empirical application

In our empirical application we use the *Milk producers* dataset Bogetoft and Otto (2020), available in the R package *Benchmarking with DEA and SFA*. The data set is composed by $n = 108$ observations, and here we estimate the production frontier using *output of milk per cow* as the response variable, and *energy expenses per cow* and *veterinary expenses per cow* as our two covariates. Here we use the additive concave (i.e. ACR) model for our estimation. The additive structure facilitates interpretation, while the (increasing and) concave constraint is used to reflect the plausible diseconomies of scale in milk production.

Figure 5 shows the two estimated additive components using the ACR model for three different α_n-levels, 1, 0.99 and 0.98. Here the concave shapes of the curves are reasonably robust against different choices of α_n-levels for the individual estimated components. That is, assuming having 0%, 1% or 2% of the outliers does not change substantially the contributions either from *energy expenses per cow* or from *veterinary expenses per cow* to estimate the production frontier measured in terms of *output of milk per cow*.

In Figure 6 we present the estimated frontier surfaces, also via the ACR model, for the three different α_n-levels, 1, 0.99 and 0.98. Similar conclusions to those from Figure 5 can be drawn here. There are little changes in the overall shape of the surfaces over the different choices of α_n, which is concave but highly non-linear. The yellow points, which are the only ones above the estimated frontier surface, are those associated with the outliers. One of those on the top right side of the figure can be clearly seen, whereas two can be identified at the bottom part. As we decrease the value of α_n, the estimated frontier surface becomes lower.

Finally, we remark that the outlier detection criterion of Fang et al. (2022) is also implemented, but with no outliers detected, so the estimated frontier remains the same as in the case with $\alpha_n = 1$.

5. CONCLUSION

In this paper we propose a robust methodology for estimating production frontiers with multi-dimensional input via a two-step nonparametric regression. We present a frontier estimation framework under a variety of flexible models which can both accommodate commonly seen shape constraints of the frontier surfaces and be robust to the presence of outliers. Monte Carlo simulations and empirical data analysis illustrate the good performance and finite properties of our proposed methods. Possible future extensions include, but are not limited to, testing strategies for different shape constraints, automatic selection of those constraints, testing the constant mean (or variance) assumption on the efficiency scores, handling of a large number of predictors in the high-dimensional setting, and handling outliers in the inputs.

ACKNOWLEDGEMENT

We thank the editor and two anonymous referees for their comments and suggestions that helped us improve the paper. Flavio A. Ziegelmann was partially supported by CNPq under Grants 310165/2018-0 and 438642/2018-0.

SUPPLEMENTARY MATERIALS

Proofs Proofs of the theoretical results from the main manuscript.

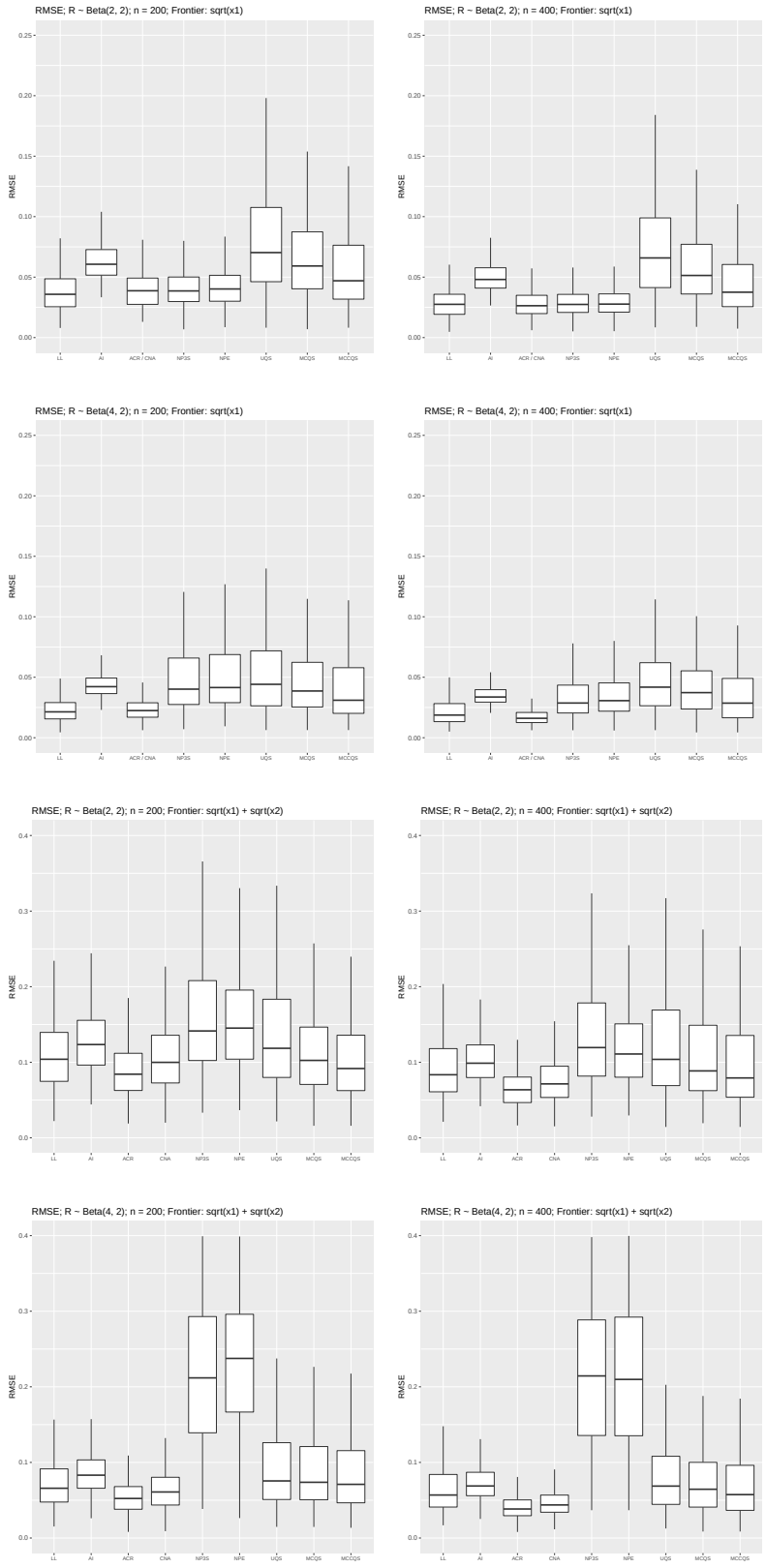

Figure 1. Boxplots of the rooted mean square errors for different estimators under different DGPs (with no outliers) as specified in the title of each panel. Larger sample size is used in all the panels on the right.

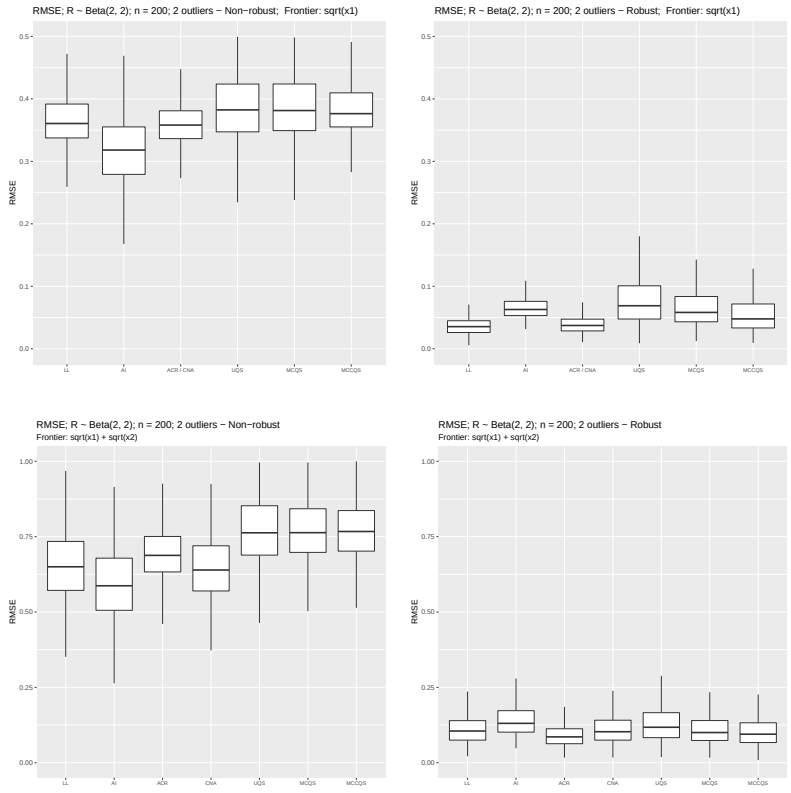

Figure 2. Boxplots of the rooted mean square error for different estimators different DGPs (with outliers) as specified in the title of each panel. In the panels on the left, the existence of outliers has not been taken into account in the (standard) estimation procedure, while in the panels on the right, the robust estimators are used.

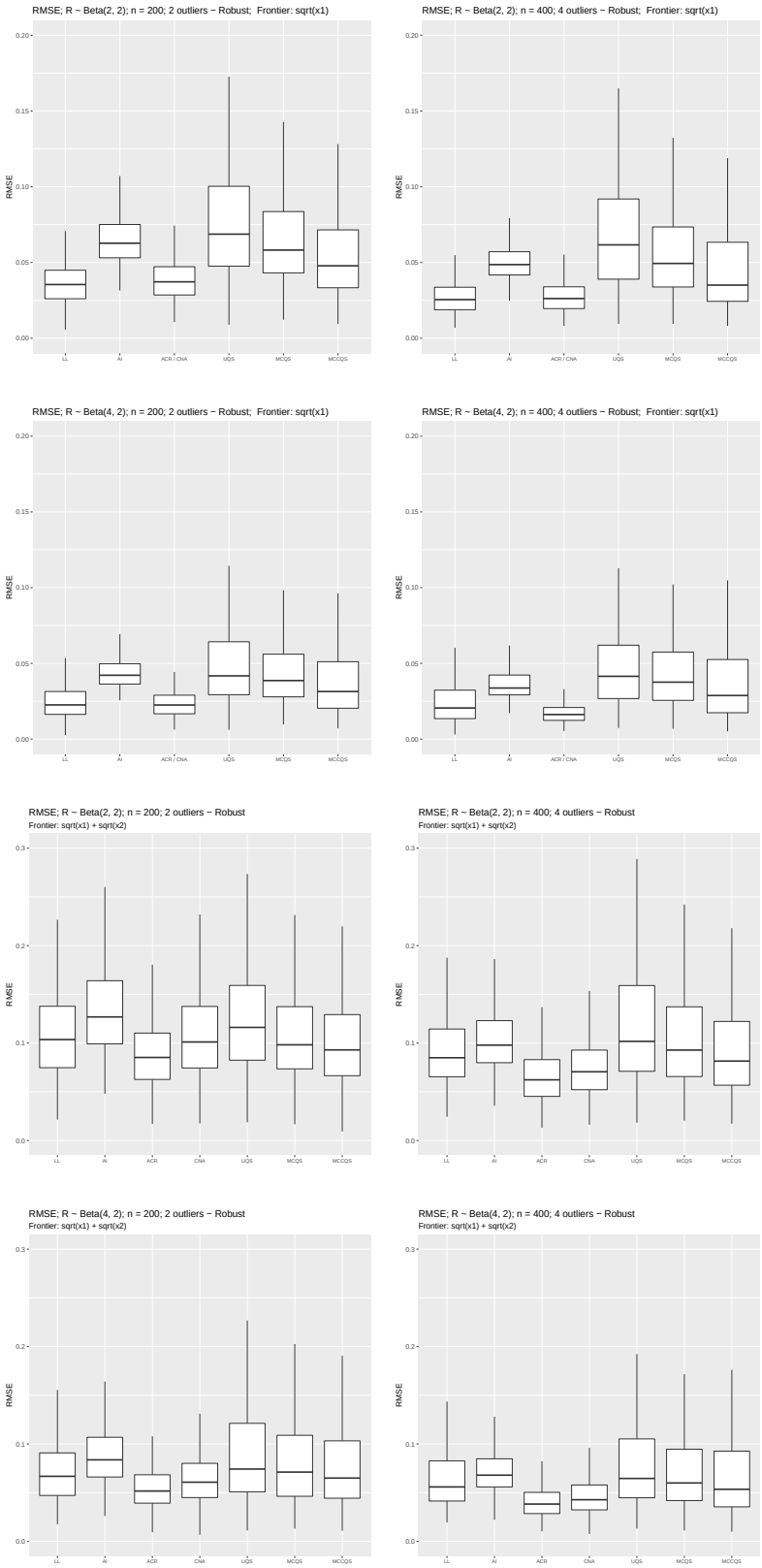

Figure 3. Boxplots of the rooted mean square error for different estimators different DGPs (with outliers) as specified in the title of each panel. The number of outliers is also increased in all the panels on the right.

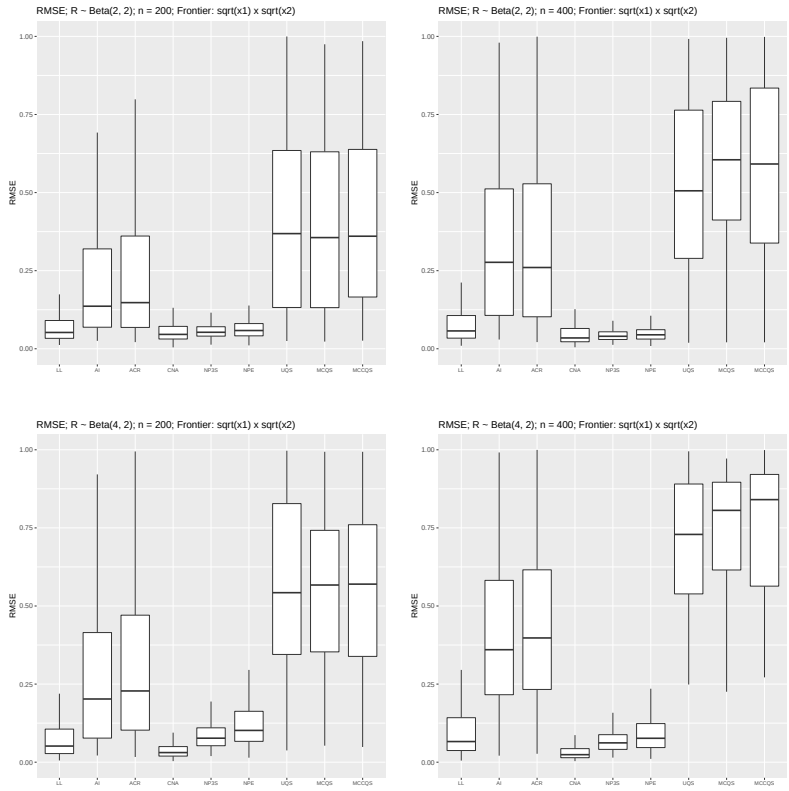

Figure 4. Boxplots of the rooted mean square error for different estimators different DGPs (Cobb–Douglas with no outliers) as specified in the title of each panel. A larger number of observations (and outliers) in the panels on the right.

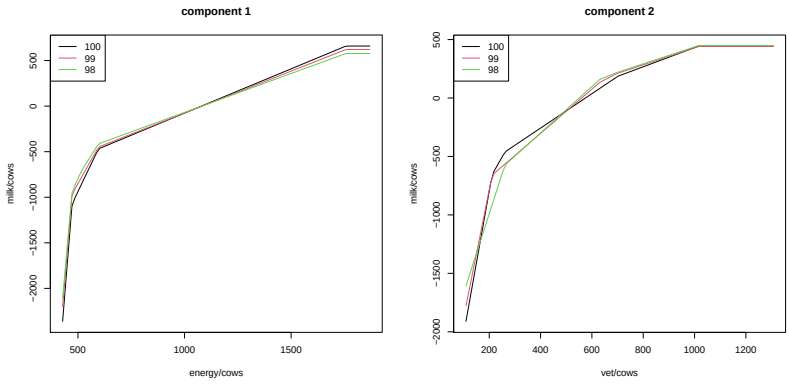

Figure 5. *Milk producers* dataset. Estimated additive components of the frontier using the ACR model for three different α_n-levels, 100%, 99% and 98%. Note that the above plots do not reflect the actual levels of the estimated frontiers because the constants of the additive functions are not included here.

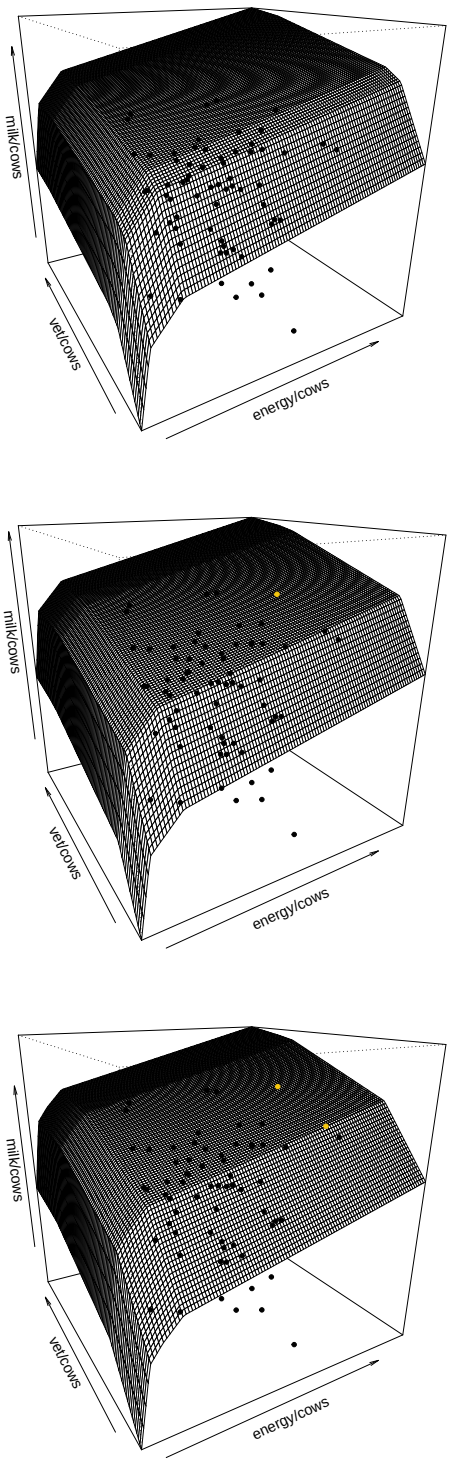

Figure 6. *Milk producers* dataset. Dispersion 3d-plots and estimated frontier surfaces using the ACR model for three different α-levels, 1, 0.99 and 0.98 (from top to bottom). Outliers, which are the only points above the estimated curve, are printed in yellow.

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.

Badunenko, O., D. J. Henderson, and S. C. Kumbhakar (2012). When, where and how to perform efficiency estimation. *Journal of Royal Statistical Society Series A 175*, 863–892.

Bogetoft, P. and L. Otto (2020). *Benchmarking with DEA and SFA*. R package version 0.29.

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

Chen, Y. and R. J. Samworth (2016). Generalized additive and index models with shape constraints. *Journal of Royal Statistical Society Series B 78*, 729–754.

Daraio, C. and L. Simar (2005). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis 24*, 93–121.

Daraio, C. and L. Simar (2007). *Advanced robust and nonparametric methods in efficiency analysis: methodology and applications*. Springer.

Deprins, D., L. Simar, and H. Tulkens (1984). Measuring labor inefficiency in post offices. In *The performance of public enterprises: concepts and measurements*, pp. 243–267.

Fan, J. (1992). Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association 87*, 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 Series B 57*, 371–394.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Now Publishers Inc.

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

Fang, Y., L. Xue, C. Martins-Filho, and L. Yang (2022). Robust estimation of additive boundaries with quantile regression and shape constraints. *Journal of Business Economics and Statistics 40*, 615–628.

Feng, O. Y., Y. Chen, Q. Han, R. J. Carroll, and R. J. Samworth (2022a). Nonparametric, tuning-free estimation of s-shaped functions. *Journal of Royal Statistical Society Series B 84*, 1324–1352.

Feng, O. Y., Y. Chen, Q. Han, R. J. Carroll, and R. J. Samworth (2022b). *Sshaped: nonparametric, tuning-Free estimation of S-Shaped functions*. R package version 1.1.

Ghosal, P. and B. Sen (2017). On univariate convex regression. *Sankhya : The Indian Journal of Statistics 79-A*, 215–253.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001a). A canonical process for estimation of convex functions: the "Invelope" of integrated Brownian motion $+t^4$. *The Annals of Statistics 29*, 1620–1652.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001b). Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics 29*, 1653–1698.

Han, Q. and J. A. Wellner (2016). Multivariate convex regression: global risk bounds and adaptation.

Horowitz, J. L. and E. Mammen (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics 32*, 2412–2443.

Johnson, A. L. and L. F. McGinnis (2008). Outlier detection in two-stage semiparametric DEA models. *European Journal of Operational Research 187*, 629–635.

Khezrimotlagha, D., W. D. Cook, and J. Zhuc (2008). A nonparametric framework to detect outliers in estimating production frontiers. *European Journal of Operational Research 286*, 375–388.

Kneip, A., L. Simar, and P. Wilson (2015). When bias kills the variance: central limit theorems for DEA and FDH efficiency scores. *Econometric Theory 31*, 4394–422.

Lim, E. and P. W. Glynn (2012). Consistency of multidimensional convex regression. *Operations Research 60*, 196–208.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics 19*, 724–740.

Mammen, E. and K. Yu (2007). Additive isotone regression. In *IMS Lecture Notes Monograph Series, Asymptotics: Particles, Processes and Inverse Problems*, Volume 55, pp. 179–195. Institute of Mathematical Statistics.

Martins-Filho, C., H. Torrent, and F. A. Ziegelmann (2013). Local exponential frontier estimation. *Brazilian Review of Econometrics 141*, 283–319.

Martins-Filho, C. and F. Yao (2007). Nonparametric frontier estimation via local linear regression. *Journal of Econometrics 141*, 283–319.

Masry, E. (1996). Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes and their Applications 65*, 81–101.

Meeusen, W. and J. van Den Broeck (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review 18*, 435–444.

Papadopoulos, A. and C. F. Parmeter (2022). Quantile Methods for Stochastic Frontier Analysis. *Foundations and Trends in Econometrics 12*, 1–120.

Parmeter, C. F. and S. C. Kumbhakar (2014). Efficiency analysis: A primer on recent advances. *Foundations and Trends® in Econometrics 7*(3–4), 191–385.

Parmeter, C. F. and J. S. Racine (2013). Smooth constrained frontier analysis. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, pp. 463–488. Springer.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*, 1257–1270.

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

Seijo, E. and B. Sen (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics 39*, 1633–1657.

Simar, L. (2003). Detecting Outliers in Frontier Models: A Simple Approach. *Journal of Productivity Analysis 20*, 391–424.

Simar, L. and P. Wilson (2013). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. In *Foundations and Trends in Econometrics*, Volume 5, pp. 411–435. Now.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics 14*, 590–606.

Sun, K., D. J. Henderson, and S. C. Kumbhakar (2011). Biases in approximating log production. *Journal of Applied Econometrics 26*(4), 708–714.

Wang, L. and L. Yang (2020). Estimation of additive frontier functions with shape constraints. *Journal of Nonparametric Statistics 32*, 262–293.

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

Supplementary Materials for Robust Nonparametric Frontier Estimation in Two Steps

YINING CHEN[†], HUDSON S. TORRENT[‡] AND FLAVIO A. ZIEGELMANN[‡]

[†] *Department of Statistics, London School of Economics and Political Science*
[‡] *Department of Statistics, Federal University of Rio Grande do Sul*

S1. PROOFS

Proof of Theorem 3.1:

First, we focus on the uniform consistency and asymptotic distribution of $\hat{m}(\mathbf{x})$. Since Θ is bounded, $\sup_{\mathbf{x}\in\Theta}|\rho(\mathbf{x})| < \infty$. As such, $\sup_{\mathbf{x}\in\Theta}\sigma(\mathbf{x}) < \infty$. Moreover, in its representation in (2.3), the noise ϵ_i has bounded q-moment for any $q \in \mathbb{N}$. Based on our assumptions, it follows from Proposition 7 of Fan and Guerre (2016) that

$$\sup_{\mathbf{x}\in\Theta}|\hat{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{p} 0.$$

as $n \to \infty$. This also implies with arbitarily high probability, $\sup_{\mathbf{x}\in\Theta}|\hat{m}(\mathbf{x})|$ is bounded above by a constant (that does not depend on n). Moreover, for any $\mathbf{x} \in \text{int}(\Theta)$, a careful application of Theorem 4 of Masry (1996) entails that

$$\sqrt{nh_n^p}\Big[\{\hat{m}(\mathbf{x}) - m(\mathbf{x})\} - h_n^2\{\Delta m(\mathbf{x})\}/2\Big] \xrightarrow{d} N\Big(0, \int_{S_p} K^2(\mathbf{u})d\mathbf{u}\ \sigma^2(\mathbf{x})/f_{\mathbf{X}}(\mathbf{x})\Big)$$

as $n \to \infty$. After rearranging the terms, we get that

$$n^{2/(p+4)}\left\{\hat{m}(\mathbf{x}) - m(\mathbf{x})\right\} \xrightarrow{d} N\left(c_h^2\{\Delta m(\mathbf{x})\}/2,\ \frac{\sigma^2(\mathbf{x})}{c_h^p f_{\mathbf{X}}(\mathbf{x})}\int_{S_p} K^2(\mathbf{u})d\mathbf{u}\right).$$

Second, we establish the convergence rate of $|\hat{\mu}_R - \mu_R|$. Note that by Proposition 7 of Fan and Guerre (2016) again, we have that for the pilot estimator $\sup_{\mathbf{x}\in\Theta}|\tilde{m}(\mathbf{x}) - m(\mathbf{x})| = o_p(n^{-2/(p+4)})$. Since $m(\mathbf{x})$ is strictly bounded away from zero over Θ, this also implies that

$$\sup_{\mathbf{x}\in\Theta}\left|\frac{\tilde{m}(\mathbf{x})}{m(\mathbf{x})} - 1\right| = o_p(n^{-2/(p+4)}) \qquad \text{and} \qquad \sup_{\mathbf{x}\in\Theta}\left|\frac{m(\mathbf{x})}{\tilde{m}(\mathbf{x})} - 1\right| = o_p(n^{-2/(p+4)}).$$

In addition, our Assumption A2.2 entails that $1 - \max_{i=1,\ldots,n} R_i = O_p(n^{-1/2})$, as $n \to \infty$. On the one hand,

$$\begin{aligned}\hat{\mu}_R &= \left(\max_{1\le i\le n}\frac{Y_i}{\tilde{m}(\mathbf{X}_i)}\right)^{-1} = \left(\max_{1\le i\le n}\frac{R_i\rho(\mathbf{X}_i)}{\tilde{m}(\mathbf{X}_i)}\right)^{-1} = \left(\max_{1\le i\le n}\frac{R_i m(\mathbf{X}_i)}{\mu_R\tilde{m}(\mathbf{X}_i)}\right)^{-1} \\ &\ge \mu_R \min_{\mathbf{x}\in\Theta}\frac{\tilde{m}(\mathbf{x})}{m(\mathbf{x})} \ge \mu_R\big\{1 - o_p(n^{-2/(p+4)})\big\}.\end{aligned}$$

On the other hand,

$$\hat{\mu}_R = \left(\max_{1\le i\le n} \frac{Y_i}{\tilde{m}(\mathbf{X}_i)}\right)^{-1} = \left(\max_{1\le i\le n} \frac{R_i\, m(\mathbf{X}_i)}{\mu_R\, \tilde{m}(\mathbf{X}_i)}\right)^{-1}$$
$$\le \mu_R\Big(\max_{1\le i\le n} R_i\Big)^{-1} \frac{\tilde{m}(\mathbf{X}_{i^*})}{m(\mathbf{X}_{i^*})} \le \mu_R\Big\{1 + O_p(n^{-1/2}) + o_p(n^{-2/(p+4)})\Big\},$$

where $i^* = \text{argmax}_{1\le i\le n} R_i$. Consequently, $|\hat{\mu}_R - \mu_R| = o_p(n^{-2/(p+4)})$. Here we have also used Assumption A2.2 to establish the fact that $\left|1 - \big(\max_{1\le i\le n} R_i\big)^{-1}\right| = O_p(n^{-1/2})$.

Third, to show the uniform consistency of $\hat{\rho}$, we note that

$$\sup_{\mathbf{x}\in\Theta} |\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})| = \sup_{\mathbf{x}\in\Theta} \left|\frac{\hat{m}(\mathbf{x})}{\hat{\mu}_R} - \frac{m(\mathbf{x})}{\mu_R}\right| \le \sup_{\mathbf{x}\in\Theta} \left|\frac{\hat{m}(\mathbf{x})}{\mu_R} - \frac{m(\mathbf{x})}{\mu_R}\right| + \sup_{\mathbf{x}\in\Theta} \left|\hat{m}(\mathbf{x}) \frac{(\hat{\mu}_R - \mu_R)}{\hat{\mu}_R \mu_R}\right| \xrightarrow{p} 0,$$

as $n \to \infty$. Similarly, we also have that

$$\begin{aligned} n^{2/(p+4)}\big(\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})\big) &= n^{2/(p+4)}\Big(\frac{\hat{m}(\mathbf{x})}{\hat{\mu}_R} - \frac{m(\mathbf{x})}{\mu_R}\Big) \\ &= \frac{n^{2/(p+4)}\big(\hat{m}(\mathbf{x}) - m(\mathbf{x})\big)}{\mu_R} + \frac{\hat{m}(\mathbf{x})}{\hat{\mu}_R \mu_R}\big\{n^{2/(p+4)}(\hat{\mu}_R - \mu_R)\big\} \\ &= \frac{n^{2/(p+4)}\big(\hat{m}(\mathbf{x}) - m(\mathbf{x})\big)}{\mu_R} + o_p(1) \\ &\xrightarrow{d} N\left(c_h^2\{\Delta\rho(\mathbf{x})\}/2,\ \frac{\sigma^2(\mathbf{x})}{c_h^p \mu_R^2 f_{\mathbf{X}}(\mathbf{x})} \int_{S_p} K^2(\mathbf{u}) d\mathbf{u}\right), \end{aligned}$$

where we also used the fact that $\Delta\rho(\mathbf{x}) = \Delta m(\mathbf{x})/\mu_R$. □

Proof of Corollary 3.1:

First, we verify that as $n \to \infty$, the asymptotic characterizations of $\hat{m}(\mathbf{x})$ and $\tilde{m}(\mathbf{x})$ stated in the proof of Theorem 3.1 remain unchanged. For the sake of brevity, below we show this for $\hat{m}(\mathbf{x})$. Let $\hat{m}(\mathbf{x})$ be the estimator based on the observations $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ with $|\mathcal{M}_n|$ outliers, and let $\hat{m}_U(\mathbf{x})$ be the estimator based on the same set of observations but with all $|\mathcal{M}_n|$ outliers replaced by the non-outliers in the output (generated by the original DGP with i.i.d. observations using the same input values). Abusing the notation in this part of the proof only, we denote the corresponding non-outliers in the output as Y_i^o for $i \in \mathcal{M}$. By the linearity of the local linear estimator, it is easy to see that for any $\mathbf{x}$, $\hat{m}(\mathbf{x}) - \hat{m}_U(\mathbf{x})$ equals the first element of

$$A_n := \left(\sum_{i\in\mathcal{M}_n} (1, \mathbf{X}_i - \mathbf{x}) K_{h_n}(\mathbf{X}_i - \mathbf{x})(Y_i - Y_i^o)\right)\left(\sum_{i=1}^n (1, \mathbf{X}_i - \mathbf{x})^\top (1, \mathbf{X}_i - \mathbf{x}) K_{h_n}(\mathbf{X}_i - \mathbf{x})\right)^{-1}.$$

Now rewrite $A_n := B_n \times C_n$, where

$$B_n := \frac{1}{nh_n^p}\left(\sum_{i\in\mathcal{M}_n}(1,\mathbf{X}_i-\mathbf{x})K_{h_n}\left(\mathbf{X}_i-\mathbf{x}\right)\left(Y_i-Y_i^o\right)\right),$$

$$C_n := \left(\frac{1}{nh_n^p}\sum_{i=1}^{n}(1,\mathbf{X}_i-\mathbf{x})^\top(1,\mathbf{X}_i-\mathbf{x})K_{h_n}\left(\mathbf{X}_i-\mathbf{x}\right)\right)^{-1}.$$

By Lemma 5 of Fan and Guerre (2016), we have that the largest eigenvalue of C_n is bounded from above (i.e. of $O_p(1)$). In addition, as $|\mathcal{M}_n| \leq n^\kappa$, and with the values of outliers bounded (so $\sup_{i\in\mathcal{M}_n}|Y_i^o - Y_i| < \infty$), we conclude that $|B_n|$ is of $O_p(\frac{n^\kappa}{nh_n^p})$, which is $o_p(n^{-2/(p+4)})$ based on our assumptions. Consequently, A_n is also $o_p(n^{-2/(p+4)})$, and therefore

$$n^{2/(p+4)}(\hat{m}(\mathbf{x}) - \hat{m}_U(\mathbf{x})) = o_p(1).$$

Next, recall that $\check{\mu}_R$ is the sample α_n-quantile of the ratios $\left\{Y_i/\tilde{m}(\mathbf{x}_i), i = 1,\ldots,n\right\}$ with $1-\alpha_n = o(n^{-4/(4+p)})$ and $n(1-\alpha_n) \geq |\mathcal{M}_n|$. Since for all $i \in \{1,\ldots,n\}\backslash\mathcal{M}_n$

$$\frac{Y_i}{\tilde{m}(\mathbf{x}_i)} = \frac{R_i\rho(\mathbf{x}_i)}{\tilde{m}(\mathbf{x}_i)} = R_i\mu_R\frac{m(\mathbf{x}_i)}{\tilde{m}(\mathbf{x}_i)} = R_i\mu_R(1+o_p(n^{-2/(p+4)})),$$

we have that

$$\check{\mu} = Q_{\alpha_n}\left(\left\{\frac{Y_i}{\tilde{m}(\mathbf{x}_i)}\middle| i = 1\ldots,n\right\}\right) = \mu_R Q_{\alpha_n}(\{R_i|i=1\ldots,n\})(1+o_p(n^{-2/(p+4)})),$$

where $Q_\alpha(\cdot)$ returns the sample α_n-quantile of a given set, and where abusing the notation slightly, we continue to define $R_i = Y_i/\rho(\mathbf{x}_i)$ even for the outliers (i.e. this would lead to $R_i > 1$ for $i \in \mathcal{M}_n$). Therefore,

$$\begin{aligned}\check{\mu} &\leq \mu_R Q_{1-\mathcal{M}_n/n}(\{R_i|i=1\ldots,n\})(1+o_p(n^{-2/(p+4)}))\\ &= \mu_R Q_1(\{R_i|i\in\{1\ldots,n\}\backslash\mathcal{M}_n\})(1+o_p(n^{-2/(p+4)}))\\ &= \mu_R(1+O_p(n^{-1/2}))(1+o_p(n^{-2/(p+4)})) = \mu_R(1+o_p(n^{-2/(p+4)})).\end{aligned}$$

On the other hand, due to Assumption A2.2, theory of empirical quantile dictates that

$$Q_{\alpha_n}(\{R_i|i=1\ldots,n\}) \geq Q_{\alpha_n}(\{R_i|i\in\{1\ldots,n\}\backslash\mathcal{M}_n\}) = 1-O_p((1-\alpha_n)^{1/2}) = 1-o_p(n^{-2/(p+4)}).$$

Thus,

$$\check{\mu} \geq \mu_R(1-o_p(n^{-2/(p+4)}))(1+o_p(n^{-2/(p+4)})) = \mu_R(1+o_p(n^{-2/(p+4)}).$$

Putting things together, we have that $|\check{\mu}_R - \mu_R| = o_p(n^{-2/(p+4)})$.

Since $\hat{\rho}_R(\mathbf{x}) = \hat{m}(\mathbf{x})/\check{\mu}_R$, the remaining of the proof regarding the consistency and asymptotic distribution of $\hat{\rho}_R$ can be established via an application of the Slutsky's theorem as shown in the proof of Theorem 3.1, thus is omitted here.

□

Proof of Theorem 3.2:

The proof is based on Le Cam's method. See Yu (1997) for an overview.

Without loss of generality, here we assume that $\Theta = [1/2, 3/2]^p$ and focus on the estimator at $\mathbf{x}_0 = (1,\ldots,1)$. For our purpose, it suffices to consider two greatly simplified

model candidates in Ψ, denoted by $\mathcal{P}_1$ and $\mathcal{P}_2$. In both models, $\mathbf{X}$ is independent of R, $\mathbf{X}$ is uniformly distributed, and R has the triangular density function $f_R(r) = 2-4|1/2-r|$ for $r \in [0,1]$ and 0 elsewhere. However, two models are different in the frontier at $\mathbf{x}_0$ by a positive δ. More specifically, in $\mathcal{P}_1$, $\rho^{\mathcal{P}_1}(\mathbf{x}) \equiv \rho^{\mathcal{P}_1}(x_1, \ldots, x_p) = x_1 + \ldots + x_p$, while in $\mathcal{P}_2$,

$$\rho^{\mathcal{P}_2}(\mathbf{x}) = \rho^{\mathcal{P}_1}(\mathbf{x}) + \big(\delta^{1/s} - \delta^{-1/s}\|\mathbf{x}-\mathbf{x}_0\|_2^2\big)_+^s,$$

where $(z)_+ = \max(z, 0)$. One could verify that both $\rho^{\mathcal{P}_1}$ and $\rho^{\mathcal{P}_2}$ are s-smooth. Moreover, in these models, the ratio between the frontier $\rho(\cdot)$ and the mean function $m(\cdot)$ is always 2 over the entire Θ.

We now provide an upper bound on the Total Variation (TV) distance between $\mathcal{P}_1^n$ and $\mathcal{P}_2^n$. Note that

$$\|\mathcal{P}_1^n - \mathcal{P}_2^n\|_{\mathrm{TV}}^2 \le 2 - 2\{1 - d_{\mathrm{hel}}(\mathcal{P}_1, \mathcal{P}_2)^2\}^n,$$

where d_{hel} is the Hellinger distance. To bound $d_{\mathrm{hel}}(\mathcal{P}_1, \mathcal{P}_2)$ from below, we focus on $\mathbf{X} \in \mathcal{S}^*$, where $\mathcal{S}^* = \{\mathbf{x} : \|\mathbf{x}-\mathbf{x}_0\|_2^2 \le \delta^{2/s}\}$, as $\rho^{\mathcal{P}_1}(\mathbf{x}) = \rho^{\mathcal{P}_2}(\mathbf{x})$ for all $\mathbf{x} \notin \mathcal{S}^*$. It follows from some algebraic manipulations that

$$\begin{aligned}
d_{\mathrm{hel}}(\mathcal{P}_1, \mathcal{P}_2)^2 &= \int_{\mathcal{S}^*} \int_0^{\rho^{\mathcal{P}_2}(\mathbf{x})} \Big(\sqrt{f_R(y/\rho^{\mathcal{P}_2}(\mathbf{x}))/\rho^{\mathcal{P}_2}(\mathbf{x})} - \sqrt{f_R(y/\rho^{\mathcal{P}_1}(\mathbf{x}))/\rho^{\mathcal{P}_1}(\mathbf{x})}\Big)^2 dy f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&\le \int_{\mathcal{S}^*} \int_0^{\rho^{\mathcal{P}_1}(\mathbf{x})/2} \Big(\frac{2}{\rho^{\mathcal{P}_1}(\mathbf{x})} - \frac{2}{\rho^{\mathcal{P}_2}(\mathbf{x})}\Big)^2 y dy d\mathbf{x} \\
&\quad + \int_{\mathcal{S}^*} \int_{\rho^{\mathcal{P}_1}(\mathbf{x})/2}^{\rho^{\mathcal{P}_1}(\mathbf{x})} \frac{16\{\rho^{\mathcal{P}_2}(\mathbf{x}) - \rho^{\mathcal{P}_1}(\mathbf{x})\}^2}{\{\rho^{\mathcal{P}_1}(\mathbf{x})\}^4} \frac{1}{f_R(y/\rho^{\mathcal{P}_2}(\mathbf{x}))/\rho^{\mathcal{P}_2}(\mathbf{x})} dy d\mathbf{x} \\
&\quad + \int_{\mathcal{S}^*} \int_{\rho^{\mathcal{P}_1}(\mathbf{x})}^{\rho^{\mathcal{P}_2}(\mathbf{x})} \frac{4}{\{\rho^{\mathcal{P}_2}(\mathbf{x})\}^2} (\rho^{\mathcal{P}_2}(\mathbf{x}) - y) dy d\mathbf{x} \\
&\le \mathrm{Vol}(S^*)\delta^2 \Big(\inf_{\boldsymbol{x} \in \mathcal{S}^*} \rho^{\mathcal{P}_1}(\mathbf{x})\Big)^{-2}/2 \\
&\quad + 64\mathrm{Vol}(S^*)\delta^2 \Big(\inf_{\boldsymbol{x} \in \mathcal{S}^*} \rho^{\mathcal{P}_1}(\mathbf{x})\Big)^{-4} \Big(\inf_{\boldsymbol{x} \in \mathcal{S}^*} \rho^{\mathcal{P}_2}(\mathbf{x})\Big)^{-2} \log\left\{\frac{\Big(\sup_{\boldsymbol{x} \in \mathcal{S}^*} \rho^{\mathcal{P}_2}(\mathbf{x})\Big)^2}{2\delta \inf_{\boldsymbol{x} \in \mathcal{S}^*} \rho^{\mathcal{P}_1}(\mathbf{x})}\right\} \\
&\quad + 2\mathrm{Vol}(S^*)\delta^2 \Big(\inf_{\boldsymbol{x} \in \mathcal{S}^*} \rho^{\mathcal{P}_2}(\mathbf{x})\Big)^{-2}
\end{aligned}$$

for sufficiently small δ. Here we used the fact that by construction

$$\sqrt{f_R(y/\rho^{\mathcal{P}_2}(\mathbf{x}))/\rho^{\mathcal{P}_2}(\mathbf{x})} - \sqrt{f_R(y/\rho^{\mathcal{P}_1}(\mathbf{x}))/\rho^{\mathcal{P}_1}(\mathbf{x})} = \begin{cases} \Big(\frac{2}{\rho^{\mathcal{P}_1}(\mathbf{x})} - \frac{2}{\rho^{\mathcal{P}_2}(\mathbf{x})}\Big)\sqrt{y}, & \text{for } y \in [0, \rho^{\mathcal{P}_1}(\mathbf{x})/2]; \\ \frac{2}{\rho^{\mathcal{P}_2}(\mathbf{x})}\sqrt{\rho^{\mathcal{P}_2}(\mathbf{x}) - y}, & \text{for } y \in [\rho^{\mathcal{P}_1}(\mathbf{x}), \rho^{\mathcal{P}_2}(\mathbf{x})]; \end{cases}$$

and

$$\begin{aligned}
\sup_{0 \le y \le \rho^{\mathcal{P}_2}(\mathbf{x})} \Big|f_R(y/\rho^{\mathcal{P}_2}(\mathbf{x}))/\rho^{\mathcal{P}_2}(\mathbf{x}) - f_R(y/\rho^{\mathcal{P}_1}(\mathbf{x}))/\rho^{\mathcal{P}_1}(\mathbf{x})\Big| &= \Big(\frac{4}{\{\rho^{\mathcal{P}_1}(\mathbf{x})\}^2} - \frac{4}{\{\rho^{\mathcal{P}_2}(\mathbf{x})\}^2}\Big)\frac{\rho^{\mathcal{P}_1}(\mathbf{x})}{2} \\
&\le \frac{4}{\{\rho^{\mathcal{P}_1}(\mathbf{x})\}^2}(\rho^{\mathcal{P}_2}(\mathbf{x}) - \rho^{\mathcal{P}_1}(\mathbf{x})),
\end{aligned}$$

and applied $|\sqrt{b} - \sqrt{a}|^2 \le |b-a|^2/a$ for $a, b > 0$.

Since the volume of a p-dimensional ball with radius r is $\frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)}r^p$, by taking $r=\delta^{1/s}$, we have that

$$d_{\text{hel}}(\mathcal{P}_1,\mathcal{P}_2)^2 \leq C'\delta^{2+p/s}\log\left(\frac{1}{\delta}\right)$$

for any sufficiently small $\delta>0$ and for some constant $C'>0$. Consequently, by picking $\delta=(n\log n/c')^{-s/(2s+p)}$ for some $c'\in(0,\log(4/3)/C']$, we see that

$$\begin{aligned}
2-2\{1-d_{\text{hel}}(\mathcal{P}_1,\mathcal{P}_2)^2\}^n &\leq 2-2\left\{1-\frac{C'c'}{n\log n}\frac{s}{2s+p}(\log n+\log\log n-\log c')\right\}^n \\
&\leq 2-2\left\{1-\frac{C'c'}{n\log n}\frac{1}{2}2\log n\right\}^n \\
&\to 2-2e^{-C'c'}\leq 1/2.
\end{aligned}$$

Consequently, for all sufficiently large n, $\|\mathcal{P}_1^n-\mathcal{P}_2^n\|_{\text{TV}}^2\leq 1/2$.

Finally, it follows from Le Cam's two-point method that for all sufficiently large n,

$$\begin{aligned}
\inf_{\hat{\rho}}\sup_{\psi\in\Psi}E_\psi\big\{\hat{\rho}(\mathbf{x}_0)-\rho^\psi(\mathbf{x}_0)\big\}^2 &\geq \frac{|\rho^{\mathcal{P}_1}(\mathbf{x}_0)-\rho^{\mathcal{P}_2}(\mathbf{x}_0)|^2}{8}(1-\|\mathcal{P}_1^n-\mathcal{P}_2^n\|_{\text{TV}}) \\
&\geq \frac{(n\log n/c')^{-2s/(2s+p)}}{8}\big(1-\sqrt{1/2}\big)\equiv C_1(n\log n)^{-2s/(2s+p)}.
\end{aligned}$$

□

Proof of Theorem 3.3:

The proof of Theorem 3.3 is similar to that of Theorem 3.1. Here we focus on the main differences. Following Masry (1996) and Gu et al. (2015), we have that

$$\sqrt{nh_n^p}\Big[\{\hat{m}(\mathbf{x})-m(\mathbf{x})\}-\text{Bias}(h_n)\Big]\xrightarrow{d}N\left(0,\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\Big[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}\Big]_{(1,1)}\right)$$

Here the bias term is of order $O_p(h_n^{\min(s,3)})$. See also Theorem 1 and Proposition 3 of Fan and Guerre (2016). This means that $\sqrt{nh_n^p}\,\text{Bias}(h_n)=O_p(n^{1/2-\eta p/2-\eta\min(s,3)})=o_p(1)$, because $\eta>(p+2s)^{-1}$ and $\eta>(p+6)^{-1}$. Consequently, we have that

$$\sqrt{nh_n^p}\Big[\{\hat{m}(\mathbf{x})-m(\mathbf{x})\}\Big]\xrightarrow{d}N\left(0,\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\Big[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}\Big]_{(1,1)}\right)$$

Moreover, Proposition 7 of Fan and Guerre (2016) implies that $\sup_{\mathbf{x}\in\Theta}|\tilde{m}(\mathbf{x})-m(\mathbf{x})|=o_p(n^{(\eta p-1)/2})$, and thus $|\hat{\mu}_R-\mu_R|=o_p(n^{(\eta p-1)/2})$ (using the argument in the second stage of proof of Theorem 3.1).The final result can then be derived by using the Slutsky's theorem together with simple algebraic manipulations.

□

Proof of Theorem 3.4:

First, it follows from Theorem 3.1 of Seijo and Sen (2011) that for any $\delta>0$,

$$\sup_{x\in[a+\delta,b-\delta]}|\hat{m}(x)-m(x)|\xrightarrow{p}0,$$

as $n \to \infty$. Note that this could also be derived by following Theorem 3.3 of Groeneboom et al. (2001), but with a small modification to deal with the fact that here $\sigma(x)$ is not constant (so $\{\sigma(X_i)\epsilon_i\}_i$ in Equation (2.3) are not identically distributed).

Second, we derive the asymptotic distribution of $\hat{m}(x)$. Note that since the (unstandardized) errors $\{\sigma(X_i)\epsilon_i\}_{i=1}^n$ in Equation (2.3) are not identically distributed, Theorem 4.2 of Ghosal and Sen (2017) could not be directly applied (N.B. see also Theorem 6.3 of Groeneboom et al. (2001) for the fixed design regression setting). However, under our settings, the errors are still independent with bounded variance (both from above and from below). Moreover, $\sigma(x)$ is continuous in a neighbourhood of x (i.e. locally "close to" a constant function). Therefore, it can be checked that the conclusion of Theorem 4.2 of Ghosal and Sen (2017) remains valid, with minimum modification needed in their proof. More precisely, by temporarily requiring $b - a = 1$ (and with $x \in [a, b]$ fixed), the process $Y_n^{\mathrm{loc}}(t)$ defined on Page 236 of Ghosal and Sen (2017) (or Pages 1692–1694 of Groeneboom et al. (2001)) still converges to $\sigma(x)\int_0^t W(s)ds + m^{(2)}(x)t^4/24$. Now by scaling, we have that

$$n^{2/5}\{\hat{m}(x) - m(x)\} \xrightarrow{d} \Big(\frac{\sigma^4(x)m^{(2)}(x)(b-a)^2}{24}\Big)^{1/5} H^{(2)}(0)$$

Next, using the argument presented in the proof of Theorem 3.1, together with the fact that $p = 1$, we see that $\hat{\mu}_R - \mu_R = o_p(n^{-2/5})$.

It now follows that

$$\begin{aligned} n^{2/5}\big(\hat{\rho}(x) - \rho(x)\big) &= n^{2/5}\Big(\frac{\hat{m}(x)}{\hat{\mu}_R} - \frac{m(x)}{\mu_R}\Big) = \frac{n^{2/5}\big(\hat{m}(x) - m(x)\big)}{\mu_R} + \frac{\hat{m}(x)}{\hat{\mu}_R\mu_R}\big\{n^{2/5}(\hat{\mu}_R - \mu_R)\big\} \\ &= \frac{n^{2/5}\big(\hat{m}(x) - m(x)\big)}{\mu_R} + o_p(1), \end{aligned}$$

where we also used the fact that $|\hat{m}(x)|$ is bounded in probability for any fixed $x \in (a, b)$.

The proof is then completed via an application of the Slutsky's theorem and by noting that $m^{(2)} = \mu_R\rho^{(2)}$. □

Proof of Theorem 3.5:

In this random design, it follows from Theorem 1 of Lim and Glynn (2012) that for any $\Theta^* \subset \mathrm{int}(\Theta)$,

$$\sup_{\mathbf{x}\in\Theta^*} |\hat{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{p} 0,$$

In addition, similar to the proof of Theorem 3.1, we derive that $\hat{\mu}_R \xrightarrow{p} \mu_R$. Consequently, the Slutsky's theorem entails that $\hat{\rho}(\mathbf{x}) \xrightarrow{p} \rho(\mathbf{x})$.

□

Proof of Theorem 3.6: The proof of this result is very similar to that of Theorem 3.4. For the sake of brevity, we only highlight the differences here.

To start off with, one could show that

$$\sup_{x\in[a+\delta]\cup[c^*-\delta,c^*+\delta]\cup[b-\delta]} |\hat{m}(x) - m(x)| \xrightarrow{p} 0,$$

as $n \to \infty$, where c^* is the unique inflection point. This could be derived by following Proposition 2 of Feng et al. (2022), but with a small modification to deal with the fact that here $\sigma(x)$ is not constant (so $\{\sigma(X_i)\epsilon_i\}_i$ in Equation (2.3) are not identically distributed). Afterwards, with regard to the asymptotic distribution of $\hat{m}(x) - m(x)$, for any $x < c^*$ one could view the problem locally as a convex regression problem, while for any $x > c^*$ one could view the problem locally as a concave regression problem. To give more details, note that the assumption of non-zero $m''(x)$ except at $x = c^*$ implies that $m(x)$ has no constant (or even linear) part over $[a, b]$, so the monotonicity constraint would never be active locally in the interior of the open intervals (a, c^*) and (c^*, b). The final steps are the same as those in the proof of Theorem 3.1, which involves showing $\hat{\mu}_R - \mu_R = o_p(n^{-2/5})$ and applying Slutsky's theorem.

Proof of Theorem 3.7:

First, we study the asymptotic behaviour of $\hat{m}$ at $\mathbf{x} = (x_1, \ldots, x_p)$. In particular, our aim is to show that

$$n^{1/3}\Big(\hat{m}(\mathbf{x}) - m(\mathbf{x})\Big) \xrightarrow{d} \sum_{j=1}^{p} \left\{ \left(\frac{\sigma_j^2(x_j)\frac{\partial m}{\partial x_j}(\mathbf{x})}{2f_{\mathbf{X}^j}(x_j)} \right)^{1/3} G_j \right\}.$$

Both m and $\hat{m}$ are additive, i.e. $m(\mathbf{x}) = \sum_{j=1}^{p} m_j(x_j) + c$ and $\hat{m}(\mathbf{x}) = \sum_{j=1}^{p} \hat{m}_j(x_j) + \hat{c}$. Without loss of generality, we assume that both satisfy the identifiability condition

$$\int m_1(z_1) f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} = \cdots \int m_p(z_p) f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} = \int \hat{m}_1(z_1) f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} = \cdots = \int \hat{m}_p(z_p) f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} = 0,$$

where $\mathbf{z} = (z_1, \ldots, z_p)$. Denote by $\hat{m}_j^{\mathrm{OR}}$ the oracle estimator of m_j given that all other m_j $(j \neq i)$ are known. More precisely, the pair $(\hat{m}_j^{\mathrm{OR}}, \hat{c}_j^{\mathrm{OR}})$ is the minimizer of

$$\sum_{i=1}^{n} \Big(Y_i - \sum_{l \neq j} m_l(\mathbf{X}_{il}) - \mu_j(\mathbf{X}_{ij}) - b \Big)^2$$

with respect to a monotone increasing function μ_j that fulfill the identifiability condition of $\int \mu_j(z_j) f_{\mathbf{X}}(\mathbf{z}) d\mathbf{z} = 0$ and a constant b. Here $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$ for $i = 1, \ldots, n$. Also note that

$$\sum_{i=1}^{n} \Big(Y_i - \sum_{l \neq j} m_l(\mathbf{X}_{il}) - \mu_j(\mathbf{X}_{ij}) - b \Big)^2 = \sum_{i=1}^{n} \Big(m_j(\mathbf{X}_{ij}) + c + \sigma(\mathbf{X}_i)\epsilon_i - \mu_j(\mathbf{X}_{ij}) - b \Big)^2.$$

Since $(\hat{m}_j^{\mathrm{OR}}, \hat{c}_j^{\mathrm{OR}})$ is the minimizer, we must have that the sum of all the observed Y_i's equals the sum of the fitted ones, which amounts to

$$\sum_{i=1}^{n} \Big(m_j(\mathbf{X}_{ij}) + c + \sigma(\mathbf{X}_i)\epsilon_i \Big) = \sum_{i=1}^{n} \Big(\hat{m}_j^{\mathrm{OR}}(\mathbf{X}_{ij}) + \hat{c}_j^{\mathrm{OR}} \Big).$$

Therefore,

$$\big|\hat{c}_j^{\mathrm{OR}} - c\big| = \left| \frac{1}{n} \sum_{i=1}^{n} \Big(m_j(\mathbf{X}_{ij}) - \hat{m}_j^{\mathrm{OR}}(\mathbf{X}_{ij}) + \sigma(\mathbf{X}_i)\epsilon_i \Big) \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \sigma(\mathbf{X}_i)\epsilon_i \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \Big(m_j(\mathbf{X}_{ij}) - \hat{m}_j^{\mathrm{OR}}(\mathbf{X}_{ij}) \Big) \right|.$$

Central limit theorem implies that $\frac{1}{n}\sum_{i=1}^{n} \sigma(\mathbf{X}_i)\epsilon_i = O_p(n^{-1/2})$. Now for any $j =$

$1,\dots,p$, let $\hat{F}_{\mathbf{X}^j}$ be the empirical distribution of $\{\mathbf{X}_{ij}\}_{i=1}^n$ and $F_{\mathbf{X}^j}$ be the cumulative distribution function with respect to the marginal density function $f_{\mathbf{X}^j}$, then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{m}_j^{\mathrm{OR}}(\mathbf{X}_{ij})-m_j(\mathbf{X}_{ij})\Big) &= \int_{z_j}\big(\hat{m}_j^{\mathrm{OR}}(z_j)-m_j(z_j)\big)d\hat{F}_{\mathbf{X}^j}(z_j)\\ &= \int_{z_j}\big(\hat{m}_j^{\mathrm{OR}}(z_j)-m_j(z_j)\big)f_{\mathbf{X}^j}(z_j)dz_j\\ &\quad+\int_{z_j}\big(\hat{m}_j^{\mathrm{OR}}(z_j)-m_j(z_j)\big)d(\hat{F}_{\mathbf{X}^j}-F_{\mathbf{X}^j})(z_j)\\ &= \mathrm{(I)}+\mathrm{(II)}.\end{aligned}$$

Here (I) = 0 by the identifiability condition of the additive model. For (II), note that $\hat{m}_j^{\mathrm{OR}}(\cdot)-m_j(\cdot)$ is the difference between two monotone functions with range at most of $O(1)$ because $\{\epsilon_i\}_i$ has bounded support. In addition, it is defined on the support of $F_{\mathbf{X}^j}$, which is again bounded. Thus by standard empirical process theory (see e.g. van der Vaart and Wellner (1996)), we have that (II) $= O_p(n^{-1/2})$. Consequently, $\hat{c}_j^{\mathrm{OR}}-c=o_p(n^{-1/3})$.

Following the well-known properties of the isotone least squares estimator (specifically, the max-min representation and its localization; see for example Robertson et al. (1988)), we have with probability tending to 1 as $n\to\infty$ that

$$\begin{aligned}\hat{m}_j^{\mathrm{OR}}(x_j)+(\hat{c}_j^{\mathrm{OR}}-c) &= \max_{u\le x_j}\min_{x_j\le v}\frac{\sum_{i:u\le\mathbf{X}_{ij}\le v}\{m_j(\mathbf{X}_{ij})+\sigma(\mathbf{X}_i)\epsilon_i\}}{\#\{i:u\le\mathbf{X}_{ij}\le v\}}\\ &= \max_{x_j-e_n\le u\le x_j}\min_{x_j\le v\le x_j+e_n}\frac{\sum_{i:u\le\mathbf{X}_{ij}\le v}\{m_j(\mathbf{X}_{ij})+\sigma(\mathbf{X}_i)\epsilon_i\}}{\#\{i:u\le\mathbf{X}_{ij}\le v\}}\end{aligned}$$

where $e_n=n^{-\beta}$ for any $\beta\in(0,1/3)$, and where $\#\{\cdot\}$ denotes the number of elements of a set. Furthermore, we let $d_n=n^{-\alpha}$ with $\alpha\in(1/3,4/9)$ and define

$$\hat{m}_j^{\mathrm{OR}-}(x_j)=\max_{x_j-e_n\le u\le x_j-d_n}\min_{x_j\le v\le x_j+e_n}\frac{\sum_{i:u\le\mathbf{X}_{ij}\le v}\{m_j(\mathbf{X}_{ij})+\sigma(\mathbf{X}_i)\epsilon_i\}}{\#\{i:u\le\mathbf{X}_{ij}\le v\}}.$$

It follows that $|\hat{m}_j^{\mathrm{OR}-}(x_j)-\hat{m}_j^{\mathrm{OR}}(x_j)|=o_p(n^{-1/3})$. In essence, this means that the max-min characterization in $\hat{m}_j^{\mathrm{OR}}(x_j)$ could be well approximated using u,v outside $[x_j-d_n,x_j]$. Also note that we dropped the term $(\hat{c}_j^{\mathrm{OR}}-c)$ in the definition of $\hat{m}_j^{\mathrm{OR}-}$ as it turns out to be negligible (i.e. of order $o_p(n^{-1/3})$). For simplicity, we shall just take $\alpha=7/18$ and $\beta=7/24$ in the remaining of the proof.

Observe that $\hat{m}_1^{\mathrm{OR}}(x_1),\dots,\hat{m}_p^{\mathrm{OR}}(x_p)$ are not jointly independent, so our next task is to modify the oracle estimators further to make them independent. To do this, define

$$\mathcal{S}_{\mathbf{x},j}=\Big\{i\Big|\mathbf{X}_{ij}\in[x_j-e_n,x_j+e_n];\mathbf{X}_{il}\notin[x_l-e_n,x_l+e_n]\text{ for every }l\in\{1,\dots,p\}\backslash\{j\}\Big\}.$$

It is easy to check that $\mathcal{S}_{\mathbf{x},1},\dots,\mathcal{S}_{\mathbf{x},p}$ are disjoint and independent sets. Then define

$$\hat{m}_j^{\mathrm{OR}*}(x_j)=\max_{x_j-e_n\le u\le x_j-d_n}\min_{x_j\le v\le x_j+e_n}\frac{\sum_{\{i\in\mathcal{S}_{\mathbf{x},j}|u\le\mathbf{X}_{ij}\le v\}}\{m_j(\mathbf{X}_{ij})+\sigma(\mathbf{X}_i)\epsilon_i\}}{\#\{i\in\mathcal{S}_{\mathbf{x},j}|u\le\mathbf{X}_{ij}\le v\}}.$$

It follows that $\hat{m}_1^{\mathrm{OR}*}(x_1),\dots,\hat{m}_p^{\mathrm{OR}*}(x_p)$ are independent.

For each $j=1,\dots,p$, denote by $\mathcal{D}_{\mathbf{x},j}(u,v)=\{i:u\le\mathbf{X}_{ij}\le v,i\notin\mathcal{S}_{\mathbf{x},j}\}$. Note that for $x_j-e_n\le u$ and $v\le x_j+e_n$, $\#\{\mathcal{D}_{\mathbf{x},j}(u,v)\}=O_p(e_n^2)$. This bound is tight in a sense that

there exists some $\underline{c} > 0$ such that $P\big(\#\{\mathcal{D}_{\mathbf{x},j}(u,v)\} \geq \underline{c} e_n^2\big) \to 1$ as $n \to \infty$. Therefore,

$$
\begin{aligned}
|\hat{m}_j^{\mathrm{OR}*}(x_j) - \hat{m}_j^{\mathrm{OR}-}(x_j)| \leq & \sup_{\substack{x_j - e_n \leq u \leq x_j - d_n \\ x_j \leq v \leq x_j + e_n}} \frac{\#\{\mathcal{D}_{\mathbf{x},j}(u,v)\}}{\#\{i \in \mathcal{S}_{\mathbf{x},j} | u \leq \mathbf{X}_{ij} \leq v\}} \\
& \times \sup_{\substack{x_j - e_n \leq u \leq x_j - d_n \\ x_j \leq v \leq x_j + e_n}} \frac{\sum_{i:u \leq \mathbf{X}_{ij} \leq v} \{m_j(\mathbf{X}_{ij}) - m_j(x_j) + \sigma(\mathbf{X}_i)\epsilon_i\}}{\#\{i : u \leq \mathbf{X}_{ij} \leq v\}} \\
& + \sup_{\substack{x_j - e_n \leq u \leq x_j - d_n \\ x_j \leq v \leq x_j + e_n}} \frac{\sum_{i \in \mathcal{D}_{\mathbf{x},j}(u,v)} \{m_j(\mathbf{X}_{ij}) - m_j(x_j) + \sigma(\mathbf{X}_i)\epsilon_i\}}{\#\{i \in \mathcal{S}_{\mathbf{x},j} | u \leq \mathbf{X}_{ij} \leq v\}} \\
\leq & \, O_p\Big(\frac{e_n^2 n}{d_n n}\Big) O_p\Big(\frac{\max(e_n^2 n, (e_n n)^{1/2})}{d_n n}\Big) + O_p\Big(\frac{\max(e_n^3 n, (e_n^2 n)^{1/2})}{d_n n}\Big) = o_p(n^{-1/3}).
\end{aligned}
$$

Consequently, $|\hat{m}_j^{\mathrm{OR}*}(x_j) - \hat{m}_j^{\mathrm{OR}}(x_j)| = o_p(n^{-1/3})$, so $n^{1/3}\big(\hat{m}_j^{\mathrm{OR}*}(x_j) - m_j(x_j)\big) \xrightarrow{d} \left(\frac{\sigma_j^2(x_j)\frac{\partial m}{\partial x_j}(\mathbf{x})}{2 f_{\mathbf{X}^j}(x_j)}\right)^{1/3} G$, where G is the distribution of the slope of the greatest convex minorant of $W(t)+t^2$ at $t = 0$. Moreover, by the independence among $\hat{m}_1^{\mathrm{OR}*}(x_1), \ldots, \hat{m}_p^{\mathrm{OR}*}(x_p)$, we have that

$$
n^{1/3}\Big(\hat{m}_1^{\mathrm{OR}*}(x_1) - m_1(x_1), \ldots, \hat{m}_p^{\mathrm{OR}*}(x_p) - m_j(x_p)\Big) \xrightarrow{d} \left(\left(\frac{\sigma_1^2(x_1)\frac{\partial m}{\partial x_1}(\mathbf{x})}{2 f_{\mathbf{X}^1}(x_1)}\right)^{1/3} G_1 \ldots, \left(\frac{\sigma_p^2(x_p)\frac{\partial m}{\partial x_p}(\mathbf{x})}{2 f_{\mathbf{X}^p}(x_p)}\right)^{1/3} G_p\right),
$$

where $G_1, \ldots, G_p \overset{\text{i.i.d.}}{\sim} G$.

Theorem 1 of Mammen and Yu (2007) states that $|\hat{m}_j(x_j) - \hat{m}_j^{\mathrm{OR}}(x_j)| = o_p(n^{-1/3})$ for $j = 1, \ldots, p$. This, combined with the last displayed equation, as well as the fact that $|\hat{m}_j^{\mathrm{OR}*}(x_j) - \hat{m}_j^{\mathrm{OR}}(x_j)| = o_p(n^{-1/3})$, entails that

$$
n^{1/3}\left(\sum_{j=1}^p \hat{m}_j(x_j) - \sum m_j(x_j)\right) \xrightarrow{d} \sum_{j=1}^p \left\{\left(\frac{\sigma_j^2(x_j)\frac{\partial m}{\partial x_j}(\mathbf{x})}{2 f_{\mathbf{X}^j}(x_j)}\right)^{1/3} G_j\right\}
$$

To complete our analysis for the quantity $\hat{m}(\mathbf{x}) - m(\mathbf{x})$, we now study the difference between the constants $\hat{c}$ and c. Since $\hat{c}$ (together with $\hat{m}_1, \ldots, \hat{m}_p$) minimises the residual sum of squares, we have that

$$
\begin{aligned}
|\hat{c} - c| &= \left|\frac{1}{n}\sum_{i=1}^n \left\{\sum_{j=1}^p \Big(m_j(\mathbf{X}_{ij}) - \hat{m}_j(\mathbf{X}_{ij})\Big) + \sigma(\mathbf{X}_i)\epsilon_i\right\}\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \sigma(\mathbf{X}_i)\epsilon_i\right| + \sum_{j=1}^p \left|\frac{1}{n}\sum_{i=1}^n \Big(m_j(\mathbf{X}_{ij}) - \hat{m}_j(\mathbf{X}_{ij})\Big)\right|.
\end{aligned}
$$

We proceed as before by noting that $\frac{1}{n}\sum_{i=1}^{n}\sigma(\mathbf{X}_i)\epsilon_i = O_p(n^{-1/2})$. Furthermore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{m}_j(\mathbf{X}_{ij}) - m_j(\mathbf{X}_{ij})\Big) &= \int_{z_j}\big(\hat{m}_j(z_j) - m_j(z_j)\big)d\hat{F}_{\mathbf{X}^j}(z_j)\\ &= \int_{z_j}\big(\hat{m}_j(z_j) - m_j(z_j)\big)f_{\mathbf{X}^j}(z_j) + \int_{z_j}\big(\hat{m}_j(z_j) - m_j(z_j)\big)d(\hat{F}_{\mathbf{X}^j} - F_{\mathbf{X}^j})(z_j)\\ &= \text{(I)} + \text{(II)} = 0 + O_p(n^{-1/2}),\end{aligned}$$

where the last line follows from the identifiability condition and empirical process theory (the detailed steps of which were explained in earlier part of the proof). Consequently, $\hat{c} - c = o_p(n^{-1/3})$, and thus

$$n^{1/3}\Big(\hat{m}(x) - m(x)\Big) \xrightarrow{d} \sum_{j=1}^{p}\left\{\left(\frac{\sigma_j^2(x_j)\frac{\partial m}{\partial x_j}(\mathbf{x})}{2f_{\mathbf{X}^j}(x_j)}\right)^{1/3} G_i\right\}.$$

This completes the first part of our proof.

Second, we study the convergence rate of $|\hat{\mu}_R - \mu_R|$. Under Assumption A6^b, Horowitz and Mammen (2004) entails that $\sup_{z_j}|\tilde{m}_j(z_j) - m_j(z_j)| = O_p((\log n)^{1/2}n^{-2/5})$ for every $j = 1, \ldots, p$. Therefore, $\sup_{\mathbf{x}\in\Theta}|\tilde{m}(\mathbf{x}) - m(\mathbf{x})| = o_p(n^{-1/3})$. Using the argument similar to that mentioned in the second stage of Theorem 3.1's proof, it is straightforward to check that $|\hat{\mu}_R - \mu_R| = o_p(n^{-1/3})$.

Finally, the proof is completed via an application of the Slutsky's theorem, and by noting that $\frac{\partial m}{\partial x_j} = \mu_R\frac{\partial \rho}{\partial x_j}$.

□

Proofs of Corollary 3.2 – Corollary 3.5:

The proofs of Corollaries 3.2 – 3.5 are similar to that of Corollary 3.1, so are omitted for brevity.

□

REFERENCES

Fan, Y. and E. Guerre (2016). Multivariate local polynomial estimators: Uniform boundary properties and asymptotic linear representation. In *Essays in Honor of Aman Ullah*, pp. 489–537. Emerald.

Feng, O. Y., Y. Chen, Q. Han, R. J. Carroll, and R. J. Samworth (2022). Nonparametric, tuning-free estimation of s-shaped functions. *Journal of Royal Statistical Society Series B 84*, 1324–1352.

Ghosal, P. and B. Sen (2017). On univariate convex regression. *Sankhya : The Indian Journal of Statistics 79-A*, 215–253.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001). Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics 29*, 1653–1698.

Gu, J., Q. Li, and J.-C. Yang (2015). Multivariate local polynomial kernel estimators: leading bias and asymptotic distribution. *Econometric Reviews 34*, 979–1010.

Lim, E. and P. W. Glynn (2012). Consistency of multidimensional convex regression. *Operations Research 60*, 196–208.

Mammen, E. and K. Yu (2007). Additive isotone regression. In *IMS Lecture Notes Monograph Series, Asymptotics: Particles, Processes and Inverse Problems*, Volume 55, pp. 179–195. Institute of Mathematical Statistics.

Masry, E. (1996). Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes and their Applications 65*, 81–101.

Robertson, T., E. T. Wright, and R. L. Dykstra (1988). *Order Restricted Statistical Inference*. Wiley.

Seijo, E. and B. Sen (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics 39*, 1633–1657.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Yu, B. (1997). Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer.