Estimating Nonparametric Conditional Frontiers and Efficiencies: A New Approach

Camilla Mastromarco[*]
camilla.mastromarco@unical.it

Léopold Simar[‡]
leopold.simar@uclouvain.be

Ingrid Van Keilegom[§]
ingrid.vankeilegom@kuleuven.be

February 27, 2024

Abstract

In production theory, conditional frontiers and conditional efficiency measures are flexible and appealing tools to investigate the role of environmental variables on the production process. Direct approaches estimate non-parametrically conditional distribution functions requiring smoothing techniques and the use of bandwidths. Traditional methods for selecting bandwidths provide bandwidths with order that may not be optimal when estimating the boundary of the distribution function. In this paper we suggest an approach that avoids this problem, by eliminating in a first step, with flexible control functions, the influence of the environmental factors on the inputs and the outputs. By doing this we produce "pure" inputs and outputs which allow to estimate a "pure" measure of efficiency, more reliable for ranking the firms, since the influence of the external factors have been eliminated. We are also able to recover the frontier and efficiencies in original units. This can be viewed as an extension of location-scale models for whitening the variables, avoiding often inappropriate restrictions. We describe the method, its statistical properties and we show in some Monte-Carlo simulations, how our new method dominates both the traditional direct and the location-scale approaches. We illustrate the usefulness of the approach with a real data set on banks.

Key Words: Nonparametric frontier models, Environmental factors, Conditional efficiency, Robust estimation of frontiers
JEL Classification: C13,C14,C49

[*] Dipartimento di Economia, Statistica e Finanza DESF University of Calabria, Italy

[‡] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), LIDAM, UCLouvain, Louvain-la-Neuve, Belgium.

[§] ORSTAT, KULeuven, Leuven, Belgium and ISBA, UCLouvain, Louvain-la-Neuve, Belgium. I. Van Keilegom acknowledges the financial support from the FWO and F.R.S.-FNRS under the Excellence of Science (EOS) programme, project ASTeRISK (grant No. 40007517).

1 Introduction

1.1 The Setup and Our Contribution

In the productivity and efficiency analysis, we analyze how production units (firms) transform a set of inputs to produce a set of outputs. The efficient frontier is then defined in the input-output space as the locus of optimal organization, e.g. the maximal attainable level of the outputs given the level of the inputs, or the minimal attainable level of inputs to produce given level of outputs. The efficiency level of a firm is then determined by an appropriate distance (in the output direction or in the input direction) of this firm to the optimal frontier. Farrell-Debreu radial distances (Debreu, 1951; Farrell, 1957) are often used to measure these efficiency scores.

Recently, the efficiency literature has focused on the role of environmental variables, these factors are neither inputs nor outputs and are typically not under the control of the managers, but they might influence the production process. Formally, we will consider the univariate input oriented case,[1] where we look for the minimal input (or cost) $X \in \mathbb{R}_+$ achievable for producing the vector of goods or services $Y \in R_+^q$ when the producer is facing environmental factors described by $Z \in \mathbb{R}^d$. The environmental conditions may affect the range of the input-output (X, Y), and hence the shape of the frontier, or Z may affect only the distribution of the inefficiencies inside the attainable set, or in some cases Z may affect both, or finally, Z might be completely independent of (X, Y). In this paper we present a novel, fully nonparametric method designed to address the influence of these external variables. Our approach enables us to capture the intricate relationships between these variables and the target outcomes, without assuming any specific functional forms governing these relationships. It allows for more comprehensive analysis of how these external variables impact diverse units and distribution patterns. This aspect is of crucial importance for applied analysis, providing a robust foundation for exploring the complexities inherent in the relationships under consideration.

A general and appealing approach to investigate the role of Z on the production process, is to consider conditional frontiers and conditional efficiency scores, as introduced by Cazals et al. (2002) and Daraio and Simar (2005). They consider a probabilistic formulation of the production process where the random variables (X, Y, Z) are defined on an appropriate probability space. The conditional distribution of the input-outputs (X, Y) given particular values of Z is of central interest. It can be described by the conditional survival function

$$\begin{aligned} S_{X,Y|Z}(x, y|z) &= \mathbb{P}(X \geq x, Y \geq y|Z = z) \\ &= S_{X|Y,Z}(x|y, z) S_{Y|Z}(y|z), \end{aligned} \tag{1.1}$$

where $S_{X|Y,Z}(x|y, z) = \mathbb{P}(X \geq x|Y \geq y, Z = z)$ and $S_{Y|Z}(y|z) = \mathbb{P}(Y \geq y|Z = z)$. Note the unusual conditioning on Y in $S_{X|Y,Z}(x|y, z)$, because Y is an output. With this notation, the conditional minimum input (or cost) frontier is then defined as the minimal achievable input level x for units producing at least the level y of outputs, facing the environmental conditions z:

$$\tau(y, z) = \inf\{x | S_{X|Y,Z}(x|y, z) < 1\}. \tag{1.2}$$

[1] The presentation is easy to adapt to the univariate output oriented case, where we search for the maximal production level, given the level of a set of inputs.

Of course if there are no environmental factors Z, or if we want to consider an unconditional to Z frontier, the survival function $S_{X|Y,Z}(x|y,z)$ has to be replaced by the unconditional to Z survival function $S_{X|Y}(x|y) = \mathbb{P}(X \geq x|Y \geq y)$ in all the equations above. This will provide the full unconditional frontier

$$\tau(y) = \inf\{x|S_{X|Y}(x|y) < 1\}. \tag{1.3}$$

Given a sample of observations $\mathcal{S}_n = \{X_i, Y_i, Z_i)\}_{i=1}^n$, traditional nonparametric estimators of the frontier functions are obtained by plugging in a nonparametric estimator of the survival function in the appropriate formula. For the unconditional to Z case, it is given by

$$\widehat{S}_{X|Y}(x|y) = \frac{\sum_{i=1}^n 1\!\!I(X_i \geq x, Y_i \geq y)}{\sum_{i=1}^n 1\!\!I(Y_i \geq y)}, \tag{1.4}$$

where $1\!\!I(\cdot)$ is the indicator function. Then the FDH estimator (see Deprins et al., 1984) of the minimal input function is given by

$$\widehat{\tau}(y) = \inf\{x|\widehat{S}_{X|Y}(x|y) < 1\} = \min_{\{i|Y_i \geq y\}}\{X_i\}, \tag{1.5}$$

The conditional to Z cases, require some smoothing in Z, so we have

$$\widehat{S}_{X|Y,Z}(x|y,z) = \frac{\sum_{i=1}^n 1\!\!I(X_i \geq x, Y_i \geq y)K_{h_z}(Z_i - z)}{\sum_{i=1}^n 1\!\!I(Y_i \geq y)K_{h_z}(Z_i - z)}, \tag{1.6}$$

where h_z is a vector of d bandwidths $(h_{z_1}, \ldots, h_{z_d})$ with $K_{h_z}(u) = \prod_{j=1}^d (1/h_{z_j})k(u_j/h_{z_j})$ and k is a univariate kernel with support $[-1, 1]$. Then we have

$$\widehat{\tau}(y,z) = \inf\{x|\widehat{S}_{X|YZ}(x|y,z) < 1\}. \tag{1.7}$$

Partial frontiers have also been introduced in the literature, considering less extreme benchmarks, and so providing estimators more robust to outliers and extreme data points. In this paper we will focus on the order-m frontier introduced by Cazals et al. (2002). In Section 1.2 we will summarize most of the available properties of these nonparametric estimators of full and order-m robust frontiers. In particular the latter have better properties than the estimators of the full frontiers, i.e. convergence to a Gaussian process and better rates of convergence.

In this traditional or "direct" approach for estimating conditional frontiers, the statistical properties rely on the properties of the bandwidths h_z used for estimating the conditional survival function (1.1). Least squares cross-validation (LSCV) techniques are available providing bandwidths with the optimal order, i.e. $h_{z_j} \propto n^{-1/(d+4)}$ (see Li et al., 2013). However, as pointed in Bădin et al. (2019), these bandwidths might not be optimal when the objective is to estimate the lower bound of the support of $S_{X|Y,Z}(x|y,z)$. The problem, already noticed by Jeong et al. (2010), is that for a given h_z, the conditional FDH estimator does not target $\tau(y,z)$ but rather $\tau^{h_z}(y,z)$ defined as

$$\tau^{h_z}(y,z) = \inf\Big\{x|S^{h_z}_{X|Y,Z}(x|y,z) = \mathbb{P}\big(X \geq x \mid Y \geq y, |Z - z| \leq h_z\big) < 1\Big\} \tag{1.8}$$

where $|Z - z| \leq h_z$ has to be understood component by component, i.e. $|Z_j - z_j| \leq h_{z_j}$, $j = 1, \ldots, d$. This introduces an additional error, similar to a bias of localization, which under mild regularity condition (smoothness of the frontier as a function of z) is of order $||h_z||$, unless the separability condition holds (see Bădin et al., 2019, for details).[2] To the best of our knowledge, this issue has not been yet solved in the literature. Bădin et al. (2019) suggest a bootstrap (subsampling) algorithms to derive an optimal bandwidth, but this involves huge numerical burden and, as fairly said in their paper, there is no theoretical justification of the procedure, although their Monte-Carlo experiments, for univariate Z, are encouraging.

Florens et al. (2014) suggested an alternative approach, avoiding this bandwidth selection issue. They consider location-scale models to describe the links between the input, the outputs and Z. They assume that

$$\begin{cases} X = \mu_x(Z) + \sigma_x(Z)\varepsilon_x \\ Y_j = \mu_{y_j}(Z) + \sigma_{y_j}(Z)\varepsilon_{y_j} \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{1.9}$$

where $(\varepsilon_x, \varepsilon_y)$ is assumed to be independent of Z, with mean zero and variance one.[3] Then if their location-scale assumptions hold, the resulting error terms of these models, ε_x and ε_y, provide, in a sense, "cleaned" versions of X and of Y, or "pure" input and outputs whitened from the effects of the environmental variables Z. In this pure input-output space we can define an efficient frontier which allows to estimate a "pure" measure of efficiency or "managerial" efficiency, more reliable for ranking the firms, since the influence of the external factors have been eliminated.

As described in Florens et al. (2014), the nonparametric estimation of the model (estimation of $\mu_\ell(Z)$ and of $\sigma_\ell(Z)$ where ℓ stands for x or y_j) involves only smoothing in the center of the data clouds and avoids the bandwidth selection issue described above. However, the location-scale approach suffers from some drawbacks. First the model (1.9) assumes that Z cannot affect the shape of the distribution of the variables (X, Y) but only their means and variances. This is not common in most of the models used in the frontier literature where e.g. skewness or kurtosis of the variables may also be dependent of Z. Second, even if the assumptions in (1.9) remains reasonable, the nonparametric estimation of the scale functions $\sigma_\ell(Z)$, is obtained by regressing on Z, the squares of the residuals obtained in a first step nonparametric estimation of the location functions $\mu_\ell(Z)$. Squaring these residuals introduces some statistical instability, in particular for the extreme data values, as will be confirmed in our Monte-Carlo experiments in Sections 4.2 and 4.3.

In this paper we propose a new approach avoiding the drawbacks and restrictions of the location-scale model. The links between the input, the outputs and the environmental factors Z, are described by fully nonparametric models based on control functions. We suppose the vector (X, Y, Z) satisfies the following model:

$$\begin{cases} X = \varphi_x(Z, U_x) \\ Y_j = \varphi_{y_j}(Z, U_{y_j}) \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{1.10}$$

[2] The separability condition, introduced by Simar and Wilson (2007), states that the support of $S_{X|Y,Z}(x|y,z)$ does not depend on z, i.e. Z has no effect on the shape or the level of the frontier.

[3] Note that we use in our paper the traditional notation in the frontier literature where X is the input and Y the outputs. Florens et al. (2014) did the opposite way.

where the functions $\varphi_\ell(\cdot, U_\ell)$ are nonseparable and monotone increasing in U_ℓ, $\ell = x, y_1, \ldots, y_q$. Without loss of generality (see below Section 2), we assume that the U_ℓ are uniformly distributed on $[0, 1]$. The model assumes that the U_ℓ are independent of Z, this assumption is part of the model and, as seen below, is needed to identify each individual equation in (1.10). This kind of models (control functions) have been used e.g. in Matzkin (2003), or Imbens and Newey (2009) in a different context where Z play the role of instruments to address endogeneity issues in regression models and in Simar et al. (2016) to identify latent heterogeneity in frontier models.

Looking to the equations (1.10), we see that due to the independence between U and Z, the variables U are constructed as being the part of the input (respectively outputs) which is independent on Z. In other words they are whitened versions of X and Y defined in a set of general nonparametric nonseparable equations. We can, in the same lines of Florens et al. (2014), interpret (U_x, U_y) as "pure" input and outputs which remain monotone transformations of the original measures. It can also be seen as the part of the input and outputs not dependent on Z. So, here again we will be able to build the efficient frontier in the pure input-output space, allowing to define an efficiency score, or a "managerial" efficiency, that will be independent of the environmental conditions. This will be done for both full frontiers and robust order-m ones. We will see below that the location-scale models of Florens et al. (2014) can be viewed as a particular semiparametric case of our model.

In this paper, we prove that, in our model (1.10), (i) the resulting estimators of the pure order-m efficiency measures are free of the curse of dimensionality due to the dimension of Z and converge with rate $\sqrt{n}$ to a Gaussian process, (ii) we are able to recover the frontiers in the original units of (X, Y), and (iii), the order-m frontier estimates in the original units, converge to a Gaussian process at rate $\sqrt{n\bar{h}_z}$ where $\bar{h}_z$ is the product of d bandwidths defined in the next sections. The latter results is similar to the properties of the estimators derived by the direct approach or by the location-scale models (see Section 1.2). These results allow to use the bootstrap for making practical inference on order-m efficiencies, as illustrated below in our real data example.

The paper is organized as follows. After Section 1.2, that summarizes known properties of nonparametric estimators of full and order-m frontiers, Section 2 presents the properties of our model and the way to compute nonparametric estimators. Section 3 gives the asymptotic properties of the estimators and Section 4 shows through some simulations how our model produces, as expected, estimators that dominates (in term of $MISE$) the traditional direct approach, but even those obtained through the location-scale models. We illustrate in Section 5 its practical use with real data. Section 6 concludes. Technical details of the proofs are given in Appendix A and more complete set of results of Monte-Carlo experiments are described in supplementary material (Appendix B).

1.2 Nonparametric Frontier Estimators, in a Nutshell

This section summarizes facts from the literature that are useful to permit comparison with the properties of the estimators obtained by our new method. The reader familiar with nonparametric frontier estimation can skip this section. More details and additional references can be found e.g. in the survey Simar and Wilson (2015).

The partial order-m cost-frontier (see Cazals et al., 2002) is defined, for a given integer

$m \geq 1$ as

$$\tau_m(y) = \mathbb{E}[\min(X_1, \ldots, X_m) | Y \geq y] = \int_0^\infty S^m_{X|Y}(x|y)\, dx. \tag{1.11}$$

The idea is to benchmark the cost of a unit producing the value y of outputs not against the minimal technically possible cost of such firms (as for the full frontier) but against the average of m peers producing at least the value y of outputs. We have $\tau_m(y) \geq \tau(y)$ and $\tau_m(y) \to \tau(y)$, as $m \to \infty$ (see Cazals et al., 2002; Daraio and Simar, 2007; Daouia and Gijbels, 2011, for details and discussions on the choice of m). In the presence of environmental variables we have, by analogy

$$\begin{aligned} \tau_m(y, z) &= \mathbb{E}[\min(X_1, \ldots, X_m) | Y \geq y, Z = z] \\ &= \int_0^\infty S^m_{X|Y,Z}(x|y, z)\, dx, \end{aligned} \tag{1.12}$$

Nonparametric estimators are obtained by plugging in (1.11) and (1.12) the appropriate nonparametric estimators (1.4) or (1.6). We have

$$\widehat{\tau}_m(y) = \int_0^\infty \widehat{S}^m_{X|Y}(x|y)\, dx. \tag{1.13}$$

$$\widehat{\tau}_m(y, z) = \int_0^\infty \widehat{S}^m_{X|YZ}(x|y, z)\, dx. \tag{1.14}$$

The statistical properties of the resulting frontier estimators for the full and order-m frontiers are well established: see Park et al. (2000), Cazals et al. (2002). We have

$$n^{1/(1+q)}\big(\widehat{\tau}(y) - \tau(y)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_y^{(1+q)}, 1 + q) \tag{1.15}$$

$$\sqrt{n}\big(\widehat{\tau}_m(y) - \tau_m(y)\big) \xrightarrow{\mathcal{L}} \text{N}(0, \sigma_y^2), \tag{1.16}$$

where exact expressions for the parameters of the limiting distributions have been derived. For the conditional to Z frontiers, it has been proven (see Cazals et al., 2002; Jeong et al., 2010) that under mild regularity conditions, we have similar results where the sample size n has to be replaced by its effective number of observations in a neighborhood of z, namely $n\bar{h}_z$ where $\bar{h}_z = \prod_{j=1}^d h_{z_j}$. So to summarize:

$$(n\bar{h}_z)^{1/(1+q)}\big(\widehat{\tau}(y, z) - \tau(y, z)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_{y,z}^{(1+q)}, 1 + q) \tag{1.17}$$

$$\sqrt{n\bar{h}_z}\big(\widehat{\tau}_m(y, z) - \tau_m(y, z)\big) \xrightarrow{\mathcal{L}} \text{N}(0, \sigma_{y,z}^2), \tag{1.18}$$

In this traditional or "direct" approach for estimating conditional frontiers, the results in (1.17) and (1.18) rely on the properties of the bandwidths h_z used for estimating the conditional survival function (1.1). Least squares cross-validation (LSCV) techniques provide bandwidths with the order, $h_{z_j} \propto n^{-1/(d+4)}$ (see Li et al., 2013) which deteriorates the rates in (1.17) and (1.18), in particular when d increases. The same happens when using the corrected bandwidths suggested by Bădin et al. (2019). We will establish in Section 3 the asymptotic properties for the estimators derived from our new method.

2 Properties and Estimation of the Model

Looking to our model (1.10), when U_ℓ are uniform on $[0,1]$ for $\ell = x, y_1, \ldots, y_q$, then φ_ℓ can be interpreted as a quantile function, as shown below. The choice of the uniform is a matter of rescaling the U_ℓ to get this nice interpretation.[4] It is known that under the above assumptions, the U_ℓ are identified by the conditional distribution of the input and the outputs given Z. The argument is going along the following lines (we do it for the input, but it is the same for the outputs):

$$\begin{aligned}
F_{X|Z}(x|z) &= \mathbb{P}(X \leq x | Z = z) \\
&= \mathbb{P}(\varphi_x(Z, U_x) \leq x | Z = z) \\
&= \mathbb{P}(U_x \leq \varphi_x^{-1}(Z, x) | Z = z) \\
&= \mathbb{P}(U_x \leq \varphi_x^{-1}(z, x)) \\
&= \varphi_x^{-1}(z, x),
\end{aligned}$$

where the last line is obtained because we assume $U_x \sim \text{Unif}([0,1])$. Since this is true for all (x, z), we have $U_x = F_{X|Z}(X|Z)$ with probability one. So, more generally we have:

$$\begin{cases} U_x = F_{X|Z}(X|Z) \\ U_{y_j} = F_{Y_j|Z}(Y_j|Z) \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{2.1}$$

So we see that $\varphi_x(Z, U_x) = F^{-1}_{X|Z}(U_x|Z)$, i.e. the conditional quantile of X given Z evaluated at $U_x \in [0,1]$. The same is true for $Y_1, \ldots, Y_q$, each function $\varphi_{y_j}(Z, U_{y_j}) = F^{-1}_{Y_j|Z}(U_{y_j}|Z)$, $j = 1, \ldots, q$ has the same conditional quantile interpretation.

Since the functions φ_ℓ are unknown, the values $U_{\ell,i}$ are not observed but they can be estimated by nonparametric methods by estimating the appropriate conditional distribution functions:

$$\widehat{U}_{x,i} = \widehat{F}_{X|Z}(X_i|Z_i) = \frac{\sum_{k=1}^n G_{h_x}(X_k - X_i) K_{h_z}(Z_k - Z_i)}{\sum_{k=1}^n K_{h_z}(Z_k - Z_i)}, \tag{2.2}$$

$$\widehat{U}_{y_j,i} = \widehat{F}_{Y_j|Z}(Y_{j,i}|Z_i) = \frac{\sum_{k=1}^n G_{h_{y_j}}(Y_{j,k} - Y_{j,i}) K_{h_z}(Z_k - Z_i)}{\sum_{k=1}^n K_{h_z}(Z_k - Z_i)}, \tag{2.3}$$

where now optimal bandwidths h_z, h_x and h_{y_j} can be obtained for each equation, by the LSCV techniques described in Li et al. (2013) and we avoid the problem of selecting optimal bandwidths h_z when estimating the boundary of some conditional distribution function and all the issues mentioned above. Here the kernels used are standard: $K_{h_z}(\cdot)$ are usual kernels for estimating densities and $G_{h_\ell}(\cdot)$ are cumulative kernels used for estimating distribution functions (cdf).[5] For example in the input case $G_{h_x}(X_k - X_i) = G\big((X_k - X_i)/h_x\big)$ and G is a cdf defined as $G(v) = \int_{-\infty}^{v} w(u)du$ for some kernel density function $w(\cdot)$ (see e.g. Li et al.,

[4] Since the functions φ_ℓ in (1.10) are monotone increasing with respect to U_ℓ, any monotone, increasing transformation of U_ℓ could be included in φ_ℓ, to get any desired continuous distribution for U_ℓ. But the interpretation of the resulting φ_ℓ will depend on the specific transformation that is used.

[5] We could avoid the smoothing in the variables X and Y, and use indicator functions in place of the kernels G. This would probably not change the practical results, but for the theory below, we need smoothed estimators of these distribution functions.

2013 for details). Note again that here we are interested in the estimation of the distribution on its full range and not only on the boundary of their support where the data can be rather sparse. We will see below that estimates of the quantile functions can also be derived from these quantities.

Having these input and outputs in pure units, we can estimate the minimal cost frontier and its order-m robust version, by usual techniques. For the full frontier, it is defined in pure units by

$$\phi(u_y) = \inf\{u_x | S_{U_x|U_y}(u_x|u_y) < 1\}, \tag{2.4}$$

where $S_{U_x|U_y}(u_x|u_y) = \mathbb{P}(U_x \geq u_x | U_y \geq u_y\}$. So $\phi(u_y)$ is the minimal achievable level of input in pure units, for units producing at least the level of output u_y in pure units. For the order-m frontier, we have for a given m

$$\begin{aligned}\phi_m(u_y) &= \mathbb{E}\left[\min(U_{x,1}, \dots, U_{x,m}) | U_y \geq u_y\right], \\ &= \int_0^1 S^m_{U_x|U_y}(u_x|u_y)\, du_x, \end{aligned} \tag{2.5}$$

which provides also, for finite m, a less extreme benchmark than the full frontier $\phi(u_y)$. Here, as shown in Cazals et al. (2002), as $m \to \infty$, we have $\phi_m(u_y) \to \phi(u_y)$.

Note that from the frontiers in the pure units we can recover the frontiers in the original units. We have indeed due to our assumptions (with some abuse of notation where multidimensional inequalities involving y have to be understood component by component):

$$\begin{aligned}\tau(y, z) &= \inf\{x | S_{X|Y,Z}(x|y, z) < 1\} \\ &= \inf\{x | \mathbb{P}(X \geq x | Y \geq y, Z = z) < 1\} \\ &= \inf\{x | \mathbb{P}(\varphi_x(Z, U_x) \geq x | \varphi_y(Z, U_y) \geq y, Z = z) < 1\} \\ &= \inf\{x | \mathbb{P}(U_x \geq u_x = \varphi_x^{-1}(z, x) | U_y \geq u_y = \varphi_y^{-1}(z, y)) < 1\} \\ &= \inf\{x | S_{U_x|U_y}(u_x|u_y) < 1\} \text{ with } u_x = \varphi_x^{-1}(z, x), u_y = \varphi_y^{-1}(z, y), \end{aligned} \tag{2.6}$$

where we used the monotonicity properties of the functions φ_ℓ^{-1} and the independence between (U_x, U_y) and Z. Now due to the monotonicity of $\varphi_x(z, u_x)$ in u_x, we have

$$\tau(y, z) = \varphi_x\left(z, \inf\{u_x | S_{U_x|U_y}(u_x|u_y) < 1\}\right) = \varphi_x(z, \phi(u_y)), \text{ where } u_y = \varphi_y^{-1}(z, y). \tag{2.7}$$

Note that here and below, $u_y = \varphi_y^{-1}(z, y)$ means $u_{y_j} = F_{Y_j|Z}(y_j|z)$ for $j = 1, \dots, q$.

For the order-m the relation is little more tricky because the transformation $\varphi_x(z, u_x)$ is not necessarily linear in u_x (as it was the case in the particular location-scale model (1.9)). Here, we obtain the following equivalence

$$\{\min(X_1, \dots, X_m) | Y \geq y, Z = z\} = \varphi_x\left(z, \min\left\{(U_{x,1}, \dots, U_{x,m}) | U_y \geq u_y\right\}\right), \tag{2.8}$$

where the last expression can be read as $\varphi_x(z, \cdot)$ is evaluated at the minimum of m iid replications of $U_{x,j}$, $j = 1, \dots, m$ generated under the condition $U_y \geq u_y$. Here again $u_y = \varphi_y^{-1}(z, y)$. So to recover the order-m frontier in original units, we need to compute the following expectation

$$\tau_m(y, z) = \mathbb{E}\left[\varphi_x(z, \min(U_{x,1}, \dots, U_{x,m}) | U_y \geq u_y)\right], \tag{2.9}$$

which leads to

$$\tau_m(y,z) = \int_0^1 \varphi_x(z,t) dF_{\tilde{U}_m|U_y}(t|u_y), \tag{2.10}$$

where $\tilde{U}_m = \min(U_{x,1},\ldots,U_{x,m})$ and $F_{\tilde{U}_m|U_y}(t|u_y) = 1 - S^m_{U_x|U_y}(t|u_y)$.

The nonparametric estimator of the frontier in the pure units is obtained from our set of predicted values of input and outputs in the space of pure units: $\{(\widehat{U}_{x,i},\widehat{U}_{y,i})\}_{i=1}^n$ by plugging in the nonparametric estimator of $S_{U_x|U_y}(u_x|u_y)$ in equations (2.4) and (2.5), respectively. This estimator is simply given by the empirical version of $S_{U_x|U_y}(u_x|u_y)$:

$$\widehat{S}_{U_x|U_y}(u_x|u_y) = \frac{\sum_{i=1}^n 1\!\!I(\widehat{U}_{x,i} \geq u_x, \widehat{U}_{y,i} \geq u_y)}{\sum_{i=1}^n 1\!\!I(\widehat{U}_{y,i} \geq u_y)}. \tag{2.11}$$

So, we recover the FDH estimator of $\phi(u_y)$ in pure units:

$$\begin{aligned}\widehat{\phi}(u_y) &= \inf\{u_x | \widehat{S}_{U_x|U_y}(u_x|u_y) < 1\},\\ &= \min_{i|\widehat{U}_{y,i} \geq u_y} \widehat{U}_{x,i}.\end{aligned} \tag{2.12}$$

The order-m in pure units is given by the integral

$$\widehat{\phi}_m(u_y) = \int_0^1 \widehat{S}^m_{U_x|U_y}(u_x|u_y)\, du_x, \tag{2.13}$$

where an exact formula is provided in Cazals et al. (2002).

Estimates of the frontiers in original units are easy to recover for the full frontier

$$\widehat{\tau}(y,z) = \widehat{F}^{-1}_{X|Z}(\widehat{\phi}(\hat{u}_y)|z), \tag{2.14}$$

where $\hat{u}_{y_j} = \widehat{F}_{Y_j|Z}(y_j|z)$, for $j = 1,\ldots,q$ and $\widehat{F}^{-1}_{X|Z}(\cdot|z)$ is the nonparametric estimator of the conditional quantile of X given $Z = z$ (see Li et al. 2013 for computations and properties). To be explicit, we have e.g.

$$\widehat{\tau}(y,z) = \inf\left\{t \mid \widehat{F}_{X|Z}(t|z) \geq \widehat{\phi}(\hat{u}_y)\right\} = \arg\min_t \mid \widehat{\phi}(\hat{u}_y) - \widehat{F}_{X|Z}(t|z) \mid . \tag{2.15}$$

For reasons explained above, in the order-m case, $\widehat{\phi}_m(u_y)$ is not useful to recover the order-m frontier in original unit due to the nonlinearity of $\varphi_x(z,u_x)$ as a function of u_x. In fact we have directly from (2.10) that

$$\widehat{\tau}_m(y,z) = \int_0^1 \widehat{\varphi}_x(z,t)\, d\widehat{F}_{\tilde{U}_m|U_y}(t|\hat{u}_y), \tag{2.16}$$

where $\widehat{F}_{\tilde{U}_m|U_y}(t|\hat{u}_y) = 1 - \widehat{S}^m_{U_x|U_y}(t|\hat{u}_y)$.

This can be computed as follows. Denote n_y the number of observations $\widehat{U}_{i,y} \geq \hat{u}_y$ and let $V^y_{(1)} \leq \ldots \leq V^y_{(n_y)}$ be the order statistics of the estimates $\widehat{U}_{j,x}$ such that $\widehat{U}_{j,y} \geq \hat{u}_y$. Then it is easy to show that

$$\widehat{\tau}_m(y,z) = \sum_{j=1}^{n_y} \widehat{\varphi}_x(z, V^y_{(j)}) \left[\left(\frac{n_y - j + 1}{n_y}\right)^m - \left(\frac{n_y - j}{n_y}\right)^m \right], \tag{2.17}$$

where for all $\gamma \in (0,1)$,

$$\widehat{\varphi}_x(z,\gamma) = \widehat{F}^{-1}_{X|Z}(\gamma|z) = \inf\left\{t \mid \widehat{F}_{X|Z}(t|z) \geq \gamma\right\} = \arg\min_t \mid \gamma - \widehat{F}_{X|Z}(t|z) \mid, \qquad (2.18)$$

i.e. the γ-quantile of $\widehat{F}_{X|Z}(\cdot|z)$.

Having the estimates of the conditional frontiers and estimates of the unconditional (marginal) frontiers, several analyses could be of interest to the practitioner to investigate the effect of the environmental factors on the production process. The methodology proposed in Bădin et al. (2012), and applied with the location-scale approach in Florens et al. (2014), allows to disentangle the role of Z on the level of the frontier and its effect of the distribution of efficiencies. As illustrated in a real data example below, our approach allows the same kind of analysis. Since it is fully nonparametric and does not rely on restrictive assumptions on the effect of Z on X and Y, this may provide promising tools for the practitioner. We will also see from our Monte-Carlo experiments below that our method provides much more reliable results than the traditional direct approach and even than the location-scale approach (even in case where the latter is true).

Remark 2.1 *As pointed out above, the Florens et al. (2014) location-scale model* (1.9) *is indeed a particular case of our model, where the functions φ_ℓ, for $\ell = x, y$ have an additive separable location scale structure $\varphi_\ell(Z, U_\ell) = \mu_\ell(Z) + \sigma_\ell(Z)\varepsilon_\ell$ with $(\varepsilon_x, \varepsilon_y)$ being independent of Z. So $\varepsilon_\ell = \eta_\ell(U_\ell)$, with $\eta_\ell(U_\ell)$ monotone in U_ℓ (as φ_ℓ), hence $U_\ell = \eta_\ell^{-1}(\varepsilon_\ell)$. We select $U_\ell = F_{\varepsilon_\ell}(\varepsilon_\ell)$ to get a uniform on* $[0,1]$ *and $\varepsilon_\ell = F^{-1}_{\varepsilon_\ell}(U_\ell)$.*

Remark 2.2 *If Z is fully independent of X and/or of Y, Florens et al. (2014) noticed that in the location-scale models, the corresponding functions $\mu(Z)$ and $\sigma(Z)$ would be constant. In our model here, for instance if Z is independent of the input X, we would have $F_{X|Z} \equiv F_X$ and so $U_x = F_X(X)$ would be just a monotone rescaling of X to become uniform on* $[0,1]$. *The same for any output. So the procedure above is still valid if Z is independent of (X,Y). Our model does not involve spurious dependencies in the estimation.*

3 Asymptotic Properties

For simplicity, we take $d = q = 1$ and $k = w$, where the functions k and w are the univariate kernels used in (2.2)–(2.3). The general case of multivariate outputs and environmental variables can be handled using higher order kernels or local polynomial smoothing methods, instead of the second order local constant smoothers that are used here. We will show below that with our general nonparametric approach to whiten the input and the outputs from the dependence on Z, we keep similar properties for the resulting estimators than the ones derived in Florens et al. (2014), but here, without the location-scale assumption.

The assumptions under which the asymptotic results are valid are:

(C1) The function k is defined on $[-1,1]$, and is a nonnegative, symmetric, and bounded second-order kernel function (i.e. $\int k(u)u\,du = 0$ and $0 < \int k(u)u^2\,du < \infty$).

(C2) The bandwidths h_x, h_y and h_z satisfy $h_x, h_y, h_z \to 0$, $nh_x \to \infty$, $nh_y \to \infty$ and $nh_z^{3+\delta} \to \infty$ for some $0 < \delta < 1$.

(C3) The functions $f_Z(z)$, $F_{X|Z}(x|z)$ and $F_{Y|Z}(y|z)$ have uniformly continuous fourth-order partial derivatives with respect to x, y and z.

(C4) The support of (X, Y, Z) in $\mathbb{R}^3$ is compact and we will denote by R_X, R_Y, R_Z the corresponding marginal supports, $\inf_{x,z} f_{X|Z}(x|z) > 0$ and $\inf_{y,z} f_{Y|Z}(y|z) > 0$, where $f_{X|Z} = F'_{X|Z}$ and $f_{Y|Z} = F'_{Y|Z}$.

(C5) The functions $z \to F^{-1}_{X|Z}(u|z)$ and $z \to F^{-1}_{Y|Z}(u|z)$ belong to $C^{1+\delta}_M(R_Z)$ for all $u \in [0, 1]$, where $C^{1+\delta}_M(R_Z)$ is the space of all functions $f : R_Z \to \mathbb{R}$ such that $\|f\|_{1+\delta} \le M$, where

$$\|f\|_{1+\delta} = \max\Big(\sup_{z\in R_Z} |f(z)|, \sup_{z\in R_Z} |f'(z)|\Big) + \sup_{z_1,z_2\in R_Z} \frac{|f'(z_1) - f'(z_2)|}{|z_1 - z_2|^\delta},$$

and where $M < \infty$ and $0 < \delta < 1$ is defined in (C2).

Theorem 3.1 *Assume (C1)-(C5).*

(i) Then,

$$\widehat{S}_{U_x|U_y}(u_x|u_y) - S_{U_x|U_y}(u_x|u_y)$$
$$= n^{-1}\sum_{i=1}^{n} \xi(X_i, Y_i, Z_i, u_x|u_y) + h_x^2 b_x(u_x|u_y) + h_y^2 b_y(u_x|u_y) + h_z^2 b_z(u_x|u_y) + R_n(u_x|u_y),$$

where

$$\begin{aligned}
\xi(X, Y, Z, u_x|u_y) &= S_{U_y}(u_y)^{-1}\big[1\!\!1(F_{X|Z}(X|Z) \ge u_x, F_{Y|Z}(Y|Z) \ge u_y) - S_{U_x,U_y}(u_x, u_y)\big] \\
&\quad + f_{U_x|U_y}(u_x|u_y)\,\big[1\!\!1(F_{X|Z}(X|Z) \le u_x) - u_x\big] \\
&\quad - \frac{\partial}{\partial u_y} S_{U_x|U_y}(u_x|u_y)\,\big[1\!\!1(F_{Y|Z}(Y|Z) \le u_y) - u_y\big] \\
b_x(u_x|u_y) &= f_{U_x|U_y}(u_x|u_y)\int B_x(F^{-1}_{X|Z}(u_x|z)|z) f_Z(z)dz \\
b_y(u_x|u_y) &= -\frac{\partial}{\partial u_y} S_{U_x|U_y}(u_x|u_y)\int B_y(F^{-1}_{Y|Z}(u_y|z)|z) f_Z(z)dz \\
b_z(u_x|u_y) &= f_{U_x|U_y}(u_x|u_y)\int B_z(F^{-1}_{X|Z}(u_x|z)|z) f_Z(z)dz \\
&\quad - \frac{\partial}{\partial u_y} S_{U_x|U_y}(u_x|u_y)\int B_z(F^{-1}_{Y|Z}(u_y|z)|z) f_Z(z)dz,
\end{aligned}$$

and $\sup_{0\le u_x,u_y\le 1} |R_n(u_x|u_y)| = o_P(n^{-1/2})$*, and where the functions* B_x, B_y *and* B_z *are given in Lemma A.1, in the Appendix.*

(ii) We have $\sup_{0\le u_x,u_y\le 1} |\widehat{S}_{U_x|U_y}(u_x|u_y) - S_{U_x|U_y}(u_x|u_y)| = O_P(n^{-1/2}) + O(h_x^2 + h_y^2 + h_z^2)$*, and if* $h_x = C_x n^{-1/4}(1 + o(1))$*,* $h_y = C_y n^{-1/4}(1 + o(1))$ *and* $h_z = C_z n^{-1/4}(1 + o(1))$ *for some* $0 \le C_x, C_y, C_z < \infty$*, then the process* $n^{1/2}(\widehat{S}_{U_x|U_y}(u_x|u_y) - S_{U_x|U_y}(u_x|u_y))$*,* $0 \le u_x, u_y \le 1$*, converges weakly to a Gaussian process* $W(u_x|u_y)$ *with covariance function given by*

$$Cov(W(u_{x1}|u_{y1}), W(u_{x2}|u_{y2})) = \mathbb{E}[\xi(X, Y, Z, u_{x1}|u_{y1})\xi(X, Y, Z, u_{x2}|u_{y2})]$$

and mean function given by $\mathbb{E}[W(u_x|u_y)] = C_x^2 b_x(u_x|u_y) + C_y^2 b_y(u_x|u_y) + C_z^2 b_z(u_x|u_y)$.

We next prove the asymptotic i.i.d. representation and the weak convergence of the estimators of the order-m frontier in pure units $\widehat{\phi}_m(u_y)$ and the order-m frontier in the original units $\widehat{\tau}_m(y,z)$.

Theorem 3.2 *Assume (C1)-(C5).*

(i) Then,

$$\widehat{\phi}_m(u_y) - \phi_m(u_y) = n^{-1}\sum_{i=1}^{n} \eta(X_i, Y_i, Z_i, u_y) \\ + h_x^2 b_x(u_y) + h_y^2 b_y(u_y) + h_z^2 b_z(u_y) + R_n(u_y),$$

where

$$\begin{aligned} \eta(X,Y,Z,u_y) &= 2^{m-1}\int_0^1 S^m_{U_x|U_y}(u_x|u_y)\xi(X,Y,Z,u_x|u_y)\,du_x \\ b_x(u_y) &= 2^{m-1}\int_0^1 S^m_{U_x|U_y}(u_x|u_y)b_x(u_x|u_y)\,du_x \\ b_y(u_y) &= 2^{m-1}\int_0^1 S^m_{U_x|U_y}(u_x|u_y)b_y(u_x|u_y)\,du_x \\ b_z(u_y) &= 2^{m-1}\int_0^1 S^m_{U_x|U_y}(u_x|u_y)b_z(u_x|u_y)\,du_x, \end{aligned}$$

and $\sup_{0\le u_y\le 1}|R_n(u_y)| = o_P(n^{-1/2})$.

(ii) We have $\sup_{0\le u_y\le 1}|\widehat{\phi}_m(u_y) - \phi_m(u_y)| = O_P(n^{-1/2}) + O(h_x^2 + h_y^2 + h_z^2)$*, and if* $h_x = C_x n^{-1/4}(1+o(1))$*,* $h_y = C_y n^{-1/4}(1+o(1))$ *and* $h_z = C_z n^{-1/4}(1+o(1))$ *for some* $0 \le C_x, C_y, C_z < \infty$*, then the process* $n^{1/2}(\widehat{\phi}_m(u_y) - \phi_m(u_y))$*,* $0 \le u_y \le 1$*, converges weakly to a Gaussian process* $W_m(u_y)$ *with covariance function given by*

$$Cov(W_m(u_{y1}), W_m(u_{y2})) = \mathbb{E}[\eta(X,Y,Z,u_{y1})\eta(X,Y,Z,u_{y2})]$$

and mean function given by $\mathbb{E}[W_m(u_y)] = C_x^2 b_x(u_y) + C_y^2 b_y(u_y) + C_z^2 b_z(u_y)$.

Theorem 3.3 *Assume (C1)-(C5).*

(i) Then,

$$\widehat{\tau}_m(y,z) - \tau_m(y,z) = (nh_z)^{-1}\sum_{i=1}^{n} k\Big(\frac{Z_i - z}{h_z}\Big) g(X_i, Y_i, Z_i, y|z) \\ + h_x^2 b_x(y|z) + h_y^2 b_y(y|z) + h_z^2 b_z(y|z) + R_n(y|z),$$

where

$$
\begin{aligned}
g(X,Y,Z,y|z) &= \int_0^1 \frac{1\!\!1(X \le F^{-1}_{X|Z}(t|Z)) - t}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z) f_Z(z)} dS^m_{U_x|U_y}(t|u_y) \\
&+ m \frac{1\!\!1(Y \le y) - F_{Y|Z}(y|Z)}{f_Z(z)} \int_0^1 F^{-1}_{X|Z}(t|z) \frac{\partial}{\partial u_y} \left[S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y) \right] dt,
\end{aligned}
$$

and

$$
\begin{aligned}
b_x(y|z) &= \int_0^1 \frac{B_x(F^{-1}_{X|Z}(t|z)|z)}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z)} dS^m_{U_x|U_y}(t|u_y) \\
b_y(y|z) &= B_y(y|z) \int_0^1 F^{-1}_{X|Z}(t|z) \frac{\partial}{\partial u_y} \left[S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y) \right] dt \\
b_z(y|z) &= \int_0^1 \frac{B_z(F^{-1}_{X|Z}(t|z)|z)}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z)} dS^m_{U_x|U_y}(t|u_y) \\
&+ B_z(y|z) \int_0^1 F^{-1}_{X|Z}(t|z) \frac{\partial}{\partial u_y} \left[S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y) \right] dt,
\end{aligned}
$$

and $\sup_{y \in R_Y, z \in R_Z} |R_n(y|z)| = o_P((nh_z)^{-1/2})$.

(*ii*) *We have* $\sup_{y \in R_Y, z \in R_Z} |\widehat{\tau}_m(y,z) - \tau_m(y,z)| = O_P((nh_z)^{-1/2}) + O(h_x^2 + h_y^2 + h_z^2)$, *and if* $h_x = C_x n^{-1/5}(1+o(1))$, $h_y = C_y n^{-1/5}(1+o(1))$ *and* $h_z = C_z n^{-1/5}(1+o(1))$ *for some* $0 \le C_x, C_y, C_z < \infty$, *then the process* $(nh_z)^{1/2}(\widehat{\tau}_m(y,z) - \tau_m(y,z))$, $y \in R_Y$ *(z and m fixed), converges weakly to a Gaussian process* $V_m(y|z)$ *with covariance function*

$$
Cov(V_m(y_1|z), V_m(y_2|z)) = \nu_0 f_Z(z) \, \mathbb{E}[g(X,Y,Z,y_1|z) g(X,Y,Z,y_2|z)|Z=z]
$$

and mean function given by $\mathbb{E}[V_m(y|z)] = C_x^2 b_x(y|z) + C_y^2 b_y(y|z) + C_z^2 b_z(y|z)$, *where* $\nu_0 = \int k^2(u)du$.

As a last asymptotic result we show the weak consistency of the estimators of the full frontier in pure units $\widehat{\phi}(u_y)$ and the full frontier in the original units $\widehat{\tau}(y,z)$.

Theorem 3.4 *Assume (C1)-(C5). Then,*

$$
\widehat{\phi}(u_y) - \phi(u_y) = o_P(1), \qquad \text{and} \qquad \widehat{\tau}(y,z) - \tau(y,z) = o_P(1),
$$

for all $0 \le u_y \le 1, y \in R_Y$ *and* $z \in R_Z$.

To summarize, we obtain convergence to Gaussian processes when order-m objects are considered, but the asymptotic bias and variances although explicit, are difficult to estimate. So, in practice bootstrap methods will allow to make inference for the order-m quantities. As in Florens et al. (2014), we will use a smoothed bootstrap on the "residuals" $(\widehat{U}_{x,i}, \widehat{U}_{y,i})$ for each Z_i producing $(U_{x,i}^{*,s}, U_{y,i}^{*,s})$, $i = 1, \ldots, n$.[6] Then we generate (X_i^*, Y_i^*) by our estimated model (1.10). So we have a bootstrap sample (X_i^*, Y_i^*, Z_i^*), $i = 1, \ldots, n$, where

$$Z_i^* = Z_i, \tag{3.1}$$

$$X_i^* = \widehat{F}_{X|Z}^{-1}(U_{x,i}^{*,s}|Z_i^*), \quad \text{and } Y_i^* = \widehat{F}_{Y|Z}^{-1}(U_{y,i}^{*,s}|Z_i^*). \tag{3.2}$$

With this bootstrap sample we compute the pure and conditional frontiers (of order-m) at the points of interest, and repeating the procedure a large number of times we obtain the bootstrap approximation of the sampling distribution of these objects.

When full frontiers are considered, we only prove weak consistency, but as conjectured in Florens et al. (2014), if the functions φ_ℓ in (1.10) are sufficiently smooth, the FDH rates of convergence would be maintained for $\widehat{\phi}(u_y)$ and $\widehat{\tau}(y, z)$.

4 Numerical Illustration

We will first illustrate on 3 simple examples how our method performs compared to the direct nonparametric method and compared to the location-scale method. Then in a more realistic example we will also consider order-m estimators and investigate the effect of the presence of outliers. In all the numerical examples below, the optimal bandwidths have been computed for each sample by least-squares cross validation, for each nonparametric regressions in the location-scale approach and for each estimation of conditional distribution functions.

4.1 A "Toy" example

We start with what we could call a "Toy" example which is a very simple classic model of frontier where the external factor Z only influences the density of the inefficiencies (see e.g. Kumbhakar and Lovell, 2000). The cost model is given by

$$X = 1 + Y^2 + \xi, \tag{4.1}$$

where $Y \sim 3\text{Beta}(1, 2)$ and $\xi|Z = z \sim \text{Exp}(3/(z + 1))$ and $Z \sim \text{Unif}(0, 4)$. So here the mean conditional efficiency is given by $\mu_\xi(z) = (z + 1)/3$ and the true cost conditional frontier $\tau(y, z) = \tau(y) = 1 + y^2$ since Z only influences the inefficiency distribution. Note that here this simple traditional frontier model does not fit the location-scale assumptions given in (1.9). We simulate i.i.d. values (X_i, Y_i, Z_i), $i = 1, \ldots, n$ according to this model.

Now the idea is to see if we can correctly estimate the true values $\tau(Y_i, Z_i)$. We will compare below the performance of our estimator compared with the traditional, direct one and the location-scale method. For doing so we can look, for each approach, to the estimates of the ISE (Integrated Squared Error) given by

$$ISE = n^{-1} \sum_{i=1}^{n} (\widehat{\tau}(Y_i, Z_i) - \tau(Y_i, Z_i))^2. \tag{4.2}$$

[6] To generate smoothed bootstrap values of $U_{\ell,i}^{*,s}$, for $\ell = x, y$ in $[0, 1]$ we can proceed as follows: define for $\ell = x, y$, $U_{\ell,i}^{*,s} = \Phi[\Phi^{-1}(U_{\ell,i}^*) + h_n \epsilon_{\ell,i}^*]$, where $(U_{x,i}^*, U_{y,i}^*)$ is a simple random draw with replacement in the set $(\widehat{U}_{x,i}, \widehat{U}_{y,i})$, $i = 1, \ldots, n$, $h_n = 1.06 n^{-1/5}$ is the optimal bandwidth for estimating a $N(0, 1)$ density, $\epsilon_{\ell,i}^* \sim N(0, 1)$ and $\Phi(\cdot)$ is the $N(0, 1)$ cdf.

We estimate the Mean Integrated Squared Error ($MISE$) by doing 500 Monte-Carlo simulations according to our DGP and averaging the ISE over the 500 trials. To check if some differences are significant we give also the estimates of the standard deviation of the Monte-Carlo estimator of the $MISE$. We do this for $n = 100, 200$ and 500. The results for this "Toy" example are reported in Table 1, where "Direct" indicates the results for the traditional method, "New Method" for our approach and the "Loc-Scale" for the approach developed in Florens et al. (2014). A quick analysis of the table reveals that for each approach the $MISE$ decreases, as it should, when the sample size increases. We also see that our new approach dominates the two others. The direct traditional approach suffers from the problems mentioned above after equation (1.8) and already pointed by Florens et al. (2014). Note that when the sample increases, the direct approach gives similar performance as our new approach, illustrating the fact that our new approach does not improve the rate of convergence of the estimators of the full frontier. However here the location-scale model is inappropriate, since the model does not fit the needed assumptions, providing in the table (bottom line) the worst case for this simple naïve model.

Table 1: $MISE$ for the simple "Toy" example.

	$n = 100$	$n = 200$	$n = 500$
Direct	0.2459	0.1749	0.1001
(std)	(0.0080)	(0.0058)	(0.0026)
New Method	0.1837	0.1407	0.1098
(std)	(0.0045)	(0.0030)	(0.0022)
Loc-Scale	0.3187	0.2491	0.1948
(std)	(0.0081)	(0.0065)	(0.0040)

4.2 A full independence case

We redo the exercise with a case where Z is fully independent of (X, Y) and so, has no influence on the production process. We have as above $\tau(y, z) = \tau(y) = 1 + y^2$ with $Y \sim 3\text{Beta}(1, 2)$ and $Z \sim \text{Unif}(0, 4)$ and however we select the cost according to $X = \tau(Y) \times \exp(\xi)$ where $\xi \sim N^+(0, 0.5^2)$ a truncated normal independent of Z. Obviously this model fits the location-scale assumptions. The results for the $MISE$ for the 3 estimators are displayed in Table 2. Here again, our new approach dominates the two other ones, note that the new method even performs better than the location-scale approach despite the fact that the latter is appropriate. This comes probably from the fact that in the location-scale approach, we first estimate the location by a local linear estimator then in a second step, we regress on Z, in a nonparametric way, the squares of the residuals obtained when estimating the location function. Squaring the residuals introduces some instability in the estimation process. But still, the location-scale approach dominates the traditional direct approach. Note that the Monte-Carlo standard deviations indicate that the differences are significant.

Table 2: $MISE$ for the full independence case example.

	$n = 100$	$n = 200$	$n = 500$
Direct	1.6502	0.9712	0.5651
(std)	(0.0762)	(0.0408)	(0.0271)
New Method	0.7281	0.4480	0.2339
(std)	(0.0227)	(0.0151)	(0.0078)
Loc-Scale	1.1930	0.6333	0.3178
(std)	(0.0606)	(0.0287)	(0.0147)

4.3 A Location-Scale model

In this model we try to simulate a more general model that fits the location-scale assumptions. The objective is to explore the performance of our new approach in scenarios where the location-scale model is applicable. The model is a bit artificial but indeed the location-scale assumptions in (1.9) are less natural in a frontier model setup, even if these models can be viewed as flexible approximations of the DGP. The model goes along the following lines.

As above, we choose $Y \sim 3\text{Beta}(1, 2)$ independent of $Z \sim \text{Unif}(0, 4)$, so $\mu_y(Z) = 1$ and $\sigma_y(Z) = \sqrt{2}/2$ and the location-scale model is true for Y. Let the random variable W be defined as $W = (1+Y)+\xi$ where $\xi \sim \text{Exp}(1)$ will govern the cost inefficiencies. Let us define $X = \mu_x(Z)+\sigma_x(Z)\varepsilon_x$ for some functions of Z, $\mu_x(Z)$ and $\sigma_x(Z)$ where $\varepsilon_x = (W-\mathbb{E}(W))/\sigma(W)$ is now standardized as required in (1.9). In our case, this defines $\varepsilon_x = (Y - 2 + \xi)/\sqrt{3/2}$. So we can rewrite the cost X as

$$X = \mu_x(Z) + \sigma_x(Z) \left[\frac{Y - 2}{\sqrt{3/2}} \right] + \sigma_x(Z)\xi/\sqrt{3/2}, \tag{4.3}$$

where the first two terms identify the true frontier in original units $\tau(y, z)$ and the last term is the inefficiency and is positive. In our example here we chose $\mu_x(Z) = 2(Z + 1)/6$ and $\sigma_x(Z) = (Z + 1)/6$. The results of the Monte-Carlo experiment for this example are in Table 3. Here again, and surprisingly, our new approach dominates significantly the location-scale approach of Florens et al. (2014), probably for the same reasons explained above in the full independence case. But of course the location-scale method dominates the traditional direct approach.

Table 3: $MISE$ for the Location-Scale model example.

	$n = 100$	$n = 200$	$n = 500$
Direct	0.0410	0.0289	0.0176
(std)	(0.0013)	(0.0009)	(0.0004)
New Method	0.0181	0.0109	0.0054
(std)	(0.0004)	(0.0003)	(0.0001)
Loc-Scale	0.0240	0.0141	0.0063
(std)	(0.0008)	(0.0004)	(0.0002)

4.4 More elaborate realistic examples

Inspired by the DGPs proposed by Bădin et al. (2019) and Simar and Wilson (2011), we aim to Illustrate how various approaches perform under different conditions where the external variable Z influences the production process. Specifically, the effect of Z may be observed on the level of the frontier, on the distribution of inefficiencies, on both, or on neither. So we consider in Appendix B of the supplementary material these four cases in various scenarios with one input X (X is a cost), and one output Y for uni- and bi-variate Z. We illustrate also in Appendix B the effect of Z on the DGP by various figures in the case of univariate Z in 3D figures.

Here we only focus on one case where Z is bivariate and influences both the frontier level and the inefficiency distribution (this is case-D in the Appendix). We also analyse the consequence of adding 3 outliers in the data set. We will look to the behavior of the order-m partial frontiers, we select, only for illustration, the values of $m = 10, 20, 50, 100$ and 200. The true values of the order-m frontiers are difficult to compute. For some given point, they can be obtained by numerical integration in y to get the true survival function $S_{X|YZ}(x \mid Y \geq y, Z = z)$ and then by integrating on x for computing $\tau_m(y, z)$ in each scenario (see (1.12)). This would be infeasible for all the data points, for all the cases and for each Monte-Carlo trial. So we investigate rather how the order-m frontier estimates the true full frontier. This will be of particular interest for investigating the cases where we added outliers. All the detailed tables of $MISE$ for various scenarios are available in the supplementary material.

To summarize the scenario presented here, we have a basic "cost" frontier $\tau(y) = 1 + y^2$ for $y \in (0, 3)$ and $Z_j \sim \text{Unif}(0, 4)$, independently for $j = 1, 2$. Then we consider that, as often the case in practice, the output Y may depend on Z_1. We define

$$Y \mid Z_1 = z_1 \sim 3\text{Beta}(1 + z_1/4, 2 - z_1/4) \tag{4.4}$$

where $\text{Beta}(a, b)$ is the beta distribution. So depending on the value of z_1, the shape of $f(y|Z_1 = z_1)$ give more weights to values of Y near its lower bound for small z_1 and near its upper bound for large values of z_1. For instance we have $\mathbb{E}(Y \mid Z_1 = z_1) = 1 + z_1/4$. We see that Z_1 may influence not only the location and the scale of the distribution of Y but also moments of higher order. So the location-scale model is not verified for Y. The inefficiency will be governed by a multiplicative factor $\exp(\xi) \geq 1$ to the frontier function, where $\xi \geq 0$. We will choose

$$\xi \mid Z = z \sim N^+(\mu_\xi(z), \sigma_\xi^2(z)), \tag{4.5}$$

where

$$\mu_\xi(z) = 1 - (z_1 + z_2)/2 \tag{4.6}$$

$$\sigma_\xi(z) = ((z_1 + z_2)/2 + 3)/10. \tag{4.7}$$

As above we did 500 MC replications and we look at the $MISE$ for samples of size $n = 100, 200$ and 500. The results are displayed in Table 4. The results speak by themselves: as the simpler examples above our new method dominates the location-scale method and also the direct approach. Note that here for this complex case, where Z affects both the frontier and the efficiency distribution, the location-scale is not correct and the approach gives even worse results than the direct traditional method. Here, there are no outliers but interestingly the order-m estimators provide better estimates of the full frontier from values of m near 50. This is because they are less sensitive to extreme data points that can jeopardize the full frontier estimates. Note also that our method also dominates the two other ones for the order-m estimates.

Table 4: $MISE$ for the simulated examples, Bivariate Z, Case D: Z_1 does affect the frontier, and Z_1 and Z_2 affect both the efficiency distribution.

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.4624	1.3764	1.2837	1.3325	1.3815	1.4169
(std)	(0.1510)	(0.1529)	(0.1518)	(0.1509)	(0.1508)	(0.1509)
New Method	0.7701	0.9894	0.8045	0.7675	0.7689	0.7700
(std)	(0.0264)	(0.0268)	(0.0257)	(0.0261)	(0.0263)	(0.0264)
Loc-Scale	2.1966	2.5026	2.0952	2.0129	2.0885	2.1649
(std)	(0.1836)	(0.1733)	(0.1637)	(0.1637)	(0.1719)	(0.1801)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.0943	1.2377	1.0373	1.0120	1.0348	1.0579
(std)	(0.0569)	(0.0590)	(0.0581)	(0.0575)	(0.0573)	(0.0572)
New Method	0.5825	0.8749	0.6306	0.5704	0.5747	0.5805
(std)	(0.0161)	(0.0182)	(0.0162)	(0.0158)	(0.0159)	(0.0161)
Loc-Scale	1.6307	2.1291	1.6425	1.4679	1.4934	1.5607
(std)	(0.0761)	(0.0684)	(0.0649)	(0.0650)	(0.0676)	(0.0714)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	0.7965	1.1906	0.8815	0.7788	0.7706	0.7764
(std)	(0.0372)	(0.0397)	(0.0388)	(0.0381)	(0.0378)	(0.0375)
New Method	0.5089	0.8962	0.5754	0.4767	0.4793	0.4937
(std)	(0.0129)	(0.0143)	(0.0122)	(0.0119)	(0.0123)	(0.0126)
Loc-Scale	1.3869	2.0295	1.4525	1.1892	1.1695	1.2228
(std)	(0.0608)	(0.0473)	(0.0432)	(0.0421)	(0.0435)	(0.0466)

Next we introduce in the same scenario three outliers. We replaced the 3 last simulated values of (X_i, Y_i, Z_i) in the preceding samples by fixing $Y_i = 1, 1.75$ and 2, keeping the same 3 simulated values of Z and then, for the 3 fixed points Y_i, we define X_i below the true cost frontier at a level of 50% of the true value. The results are displayed in Table 5.

We observe indeed that the order-m frontiers are resistant to these outliers when m is not too large (otherwise the order-m frontiers and estimates converge to the full frontiers, see e.g. Cazals et al. (2002)). But again our new method performs better than the two others and in particular than the location-scale approach which is inappropriate. In this scenario, it seems that an optimal order-m would be around 10 or 20 depending on the value of n. In practice for a real data example, several techniques have been proposed to select an appropriate value of m for a given sample (see e.g. Daouia and Gijbels, 2011).

4.5 Conclusions from these Monte-Carlo experiments

To summarize, we observed in our various scenarios (including all the cases in the Appendix B) that our new method based on the control functions approach dominates, as expected,

Table 5: $MISE$ for the simulated examples, Bivariate Z, Case D: Z_1 does affect the frontier, and Z_1 and Z_2 affect both the efficiency distribution, with 3 outliers.

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.4774	1.6725	1.8832	2.1840	2.3239	2.3990
(std)	(0.1509)	(0.1503)	(0.1497)	(0.1501)	(0.1506)	(0.1508)
New Method	2.2479	1.0634	1.3265	1.8663	2.1312	2.2292
(std)	(0.0340)	(0.0269)	(0.0286)	(0.0316)	(0.0334)	(0.0339)
Loc-Scale	3.6885	2.3308	2.4772	3.0768	3.4495	3.6351
(std)	(0.1981)	(0.1603)	(0.1562)	(0.1652)	(0.1802)	(0.1930)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.8295	1.3545	1.3663	1.5596	1.6790	1.7499
(std)	(0.0574)	(0.0582)	(0.0572)	(0.0567)	(0.0568)	(0.0571)
New Method	2.0350	0.8016	0.8282	1.2787	1.6700	1.9216
(std)	(0.0271)	(0.0163)	(0.0176)	(0.0216)	(0.0247)	(0.0265)
Loc-Scale	3.1665	1.8353	1.6858	2.1433	2.6255	2.9702
(std)	(0.0824)	(0.0607)	(0.0598)	(0.0652)	(0.0710)	(0.0766)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.2243	1.1636	0.9662	1.0121	1.0906	1.1501
(std)	(0.0348)	(0.0383)	(0.0370)	(0.0358)	(0.0352)	(0.0349)
New Method	1.8634	0.7942	0.5681	0.7295	1.0538	1.4278
(std)	(0.0234)	(0.0136)	(0.0126)	(0.0155)	(0.0186)	(0.0209)
Loc-Scale	2.8680	1.8621	1.4008	1.4656	1.8214	2.2622
(std)	(0.0704)	(0.0478)	(0.0441)	(0.0453)	(0.0493)	(0.0548)

the traditional direct approach defined by (1.7) and (1.14) but also, in most of the cases, the location-scale model, which is too restrictive since Z can only influence the first two moments of the input and the outputs. But even if the location-scale model is true, we have seen in our scenarios above that the new method has better $MISE$ performances, we pointed above that this may come from the second stage nonparametric regression for estimating the scale functions $\sigma(Z)$ (squaring the residuals of the first stage regression introduces more instability). We see also that our new method has in most of the cases, the smallest Monte-Carlo standard deviation of the $MISE$ indicating a greater statistical stability. These remarks apply also to the order-m estimates.

5 Real Data Illustration

To illustrate our method with a real data set, we use data from the banking sector, also used in Bădin et al. (2012) applying the direct traditional approach and in Florens et al. (2014), implementing the location-scale model to clean the input (a cost) and the outputs. The aim here is to show to practitioners the kind of useful results that can be exploited by using our approach to estimate conditional cost frontiers. It is safer to opt for our control function approach as described in (1.10), since, as discussed above, it allows more general structure for the effects of environmental variables compared to the location-scale models. Also the simulation section above indicates that our method is much more stable and dominates the other two.

The original dataset comes from Simar and Wilson (2007) and contains 3 inputs (purchased funds, core deposits and labor) and 4 outputs (consumer loans, business loans, real estate loans and securities held) for banks. Two environmental factors are considered, the size of the banks Z_1 (the log of the total assets = SIZE) and a measure of the diversity of the services proposed by the banks Z_2 (DIVERSITY). Daraio et al. (2018), using the same data set, rejected the separability condition, advocating the use of conditional efficiency measures.

We select a subsample of 303 banks, as in Simar and Wilson (2007), Bădin et al. (2012) and Florens et al. (2014). In the two latter papers, it is explained that, by using the methodology described in Daraio and Simar (2007), the inputs can be aggregated in a one-dimensional input measure, without losing much information and the same is true for the outputs. The final output Y is highly correlated (more than 0.93) with the 4 original outputs and the same is true for the final input X (correlation with the original inputs more than 0.97). This allows to illustrate the results in low-dimensional pictures.

Figure 1 displays the values of the 303 banks in "pure" units, whitened from the influence of (Z_1, Z_2) and the estimated full frontier $\widehat{\phi}(\hat{u}_y)$ and the order-m frontier $\widehat{\phi}_m(\hat{u}_y)$, with $m = 30$. We see that the full frontier in Figure 1 envelops nicely the cloud of data in pure units, with the same qualitative remark for the order-m frontier. We see also that with $m = 30$ we have only very few points below the m-frontier.[7]

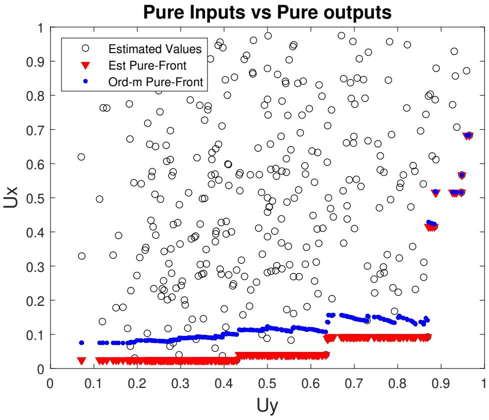

Figure 1: Bank example: Estimated "pure" inputs and outputs and the estimated efficient frontiers $\widehat{\phi}(\hat{u}_y)$ and $\widehat{\phi}_m(\hat{u}_y)$, here $m = 30$.

Finally, for practitioners, a detailed analysis of the individual efficiency scores (in pure units and conditional on z) would be very informative, in particular to detect some inefficient units and measure how far they are from the minimal cost frontiers in original units (full and of order-m). For the bootstrap confidence intervals we used the procedure described in (3.1) and (3.2). In Tables 6 and 7 we display the results for 15 randomly chosen banks. In Table 6, we only display the results for the efficiencies in pure units, i.e $\widehat{\delta}_i = \hat{u}_{x_i} - \widehat{\phi}(\hat{u}_{y_i})$ and $\widehat{\delta}_{i,m} = \hat{u}_{x_i} - \widehat{\phi}_m(\hat{u}_{y_i})$. For instance, we see that the unit "237" has rank 1 and lies on the efficient full frontier, then as expected it lies below the cost frontier of order-m. The 95%

[7] As in Florens et al. (2014), we selected $m = 30$ just for illustration. This provides, in our sample, less than 6% of data points below the order-m cost frontier.

confidence bounds for this unit indicate it is significantly below the order-m cost frontier, so it behaves like a super-efficient unit. Unit "170" is one of the worst with respect to his cost.

Table 6: Pure and conditional efficiency scores of order-m, for 15 randomly selected banks. The ranks R_i are computed relative to the pure order-m efficiency scores with $m = 30$. The 95% confidence bounds for $\delta_{m,i}$ are computed by bootstrap ($B = 1000$ replications).

Unit i	$\widehat{\delta}_i$	R_i	$\widehat{\delta}_{m,i}$	low_{δ_m}	up_{δ_m}
259	0.1439	41	0.1017	0.0171	0.2305
237	0.0000	1	-0.0745	-0.1275	-0.0487
258	0.1496	37	0.0743	0.0200	0.0967
1	0.3617	120	0.2848	0.2168	0.3086
241	0.4089	150	0.3514	0.3132	0.3699
66	0.5870	218	0.5225	0.4783	0.5495
164	0.3177	104	0.2478	0.1857	0.2849
274	0.5396	206	0.4826	0.4460	0.5078
303	0.1879	53	0.1378	0.0413	0.2341
199	0.2567	107	0.2558	-0.2063	0.5442
216	0.7948	271	0.7154	0.6439	0.7489
125	0.3263	108	0.2562	0.1954	0.3118
239	0.4024	145	0.3354	0.2890	0.3633
170	0.9079	297	0.8274	0.7543	0.8571
242	0.0150	7	-0.0438	-0.0825	-0.0214

In Table 7, we provide useful results in the original units. Comparing the cost of a bank (column X_i) with the full cost frontier (the column $\widehat{\tau}_i$), we can appreciate the gap (excess of cost) that each unit has to do for reaching the optimal cost function. Being less extreme, the order-m frontier provides a less severe benchmark (column $\widehat{\tau}_{m,i}$). Also we provide an efficiency measures in proportion, i.e. $\widehat{\theta}_i = \widehat{\tau}_i / X_i$, which gives the percentage of reduction of the cost a bank should perform to achieve the optimal cost function. Their order-m versions $\widehat{\theta}_{m,i} = \widehat{\tau}_{m,i}/X_i$ are less severe. For points below the order-m cost frontier, this number can exceed 1 (as for the super-efficient unit "237"). Again, for the order-m objects, we can provide confidence intervals by bootstrap methods, as given in the Table.

Table 7: Results for the same 15 banks in original units. X_i is the observed cost, $\widehat{\tau}_i$ is its full frontier estimate with $\widehat{\tau}_{m,i}$ for the order-m with $m = 30$. The proportionate cost reductions are given by $\widehat{\theta}_i$ and $\widehat{\theta}_{m,i}$. 95% confidence bounds are for the order-m objects.

Unit i	X_i	$\widehat{\tau}_i$	$\widehat{\tau}_{m,i}$	low_{τ_m}	up_{τ_m}	$\widehat{\theta}_i$	$\widehat{\theta}_{m,i}$	low_{θ_m}	up_{θ_m}
259	7.2986	6.2393	6.2432	5.0774	7.2337	0.8549	0.8554	0.6957	0.9911
237	0.3505	0.3505	0.4047	0.3573	0.4616	1.0000	1.1548	1.0196	1.3173
258	0.1998	0.1604	0.1830	0.1725	0.2433	0.8026	0.9157	0.8632	1.2177
1	1.1985	1.0134	1.0565	1.0156	1.1909	0.8456	0.8816	0.8474	0.9937
241	0.8693	0.6903	0.7254	0.6714	0.8077	0.7940	0.8344	0.7723	0.9291
66	0.3421	0.2396	0.2592	0.2413	0.3164	0.7005	0.7577	0.7055	0.9249
164	1.8694	1.6086	1.6930	1.5929	1.9406	0.8605	0.9056	0.8521	1.0381
274	0.4026	0.2715	0.2990	0.2842	0.3646	0.6743	0.7428	0.7061	0.9057
303	0.2969	0.2625	0.2725	0.2456	0.3324	0.8841	0.9177	0.8273	1.1196
199	2.6751	2.5152	2.5158	2.3157	3.0799	0.9402	0.9404	0.8656	1.1513
216	7.2741	5.7449	5.8366	5.4627	7.0507	0.7898	0.8024	0.7510	0.9693
125	1.0559	0.8559	0.9117	0.8257	0.9934	0.8106	0.8634	0.7820	0.9408
239	1.3945	1.1799	1.2690	1.2240	1.3916	0.8461	0.9100	0.8778	0.9980
170	2.9572	1.9072	2.1360	1.9677	2.3823	0.6449	0.7223	0.6654	0.8056
242	1.8388	1.8138	1.9190	1.7612	2.0793	0.9864	1.0436	0.9578	1.1308

It is quite interesting to analyze the shape of the frontier in the (Y, X) space for fixed values of the environmental conditions. So, we selected 9 pairs $(QZ_{1k}, QZ_{2\ell})$, for $k, \ell = 1, 2, 3$, for each quartile QZ_{1k} of Z_1 and $QZ_{2\ell}$ of Z_2. In Figure 2, we see the case for Z_2 fixed at its median and the 3 different quartiles for Z_1. The conditional frontiers did not change so much when changing the value of Z_2, so to save space we do not display these figures. We see again that with our approach the frontiers envelop nicely the cloud of points. The picture indicates, as expected, that the SIZE (i.e. Z_1) impacts the level of the frontier.

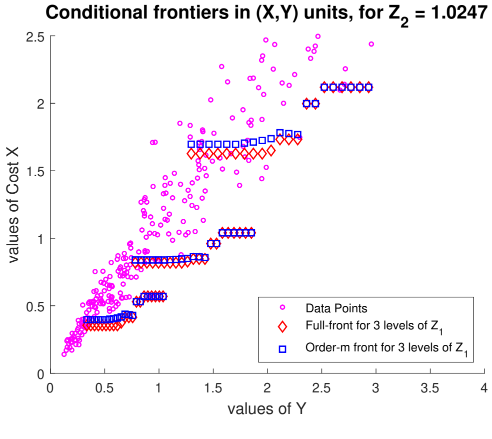

Figure 2: Bank example: data points in original units and frontier estimates when fixing the level of Z. Here Z_2 (DIVERSITY) is fixed at its median value, and Z_1 (SIZE) is fixed at its 3 quartiles (from the left to the right).

A final point of interest is to investigate how the support of (Y, X) is impacted by the values of Z. This is usually done in this literature by looking to the ratios of $\widehat{\tau}(y, z)/\widehat{\tau}(y)$

as a function of z for fixed values of Y defined at its 3 quartiles. This is displayed Figure 3. We provide the surface for the 3 quartiles of Y, because even if the global shape of the surfaces is comparable, we can detect slight different behavior for large or small values of Y. For the middle panel, Y equal its median, the effect of the SIZE (Z_1) seems not being important and the effect of DIVERSITY (Z_2) seems slightly more important for small values of Z_1. We see also that the effect of Z_2 is slightly more important for smaller banks (Z_1 small) than for large banks at each quartile of Y. For the upper quartile of Y this effect seems to be slightly reversed: for small banks with low level of output (up to the median of Y) too much diversity seems to be favourable (smaller level for the conditional cost frontier compared to the marginal one), but this effect seems to disappear for banks having the largest output.

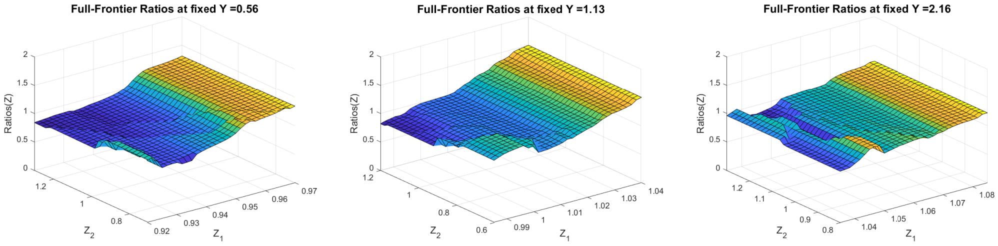

Figure 3: Bank example: Analysis of the ratios $\widehat{\tau}(y,z)/\widehat{\tau}(y)$ for fixed values of y. From left to right y is given by the 3 quartiles of Y. Z_1 is the SIZE and Z_2 is DIVERSITY.

6 Conclusions

Conditional frontiers and conditional efficiency measures are useful tools for the practitioner to investigate the role of environmental variables on a production process. The traditional, direct, approach implies the estimation of the support of appropriate conditional survival function, implying the selection of optimal bandwidths. The latter is often obtained by least-squares cross validation techniques which are optimal for estimating a survival function over its full range, but, as pointed in Bădin et al. (2019) not necessarily optimal for estimating its boundary.

We propose in this paper a way to overcome this difficulty without imposing any additional assumptions on the model. The idea is to pre-whiten the variables (input and outputs) from the influence of these environmental factors in a first nonparametric stage by control functions type model. By doing so, we first produce a version of pure input and outputs, which can be viewed as part of the original variables independent of the environmental factors, allowing to measure “pure” or “managerial” efficiency measures, more appropriate to rank the firms. Second, we are able to recover the frontiers, full and robust versions, in the original units. We prove that for the order-m robust frontier in pure units, the derived estimator does not suffer from the curse of dimensionality due to the dimension of Z, proving the $\sqrt{n}$ convergence to a Gaussian process. For this frontier in original units, we keep the nice properties of its correspondent obtained by traditional methods (i.e. $\sqrt{nh_z}$ convergence to a Gaussian process). This is similar to the results obtained by Florens et al. (2014), but in our case we do not have to rely to the restricted location-scale assumptions, which assume that Z

can only influence the mean and the scale of the input and outputs, and not other shape parameters. Our fully nonparametric approach, characterized by its non-separability in Z, enables the practitioners to analyse the intricate, non-separable connections between these variables and the target outcomes, avoiding assuming any specific functional forms governing these relationships. Hence, it is very important in applied analysis to aim at exploring the complexities inherent in the relationships under investigation.

Various Monte-Carlo experiments demonstrate that our approach yields much more reliable estimators of the true frontier (full and robust order-m) when compared to the direct approach, but even compared with the location-scale approach. When the location-scale assumptions are not verified the latter approach can provide very poor results, as indicated in our experiments. Finally we illustrate the practical use of our approach in a real data set, showing its great flexibility and usefulness.

References

Bădin, L., C. Daraio, and L. Simar (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research 223*, 818–833.

Bădin, L., C. Daraio, and L. Simar (2019). A bootstrap approach for bandwidth selection in estimating conditional efficiency measures. *European Journal of Operational Research 277*, 784–797.

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

Daouia, A. and I. Gijbels (2011). Robustness and inference in nonparametric partial frontier modeling. *Journal of Econometrics 161*, 147–165.

Daraio, C. and L. Simar (2005). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis 24*(1), 93–121.

Daraio, C. and L. Simar (2007). *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*. New York: Springer.

Daraio, C., L. Simar, and P. Wilson (2018). Central limit theorems for conditional efficiency measures and tests of the "separability" condition in nonparametric, two-stage models of production. *Econometrics Journal 21*, 170–191.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica 19*(3), 273–292.

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

Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A 120*, 253–281.

Florens, J., L. Simar, and I. Van Keilegom (2014). Frontier estimation in nonparametric location-scale models. *Journal of Econometrics 178*, 456–470.

Guerre, E. and C. Sabbah (2012). Uniform bias study and bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory 28*, 87–129.

Imbens, G. and W. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*, 1481–1512.

Jeong, S., B. Park, and L. Simar (2010). Nonparametric conditional efficiency measures: asymptotic properties. *Annals of Operations Research 173*, 105–122.

Kumbhakar, S. and C. Lovell (2000). *Stochastic Frontier Analysis.* Cambridge University Press.

Li, Q., J. Lin, and J. Racine (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics 31*(1), 57–65.

Matzkin, R. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica 71*, 1339–1375.

Park, B., L. Simar, and C. Weiner (2000). The fdh estimator for productivity efficiency scores: Asymptotic properties. *Econometric Theory 16*, 855–877.

Simar, L., A. Vanhems, and I. Van Keilegom (2016). Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *Journal of Econometrics 190*, 360–373.

Simar, L. and P. Wilson (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics 136*(1), 31–64.

Simar, L. and P. Wilson (2011). Two-stage dea: Caveat emptor. *Journal of Productivity Analysis 36*, 205–218.

Simar, L. and P. Wilson (2015). Statistical approaches for nonparametric frontier models: A guided tour. *International Statistical Review 83*(1), 77–110.

Van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes.* New York: Springer.

A Appendix: Proofs

The proofs rely on two lemmas. We start with a lemma regarding the estimators $\widehat{F}_{Y|Z}(y|z)$ and $\widehat{F}^{-1}_{Y|Z}(u|z)$, that will be used multiple times in the proofs of the main results. The result is e.g. given in Theorems 2.4 and 4.1 in Li et al. (2013). Similar results hold for Y replaced by X.

Lemma A.1 *Assume (C1)-(C5). Then,*

$$\widehat{F}_{Y|Z}(y|z) - F_{Y|Z}(y|z) = (nh_z)^{-1} \sum_{i=1}^{n} k\Big(\frac{Z_i - z}{h_z}\Big) \frac{I\!I(Y_i \le y) - F_{Y|Z}(y|Z_i)}{f_Z(z)} \tag{A.1}$$
$$+ h_y^2 B_y(y|z) + h_z^2 B_z(y|z) + O_P((nh_z)^{-1}\log n),$$

and

$$\widehat{F}^{-1}_{Y|Z}(u|z) - F^{-1}_{Y|Z}(u|z) = -(nh_z)^{-1} \sum_{i=1}^{n} k\Big(\frac{Z_i - z}{h_z}\Big) \frac{I\!I(Y_i \le F^{-1}_{Y|Z}(u|Z_i)) - u}{f_{Y|Z}(F^{-1}_{Y|Z}(u|z)|z) f_Z(z)} \tag{A.2}$$
$$- h_y^2 \frac{B_y(F^{-1}_{Y|Z}(u|z)|z)}{f_{Y|Z}(F^{-1}_{Y|Z}(u|z)|z)} - h_z^2 \frac{B_z(F^{-1}_{Y|Z}(u|z)|z)}{f_{Y|Z}(F^{-1}_{Y|Z}(u|z)|z)} + O_P((nh_z)^{-1}\log n),$$

uniformly in $u \in [0,1], y \in R_Y, z \in R_Z$, where $B_y(y|z) = \frac{\kappa_2}{2}\frac{\partial^2}{\partial y^2}F_{Y|Z}(y|z)$, $B_z(y|z) = \frac{\kappa_2}{2}\frac{\partial^2}{\partial z^2}F_{Y|Z}(y|z)$, and $\kappa_2 = \int k(u)u^2du$.

Lemma A.2 *Assume (C1)-(C5). Then,*

$$\sup_{0 \le u_y \le 1} \Big| n^{-1} \sum_{i=1}^{n} \big\{ I\!I(\widehat{U}_{y,i} \ge u_y) - I\!I(U_{y,i} \ge u_y) \big\} - \mathbb{P}(\widehat{U}_y \ge u_y) + \mathbb{P}(U_y \ge u_y) \Big| = o_P(n^{-1/2}),$$

and

$$\sup_{0 \le u_x, u_y \le 1} \Big| n^{-1} \sum_{i=1}^{n} \big\{ I\!I(\widehat{U}_{x,i} \ge u_x, \widehat{U}_{y,i} \ge u_y) - I\!I(U_{x,i} \ge u_x, U_{y,i} \ge u_y) \big\}$$
$$- \mathbb{P}(\widehat{U}_x \ge u_x, \widehat{U}_y \ge u_y) + \mathbb{P}(U_x \ge u_x, U_y \ge u_y) \Big| = o_P(n^{-1/2}),$$

where $\mathbb{P}(\widehat{U}_y \ge u_y)$ and $\mathbb{P}(\widehat{U}_x \ge u_x, \widehat{U}_y \ge u_y)$ are the survival function of $\widehat{U}_y$ and $(\widehat{U}_x, \widehat{U}_y)$, respectively, conditional on the data.

Proof. We will show the first statement, the second one can be shown in a similar way. Define the class of functions

$$\mathcal{F} = \Big\{ (y,z) \to I\!I(F(y|z) \ge u) : u \in [0,1], F(\cdot|z) \text{ monotone onto } [0,1]$$
$$\text{for all } z \in R_Z, F^{-1}(u|\cdot) \in C_M^{1+\delta}(R_Z) \text{ for all } u \in [0,1] \Big\},$$

where δ, M and $C_M^{1+\delta}(R_Z)$ are defined in condition (C5). We will first show that $\mathcal{F}$ is Donsker. Since the function $z \to F^{-1}(u|z) \in C_M^{1+\delta}(R_Z)$ for all $u \in [0,1]$ and all F satisfying the above conditions, there exist ϵ^2-brackets $b_j^L \le b_j^U$, $j = 1, \ldots, M_\epsilon = O(\exp(\epsilon^{-2/(1+\delta)}))$, such that for a given u and F there exists $1 \le j \le M_\epsilon$ satisfying $b_j^L(\cdot) \le F^{-1}(u|\cdot) \le b_j^U(\cdot)$ (see Corollary 2.7.2 in Van der Vaart and Wellner, 1996). Hence, we also have that

$$I\!I(b_j^U(\cdot) \le y) \le I\!I(F^{-1}(u|\cdot) \le y) = I\!I(F(y|\cdot) \ge u) \le I\!I(b_j^L(\cdot) \le y),$$

where the equality holds thanks to the monotonicity of F in the first argument. Next, calculate

$$\int \big[1\!\!1(b_j^L(z) \leq y) - 1\!\!1(b_j^U(z) \leq y)\big]^2 dF_{Y,Z}(y,z) = \int \big[F_{Y|Z}(b_j^U(z)|z) - F_{Y|Z}(b_j^L(z)|z)\big] dF_Z(z)$$
$$\leq K\|b_j^U - b_j^L\|_{L_1^{\mathbb{P}}} \leq K\|b_j^U - b_j^L\|_{L_2^{\mathbb{P}}} \leq L\epsilon^2,$$

for some $K, L < \infty$, provided $\sup_{y,z} f_{Y|Z}(y|z) < \infty$. This shows that $N_{[]}(\mathcal{F}, \epsilon, L_2(\mathbb{P})) \leq M_\epsilon$, where $N_{[]}(\mathcal{F}, \epsilon, L_2(\mathbb{P}))$ is the ϵ-bracketing number of the class $\mathcal{F}$ with respect to the $L_2(\mathbb{P})$ measure (with $\mathbb{P}$ the joint probability measure of (Y,Z)), i.e. the smallest number of ϵ-brackets needed to cover the space $\mathcal{F}$. Hence,

$$\int_0^{2M} \big(\log N_{[]}(\mathcal{F}, \epsilon, L_2(\mathbb{P}))\big)^{1/2} d\epsilon \leq K_1 \int_0^{2M} \epsilon^{-1/(1+\delta)} d\epsilon = K_2 (2M)^{\delta/(1+\delta)} < \infty,$$

for some $K_1, K_2 < \infty$. It now follows from Theorem 2.5.6 in Van der Vaart and Wellner (1996) that the class $\mathcal{F}$ is $\mathbb{P}$-Donsker, and hence

$$\lim_{\alpha\to 0}\lim_{\epsilon \to 0} \mathbb{P}\Big(\sup_{f,g\in\mathcal{F},\rho_{\mathbb{P}}(f-g)<\alpha} \Big| n^{-1/2}\sum_{i=1}^n \Big\{f(Y_i,Z_i) - g(Y_i,Z_i)$$
$$- \mathbb{E}f(Y,Z) + \mathbb{E}g(Y,Z)\Big\}\Big| > \epsilon\Big) = 0, \tag{A.3}$$

where $\rho_{\mathbb{P}}^2(f) = \text{Var} f(Y,Z)$ (see Corollary 2.3.12 in Van der Vaart and Wellner, 1996).

Next, we show that for all $0 \leq u \leq 1$, the function $(y,z) \to 1\!\!1(\widehat{F}_{Y|Z}(y|z) \geq u)$ belongs to $\mathcal{F}$ with probability tending to one. For this it suffices to show that $\sup_{0\leq u\leq 1} \|\widehat{F}_{Y|Z}^{-1}(u|\cdot) - F_{Y|Z}^{-1}(u|\cdot)\|_{1+\delta} = o_P(1)$, since we assume that $F_{Y|Z}^{-1}(u|\cdot) \in C_M^{1+\delta}(R_Z)$ for all $u \in [0,1]$. This follows from Guerre and Sabbah (2012).

Finally we calculate

$$\text{Var}\Big(1\!\!1(\widehat{F}_{Y|Z}(Y|Z) \geq u) - 1\!\!1(F_{Y|Z}(Y|Z) \geq u)\Big)$$
$$\leq \mathbb{E}\Big(1\!\!1(\widehat{F}_{Y|Z}(Y|Z) \geq u) - 1\!\!1(F_{Y|Z}(Y|Z) \geq u)\Big)^2$$
$$= \int \Big[F_{Y|Z}(\widehat{F}_{Y|Z}^{-1}(u|z)|z) - F_{Y|Z}(\widehat{F}_{Y|Z}^{-1}(u|z) \wedge F_{Y|Z}^{-1}(u|z)|z)\Big] dF_Z(z)$$
$$+ \int \Big[F_{Y|Z}(F_{Y|Z}^{-1}(u|z)|z) - F_{Y|Z}(\widehat{F}_{Y|Z}^{-1}(u|z) \wedge F_{Y|Z}^{-1}(u|z)|z)\Big] dF_Z(z)$$
$$= \sup_{0\leq u\leq 1, z\in R_Z} |\widehat{F}_{Y|Z}^{-1}(u|z) - F_{Y|Z}^{-1}(u|z)| = o(1) \quad a.s.$$

Hence, the result follows from (A.3). □

Proof of Theorem 3.1.

$$\widehat{S}_{U_x|U_y}(u_x|u_y) - S_{U_x|U_y}(u_x|u_y) \tag{A.4}$$
$$= \widehat{S}_{U_x,U_y}(u_x,u_y)\Big[\frac{1}{\widehat{S}_{U_y}(u_y)} - \frac{1}{S_{U_y}(u_y)}\Big] + \frac{1}{S_{U_y}(u_y)}\Big[\widehat{S}_{U_x,U_y}(u_x,u_y) - S_{U_x,U_y}(u_x,u_y)\Big],$$

where $\widehat{S}_{U_y}(u_y) = n^{-1}\sum_{i=1}^n \mathbb{1}(\widehat{U}_{y,i} \geq u_y)$ and $\widehat{S}_{U_x,U_x}(u_x,u_y) = n^{-1}\sum_{i=1}^n \mathbb{1}(\widehat{U}_{x,i} \geq u_x, \widehat{U}_{y,i} \geq u_y)$. It follows from Lemma A.2 that

$$
\begin{aligned}
&\widehat{S}_{U_x,U_y}(u_x,u_y) - S_{U_x,U_y}(u_x,u_y) \\
&= n^{-1}\sum_{i=1}^n \big\{ \mathbb{1}(U_{x,i} \geq u_x, U_{y,i} \geq u_y) - S_{U_x,U_y}(u_x,u_y)\big\} \\
&\quad + \mathbb{P}\Big(X \geq \widehat{F}^{-1}_{X|Z}(u_x|Z), Y \geq \widehat{F}^{-1}_{Y|Z}(u_y|Z)\Big) - \mathbb{P}\Big(X \geq F^{-1}_{X|Z}(u_x|Z), Y \geq F^{-1}_{Y|Z}(u_y|Z)\Big) + o_P(n^{-1/2}).
\end{aligned}
$$

The first term on the right hand side is a sum of i.i.d. terms. The second term equals

$$
\begin{aligned}
&\int \Big[\mathbb{P}\big(X \geq \widehat{F}^{-1}_{X|Z}(u_x|Z), Y \geq \widehat{F}^{-1}_{Y|Z}(u_y|Z)\big|Z=z\big) \\
&\qquad - \mathbb{P}\big(X \geq F^{-1}_{X|Z}(u_x|Z), Y \geq F^{-1}_{Y|Z}(u_y|Z)\big|Z=z\big)\Big]dF_Z(z) \\
&= \int \frac{\partial}{\partial t_1} S_{X,Y|Z}(t_1, F^{-1}_{Y|Z}(u_y|z)|z)\big|_{t_1=F^{-1}_{X|Z}(u_x|z)}\big[\widehat{F}^{-1}_{X|Z}(u_x|z) - F^{-1}_{X|Z}(u_x|z)\big]dF_Z(z) \\
&\quad + \int \frac{\partial}{\partial t_2} S_{X,Y|Z}(F^{-1}_{X|Z}(u_x|z), t_2|z)\big|_{t_2=F^{-1}_{Y|Z}(u_y|z)}\big[\widehat{F}^{-1}_{Y|Z}(u_y|z) - F^{-1}_{Y|Z}(u_y|z)\big]dF_Z(z) + o_P(n^{-1/2}) \\
&= \frac{\partial}{\partial u_x} S_{U_x,U_y}(u_x,u_y)\int f_{X|Z}(F^{-1}_{X|Z}(u_x|z)|z)\big[\widehat{F}^{-1}_{X|Z}(u_x|z) - F^{-1}_{X|Z}(u_x|z)\big]dF_Z(z) \\
&\quad + \frac{\partial}{\partial u_y} S_{U_x,U_y}(u_x,u_y)\int f_{Y|Z}(F^{-1}_{Y|Z}(u_y|z)|z)\big[\widehat{F}^{-1}_{Y|Z}(u_y|z) - F^{-1}_{Y|Z}(u_y|z)\big]dF_Z(z) + o_P(n^{-1/2}) \\
&= -\frac{\partial}{\partial u_x} S_{U_x,U_y}(u_x,u_y)\Big\{(nh_z)^{-1}\sum_{i=1}^n \int k\Big(\frac{Z_i - z}{h_z}\Big)\big[\mathbb{1}(X_i \leq F^{-1}_{X|Z}(u_x|Z_i)) - u_x\big]dz \\
&\quad + \int \Big[h_x^2 B_x(F^{-1}_{X|Z}(u_x|z)|z) + h_z^2 B_z(F^{-1}_{X|Z}(u_x|z)|z)\Big] f_Z(z)dz\Big\} \\
&\quad - \frac{\partial}{\partial u_y} S_{U_x,U_y}(u_x,u_y)\Big\{(nh_z)^{-1}\sum_{i=1}^n \int k\Big(\frac{Z_i - z}{h_z}\Big)\big[\mathbb{1}(Y_i \leq F^{-1}_{Y|Z}(u_y|Z_i)) - u_y\big]dz \\
&\quad + \int \Big[h_y^2 B_y(F^{-1}_{Y|Z}(u_y|z)|z) + h_z^2 B_z(F^{-1}_{Y|Z}(u_y|z)|z)\Big] f_Z(z)dz\Big\} + o_P(n^{-1/2}),
\end{aligned}
$$

where $S_{X,Y|Z}(x,y|z) = \mathbb{P}(X \geq x, Y \geq y|Z=z)$, and where the last equality follows from Lemma A.1. The latter expression can be further simplified as follows:

$$
\begin{aligned}
&-\frac{\partial}{\partial u_x} S_{U_x,U_y}(u_x,u_y)\, n^{-1}\sum_{i=1}^n \big[\mathbb{1}(F_{X|Z}(X_i|Z_i) \leq u_x) - u_x\big] \\
&-\frac{\partial}{\partial u_y} S_{U_x,U_y}(u_x,u_y)\, n^{-1}\sum_{i=1}^n \big[\mathbb{1}(F_{Y|Z}(Y_i|Z_i) \leq u_y) - u_y\big] \\
&-\frac{\partial}{\partial u_x} S_{U_x,U_y}(u_x,u_y)\int \Big[h_x^2 B_x(F^{-1}_{X|Z}(u_x|z)|z) + h_z^2 B_z(F^{-1}_{X|Z}(u_x|z)|z)\Big] f_Z(z)dz \\
&-\frac{\partial}{\partial u_y} S_{U_x,U_y}(u_x,u_y)\int \Big[h_y^2 B_y(F^{-1}_{Y|Z}(u_y|z)|z) + h_z^2 B_z(F^{-1}_{Y|Z}(u_y|z)|z)\Big] f_Z(z)dz + o_P(n^{-1/2}),
\end{aligned}
$$

On the other hand, using a similar decomposition, the first term of (A.4) equals

$$\begin{aligned}
&\widehat{S}_{U_x,U_y}(u_x,u_y)\Big[\frac{1}{\widehat{S}_{U_y}(u_y)}-\frac{1}{S_{U_y}(u_y)}\Big]\\
&=-\frac{S_{U_x,U_y}(u_x,u_y)}{S^2_{U_y}(u_y)}f_{U_y}(u_y)\,n^{-1}\sum_{i=1}^n\big[1\!\!1(F_{Y|Z}(Y_i|Z_i)\le u_y)-u_y\big]\\
&\quad-\frac{S_{U_x,U_y}(u_x,u_y)}{S^2_{U_y}(u_y)}f_{U_y}(u_y)\int\Big[h_y^2B_y(F^{-1}_{Y|Z}(u_y|z)|z)+h_z^2B_z(F^{-1}_{Y|Z}(u_y|z)|z)\Big]f_Z(z)dz\\
&\quad+o_P(n^{-1/2}).
\end{aligned}$$

Finally, note that

$$\frac{S_{U_x,U_y}(u_x,u_y)}{S^2_{U_y}(u_y)}f_{U_y}(u_y)+\frac{1}{S_{U_y}(u_y)}\frac{\partial}{\partial u_y}S_{U_x,U_y}(u_x,u_y)=\frac{\partial}{\partial u_y}S_{U_x|U_y}(u_x|u_y)$$

and $S_{U_y}(u_y)^{-1}\frac{\partial}{\partial u_x}S_{U_x,U_y}(u_x,u_y)=-f_{U_x|U_y}(u_x|u_y)$, from which the result follows. □

Proof of Theorem 3.2. Write

$$\begin{aligned}
&\widehat{\phi}_m(u_y)-\phi_m(u_y)\\
&=\int_0^1\big[\widehat{S}^m_{U_x|U_y}(u_x|u_y)-S^m_{U_x|U_y}(u_x|u_y)\big]\,du_x\\
&=2^{m-1}\int_0^1 S^{m-1}_{U_x|U_y}(u_x|u_y)\big[\widehat{S}_{U_x|U_y}(u_x|u_y)-S_{U_x|U_y}(u_x|u_y)\big]\,du_x+o_P(n^{-1/2}),
\end{aligned}$$

since $a^m-b^m=(a-b)(\sum_{k=0}^{m-1}c_ka^{m-1-k}b^k)$ for any a,b and for certain coefficients c_k that are such that $\sum_{k=0}^{m-1}c_k=2^{m-1}$. The result now follows from the first part of Theorem 3.1. □

Proof of Theorem 3.3. Write

$$\begin{aligned}
&\widehat{\tau}_m(y,z)-\tau_m(y,z)\\
&=-\int_0^1\widehat{F}^{-1}_{X|Z}(t|z)d\widehat{S}^m_{U_x|U_y}(t|\hat{u}_y)+\int_0^1F^{-1}_{X|Z}(t|z)dS^m_{U_x|U_y}(t|u_y)\\
&=-\int_0^1\big[\widehat{F}^{-1}_{X|Z}-F^{-1}_{X|Z}\big](t|z)dS^m_{U_x|U_y}(t|u_y)\\
&\quad-\int_0^1F^{-1}_{X|Z}(t|z)d\big[\widehat{S}^m_{U_x|U_y}(t|\hat{u}_y)-S^m_{U_x|U_y}(t|\hat{u}_y)\big]\\
&\quad-\int_0^1F^{-1}_{X|Z}(t|z)d\big[S^m_{U_x|U_y}(t|\hat{u}_y)-S^m_{U_x|U_y}(t|u_y)\big]+o_P((nh_z)^{-1/2})\\
&=T_1(y,z)+T_2(y,z)+T_3(y,z)+o_P((nh_z)^{-1/2}),
\end{aligned}$$

uniformly in y. First note that

$$T_2(y,z)=\int_0^1\big[\widehat{S}^m_{U_x|U_y}(t|\hat{u}_y)-S^m_{U_x|U_y}(t|\hat{u}_y)\big]dF^{-1}_{X|Z}(t|z)=o_P((nh_z)^{-1/2}),$$

uniformly in y, thanks to Theorem 3.1. Next, (A.2) allows us to write $T_1(y,z)$ as a sum of i.i.d. terms, up to a negligible term:

$$\begin{aligned}
&T_1(y,z)\\
&= (nh_z)^{-1}\sum_{i=1}^{n} k\Big(\frac{Z_i-z}{h_z}\Big)\int_0^1 \frac{1\!\!1(X_i \le F^{-1}_{X|Z}(t|Z_i)) - t}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z)f_Z(z)} dS^m_{U_x|U_y}(t|u_y)\\
&\quad + \int_0^1 \Big[h_x^2 \frac{B_x(F^{-1}_{X|Z}(t|z)|z)}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z)} + h_z^2 \frac{B_z(F^{-1}_{X|Z}(t|z)|z)}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z)}\Big] dS^m_{U_x|U_y}(t|u_y) + o_P((nh_z)^{-1/2}).
\end{aligned}$$

Finally, we consider the term $T_3(y,z)$:

$$\begin{aligned}
&T_3(y,z)\\
&= m\int_0^1 F^{-1}_{X|Z}(t|z)\big[S^{m-1}_{U_x|U_y}(t|\hat u_y) f_{U_x|U_y}(t|\hat u_y) - S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y)\big]dt\\
&= m\int_0^1 F^{-1}_{X|Z}(t|z)\frac{\partial}{\partial u_y}\big[S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y)\big](\hat u_y - u_y)\,dt + o_P((nh_z)^{-1/2})\\
&= m\Big[(nh_z)^{-1}\sum_{i=1}^{n} k\Big(\frac{Z_i-z}{h_z}\Big)\frac{1\!\!1(Y_i \le y) - F_{Y|Z}(y|Z_i)}{f_Z(z)} + h_y^2 B_y(y|z) + h_z^2 B_z(y|z)\Big]\\
&\quad \times \int_0^1 F^{-1}_{X|Z}(t|z)\frac{\partial}{\partial u_y}\big[S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y)\big]\,dt + o_P((nh_z)^{-1/2}),
\end{aligned}$$

where the last equality follows from (A.1) in Lemma A.1. □

Proof of Theorem 3.4. Let $i_1 = \text{argmin}_j Z_j$, $i_2 = \text{argmin}_{j:Z_j - Z_{i_1} > 2h}(Z_j - Z_{i_1})$, $i_3 = \text{argmin}_{j:Z_j - Z_{i_2} > 2h}(Z_j - Z_{i_2})$, etc., and note that there are $r_n = O(h^{-1})$ such indices. By construction, the variables $\widehat U_{x,i_1}, \widehat U_{x,i_2}, ..., \widehat U_{x,i_{r_n}}$ are mutually independent, and similarly with x replaced by y. Also, note that $0 \le \widehat\phi(u_y) - \phi(u_y) \le \min\{\widehat U_{x,i_k} : k = 1,\dots,r_n, \widehat U_{y,i_k} \ge u_y\} - \phi(u_y)$.

Now, fix $s > 0$ and consider

$$\begin{aligned}
&\mathbb{P}\big(\widehat\phi(u_y) - \phi(u_y) \le s\big)\\
&\ge \mathbb{P}\big(\min\{\widehat U_{x,i_k} : k = 1,\dots,r_n, \widehat U_{y,i_k} \ge u_y\} \le \phi(u_y) + s\big)\\
&= 1 - \mathbb{P}\big(\widehat U_{y,i_1} < u_y \text{ or } \widehat U_{x,i_1} > \phi(u_y) + s\big) \times \ldots \times \mathbb{P}\big(\widehat U_{y,i_{r_n}} < u_y \text{ or } \widehat U_{x,i_{r_n}} > \phi(u_y) + s\big)\\
&\ge 1 - \pi_n^{r_n}, \qquad (A.5)
\end{aligned}$$

where $\pi_n = \max_{k=1,\dots,r_n}\mathbb{P}\big(\widehat U_{y,i_k} < u_y \text{ or } \widehat U_{x,i_k} > \phi(u_y) + s\big)$. For $\delta_n \to 0$, for ν_n sufficiently small and $k = 1, \ldots, r_n$, write

$$\begin{aligned}
&\mathbb{P}\big(\widehat{U}_{y,i_k} < u_y \text{ or } \widehat{U}_{x,i_k} > \phi(u_y) + s\big) \\
&\leq \mathbb{P}\big(\big\{\widehat{U}_{y,i_k} < u_y \text{ or } \widehat{U}_{x,i_k} > \phi(u_y) + s\big\} \text{ and } \sup_{j=x,y} \max_{k=1}^{r_n} \big|\widehat{U}_{j,i_k} - U_{j,i_k}\big| \leq \delta_n\big) \\
&\quad + \mathbb{P}\big(\sup_{j=x,y} \max_{k=1}^{r_n} \big|\widehat{U}_{j,i_k} - U_{j,i_k}\big| > \delta_n\big) \\
&\leq \mathbb{P}\big(U_{y,i_k} < u_y + \delta_n \text{ or } U_{x,i_k} > \phi(u_y) + s - \delta_n\big) + \nu_n \\
&\leq \mathbb{P}\big(U_{y,i_k} < u_y + \delta_n \text{ or } U_{x,i_k} > \phi\big(u_y + \delta_n\big) + \frac{s}{2}\big) + \nu_n \\
&:= 1 - q + \nu_n \leq 1 - \frac{q}{2} < 1,
\end{aligned}$$

since $q > 0$. The second inequality holds since $\max_{i=1}^{n} |\widehat{U}_{y,i} - U_{y,i}| = o_P(1)$ and similarly for $U_{x,i}$. It follows that $\pi_n \to \pi < 1$, and hence (A.5) converges to one, since $r_n \to \infty$ as $n \to \infty$. Hence, $\widehat{\phi}(u_y) - \phi(u_y) \xrightarrow{\mathbb{P}} 0$.

Next, consider

$$\begin{aligned}
\widehat{\tau}(y, z) - \tau(y, z) &= \widehat{F}_{X|Z}^{-1}\big(\widehat{\phi}(\hat{u}_y)|z\big) - F_{X|Z}^{-1}\big(\phi(u_y)|z\big) \\
&= \frac{1}{f_{X|Z}\big(\phi(u_y)|z\big)}\big(\widehat{\phi}(\hat{u}_y) - \phi(u_y)\big) + o_P(1),
\end{aligned}$$

and this converges to zero in probability thanks to the weak consistency of $\hat{u}_y$. □