ESTIMATING NONPARAMETRIC CONDITIONAL FRONTIERS AND EFFICIENCIES: A NEW APPROACH

Camilla Mastromarco, Léopold Simar, Ingrid Van Keilegom

LIDAM Discussion Paper ISBA 2022 / 35



UCLouvain

LIDAM ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

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

November 2022

Abstract

In production theory, conditional frontiers and conditional efficiency measures are a flexible and appealing approach to consider the role of environmental variables on the production process. Direct approaches estimate non-parametrically conditional distribution functions requiring smoothing techniques and the use of selected bandwidths. The statistical literature produces way to derive bandwidths of optimal order, by using e.g. least-squares-cross-validation techniques. However, it has been shown that the resulting order may not be optimal when estimating the boundary of the distribution function. As a consequence the direct approaches may suffer from some statistical instability. In this paper we suggest a full nonparametric approach which avoids the problem of estimating these bandwidths, by eliminating in a first step 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. This can be viewed as an extension of the use of location-scale models (implying some semi-parametric structure) to full nonparametric models, based on nonseparable, nonparametric models. We are also able to recover the frontier and efficiencies in original units. We describe the method, its statistical properties and we show in some Monte-Carlo simulations, how our new method dominates the traditional direct approach.

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 The Setup and Basic Notations

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. From an empirical perspective, the attainable set in the input-output space is unknown as well as its efficient frontier and the efficiency scores, so they have to be estimated from a set of observed units. Nonparametric estimators are quite attractive since they do not rely on restrictive parametric assumptions on the data generating process, in particular, the shape of the frontier and the distribution of the inputs and outputs in the attainable set. They are mainly based on envelopment of the cloud of observed points, like the Data Envelopment Analysis (DEA) initiated by Farrell (1957) and popularized by Charnes et al. (1978), assuming convexity of the attainable set or the Free Disposal Hull (FDH), relaxing convexity and initiated by Deprins et al. (1984). Today, these nonparametric estimators have well-known statistical properties, as described e.g. in Kneip et al. (2015) and in the references therein.

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 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).

A traditional approach in the literature focuses on "two-stage" approaches. The efficiencies are estimated for each observed unit, in a first stage considering only the space of inputs and outputs (X, Y) ignoring Z. Then in a second stage, the estimated efficiencies are regressed on Z (in the best cases, taking into account the limited range of the dependent variable, the efficiency estimates). By formalizing this procedure, Simar and Wilson

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

(2007, 2011) have shown that (i) this approach is meaningful only if a "separability condition" between Z and (X, Y) holds and (ii) even if the second stage regression is meaningful, traditional inference is flawed by the inherent bias of the efficiency estimates and by their nonparametric rate of convergence which may be far from $\sqrt{n}$ if the dimension of (X, Y) increases. The separability condition states that the attainable set in the input-output space does not depend on Z and Daraio et al. (2018) propose a test for validating this assumption, but still, even if validated, usual inference is not available in the second stage regression and if used, this may lead to wrong inference, see Section 5 in Kneip et al., (2015) for details and a formal theory.

A more general and appealing approach is to consider conditional frontiers and conditional efficiency scores 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:

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

Partial frontiers have also been introduced, considering less extreme benchmarks, and so providing estimators more robust to outliers and extreme data points. We have the order-α quantile frontier introduced by Aragon et al. (2005) and Daouia and Simar (2007). Their conditional version are defined for any $\alpha \in [0, 1]$

$$\tau_\alpha(y, z) = \inf\{x|S_{X|Y,Z}(x|y, z) < \alpha\}, \tag{1.3}$$

where we see clearly that when $\alpha \to 1$, $\tau_\alpha(y, z) \to \tau(y, z)$. We have also, in the same spirit the order-m conditional frontier introduced by Cazals et al. (2002). Here for a given integer $m \geq 1$ we have

$$\begin{aligned} \tau_m(y, z) &= \mathbb{E}[\min(X_1, \ldots, X_m)|Y \geq y, Z = z] \\ &= \int_0^\infty S^m_{X|Y,Z}(x|y, z)\, dx, \end{aligned} \tag{1.4}$$

where it is easy to see that as $m \to \infty$, $\tau_m(y,z) \to \tau(y,z)$. The idea here is to benchmark the input of a unit producing the value y of outputs and facing environmental conditions z not against the minimal possible input of such firms but the average of m peers facing the same conditions z and producing at least the value y of outputs (see Cazals et al., 2002 and Daraio and Simar, 2007 for details and discussions on the choice of m, see also Daouia and Gijbels, 2011).

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 frontier $\tau(y)$ and the partial frontiers $\tau_\alpha(y)$ and $\tau_m(y)$, respectively (see e.g. Simar and Wilson, 2015, for a recent survey on these techniques).

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

$$\widehat{S}_{X|Y}(x|y) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \geq x, Y_i \geq y)}{\sum_{i=1}^n \mathbb{1}(Y_i \geq y)}, \tag{1.5}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Then the FDH estimator of the minimal input function is given by

$$\widehat{\tau}(y) = \inf\{x | \widehat{S}_{X|Y}(x|y) < 1\} = \min_{\{i|Y_i \geq y\}} \{X_i\}, \tag{1.6}$$

and, for instance, the order-m estimator is given by

$$\widehat{\tau}_m(y) = \int_0^\infty \widehat{S}^m_{X|Y}(x|y)\, dx. \tag{1.7}$$

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

$$\widehat{S}_{X|Y,Z}(x|y,z) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \geq x, Y_i \geq y) K_{h_z}(Z_i - z)}{\sum_{i=1}^n \mathbb{1}(Y_i \geq y) K_{h_z}(Z_i - z)}, \tag{1.8}$$

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

$$\widehat{\tau}(y,z) = \inf\{x | \widehat{S}_{X|YZ}(x|y,z) < 1\}, \tag{1.9}$$

and, for instance, the order-m estimator is given by

$$\widehat{\tau}_m(y,z) = \int_0^\infty \widehat{S}^m_{X|YZ}(x|y,z)\, dx. \tag{1.10}$$

The statistical properties of the resulting frontier estimators are well established: see Park et al. (2000), Cazals et al. (2002). In a nutshell, we have

$$n^{1/(1+q)}\big(\widehat{\tau}(y)-\tau(y)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_y^{(1+q)}, 1+q) \tag{1.11}$$

$$\sqrt{n}\big(\widehat{\tau}_m(y)-\tau_m(y)\big) \xrightarrow{\mathcal{L}} \text{N}(0,\sigma_y^2), \tag{1.12}$$

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 and 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:

$$(n\bar{h}_z)^{1/(1+q)}\big(\widehat{\tau}(y,z)-\tau(y,z)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_{y,z}^{(1+q)}, 1+q) \tag{1.13}$$

$$\sqrt{n\bar{h}_z}\big(\widehat{\tau}_m(y,z)-\tau_m(y,z)\big) \xrightarrow{\mathcal{L}} \text{N}(0,\sigma_{y,z}^2), \tag{1.14}$$

Similar results hold for the order-α and conditional order-α frontiers to their equivalent order-m (see Daouia and Simar, 2006, for details).

In this traditional or "direct" approach for estimating conditional frontiers, the results in (1.13) and (1.14) 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) which deteriorates the rates in (1.13) and (1.14), since n is replaced by $n^{4/(d+4)}$, in particular when d increases. 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 $S_{X,Y|Z}(x,y|z)$ in the x direction. 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

$$\tau^{h_z}(y,z) = \inf\Big\{x | S_{X|Y,Z}^{h_z}(x|y,z) = \mathbb{P}\big(X \geq x \mid Y \geq y, |Z-z| \leq h_z\big) < 1\Big\} \tag{1.15}$$

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 $||h||$, unless the separability condition holds (see Bădin et al. 2019 for details). Since in practice, with real data, we do not know if this separability condition holds, we may have problem and the LSCV might not be optimal. 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 are encouraging.

An alternative, avoiding these difficulties was suggested by Florens et al. (2014), who propose an indirect approach, avoiding smoothing near the frontier where the data can be rather sparse and avoiding the bandwidth selection issue for estimating the boundary of a conditional distribution. They consider location-scale models to describe the link between the input and the outputs with Z. Then if their location-scale assumptions hold, the resulting errors terms of these models, ε_x and ε_y defined below in (1.16), 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. In our setup, the location-scale model can be defined as follows. We assume that

$$\begin{cases} X = \mu_x(Z) + \sigma_x(Z)\varepsilon_x \\ Y_j = \mu_{y_j}(Z) + \sigma_{y_j}(Z)\varepsilon_{y_j} \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{1.16}$$

where $(\varepsilon_x, \varepsilon_y)$ is assumed to be independent of Z, with mean zero and variance one.

As described in Florens et al. (2014), the nonparametric estimation of the model involves only smoothing in the center of the data cloud and avoids the bandwidth selection issue described above. They show how the original conditional frontiers $\tau(y, z)$ and $\tau_m(y, z)$ can be recovered form the corresponding estimated frontiers in the "pure" units space.

The location model (1.16) is quite flexible, semiparametric but still assume this additive form for the input and each output and more importantly the joint independence between $(\varepsilon_x, \varepsilon_y)$ and Z. In this paper we want to extend the location-scale model to a more general structure where the link between the input and outputs and the environmental factors Z, is described by fully nonparametric models. We will see below that the location-scale models of Florens et al. (2014) can be viewed as a particular semiparametric case of our models. We will show that by doing this extension we do not lose the nice properties of the location-scale approach, i.e. (i) the resulting estimators of the pure 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, and (ii), for the robust frontier estimates, we keep the $\sqrt{n\bar{h}_z}$-convergence to a Gaussian process, as is the case for the estimators derived by the direct approach. However, in our case here, we do not have to rely on the location-scale assumption.

In our paper, as in Florens et al. (2014), we will only focus on the robust order-m frontiers and we let the analysis of the order-α robust frontiers for further research. Comparisons of the merits and drawbacks of the two approaches from a robustness and an asymptotic points of view can be found in Daouia and Ruiz-Gazen (2006) and Daouia and Gijbels (2011). In

particular, it is known that once the order-α quantile frontiers break down for large chosen tail probability levels, they become definitely less resistant to outliers than the order-m frontiers.

The paper is organized as follows. Section 2 presents our model and the way to compute the nonparametric estimators. Section 3 gives the asymptotic properties of the estimators and Section 4 shows through some simulations how our model produces, as expected, more stable estimators than the traditional direct approach. We illustrate in Section 5 its practical use with real data, comparing to the results of Florens et al. (2014) on the same date set. Section 6 concludes. Technical details of the proofs are given in Appendix A.

2 The Model

We suppose the vector (X, Y, Z) follows the following model:

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

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 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 (2.1). 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 (2.1), 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.

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 seen below. The choice of the uniform is a matter of rescaling the U_ℓ to get this nice interpretation.[2]

[2]Since the functions φ_ℓ in (2.1) are monotone increasing with respect to U_ℓ, any monotone, increasing

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:

$$\begin{cases} U_x = F_{X|Z}(X|Z) \\ U_{y_j} = F_{Y_j|Z}(Y_j|Z) \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{2.2}$$

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:

$$\widehat{U}_{x,i} = \widehat{F}_{X|Z}(X_i|Z_i) = \frac{\sum_{k=1}^n G_{h_x}(X_k - X_i)K_{h_z}(Z_k - Z_i)}{\sum_{k=1}^n K_{h_z}(Z_k - Z_i)}, \tag{2.3}$$

$$\widehat{U}_{y_j,i} = \widehat{F}_{Y_j|Z}(Y_{j,i}|Z_i) = \frac{\sum_{k=1}^n G_{h_{y_j}}(Y_{j,k} - Y_{j,i})K_{h_z}(Z_k - Z_i)}{\sum_{k=1}^n K_{h_z}(Z_k - Z_i)}, \tag{2.4}$$

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). 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., 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

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.

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

$$\phi(u_y) = \inf\{u_x | S_{U_x|U_y}(u_x|u_y) < 1\}, \tag{2.5}$$

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

$$\begin{aligned}\phi_m(u_y) &= \mathbb{E}\left[\min(U_{x,1}, \ldots, U_{x,m}) | U_y \geq u_y\right], \\ &= \int_0^1 S^m_{U_x|U_y}(u_x|u_y)\, du_x,\end{aligned} \tag{2.6}$$

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):

$$\begin{aligned}\tau(y,z) &= \inf\{x | S_{X|Y,Z}(x|y,z) < 1\} \\ &= \inf\{x | \mathbb{P}(X \geq x | Y \geq y, Z = z) < 1\} \\ &= \inf\{x | \mathbb{P}(\varphi_x(Z, U_x) \geq x | \varphi_y(Z, U_y) \geq y, Z = z) < 1\} \\ &= \inf\{x | \mathbb{P}(U_x \geq u_x = \varphi_x^{-1}(z,x) | U_y \geq u_y = \varphi_y^{-1}(z,y)) < 1\} \\ &= \inf\{x | S_{U_x|U_y}(u_x|u_y) < 1\} \text{ with } u_x = \varphi_x^{-1}(z,x), u_y = \varphi_y^{-1}(z,y),\end{aligned} \tag{2.7}$$

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

$$\tau(y,z) = \varphi_x\left(z, \inf\{u_x | S_{U_x|U_y}(u_x|u_y) < 1\}\right) = \varphi_x(z, \phi(u_y)), \text{ where } u_y = \varphi_y^{-1}(z,y). \tag{2.8}$$

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

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). Here, we obtain the following equivalence

$$\{\min(X_1, \ldots, X_m) | Y \geq y, Z = z\} = \varphi_x\left(z, \min\left\{(U_{x,1}, \ldots, U_{x,m}) | U_y \geq u_y\right\}\right), \tag{2.9}$$

where the last expression can be read as $\varphi_x(z, \cdot)$ is evaluated at the minimum of m iid replications of $U_{x,j}$, $j = 1, \ldots, m$ 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

$$\tau_m(y, z) = \mathbb{E}\left[\varphi_x(z, \min(U_{x,1}, \ldots, U_{x,m}) | U_y \geq u_y)\right], \tag{2.10}$$

which leads to

$$\tau_m(y, z) = \int_0^1 \varphi_x(z, t) dF_{\tilde{U}_m|U_y}(t|u_y), \tag{2.11}$$

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 (2.5) and (2.6), respectively. This estimator is simply given by the empirical version of $S_{U_x|U_y}(u_x|u_y)$:

$$\widehat{S}_{U_x|U_y}(u_x|u_y) = \frac{\sum_{i=1}^n \mathbb{1}(\widehat{U}_{x,i} \geq u_x, \widehat{U}_{y,i} \geq u_y)}{\sum_{i=1}^n \mathbb{1}(\widehat{U}_{y,i} \geq u_y)}. \tag{2.12}$$

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

$$\begin{aligned} \widehat{\phi}(u_y) &= \inf\{u_x | \widehat{S}_{U_x|U_y}(u_x|u_y) < 1\}, \\ &= \min_{i|\widehat{U}_{y,i} \geq u_y} \widehat{U}_{x,i}. \end{aligned} \tag{2.13}$$

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

$$\widehat{\phi}_m(u_y) = \int_0^1 \widehat{S}^m_{U_x|U_y}(u_x|u_y)\, du_x, \tag{2.14}$$

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

$$\widehat{\tau}(y, z) = \widehat{F}^{-1}_{X|Z}(\widehat{\phi}(\hat{u}_y)|z), \tag{2.15}$$

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.

$$\widehat{\tau}(y, z) = \inf\left\{t \mid \widehat{F}_{X|Z}(t|z) \geq \widehat{\phi}(\hat{u}_y)\right\} = \arg\min_t \mid \widehat{\phi}(\hat{u}_y) - \widehat{F}_{X|Z}(t|z) \mid . \tag{2.16}$$

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 non linearity of $\varphi_x(z, u_x)$ as a function of u_x. In fact we have directly from (2.11) that

$$\widehat{\tau}_m(y, z) = \int_0^1 \widehat{\varphi}_x(z, t)\, d\widehat{F}_{\tilde{U}_m|U_y}(t|\hat{u}_y), \tag{2.17}$$

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

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

$$\widehat{\tau}_m(y,z) = \sum_{j=1}^{n_y} \widehat{\varphi}_x(z, V^y_{(j)}) \left[\left(\frac{n_y - j + 1}{n_y} \right)^m - \left(\frac{n_y - j}{n_y} \right)^m \right], \tag{2.18}$$

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

$$\widehat{\varphi}_x(z,\gamma) = \widehat{F}^{-1}_{X|Z}(\gamma|z) = \inf \left\{ t \mid \widehat{F}_{X|Z}(t|z) \geq \gamma \right\} = \arg\min_t \mid \gamma - \widehat{F}_{X|Z}(t|z) \mid, \tag{2.19}$$

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

Remark 2.1. *Having the estimates of the conditional frontiers and estimates of the unconditional (marginal) frontiers, several analysis 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 since it is expected to be statistically much more stable than the traditional direct approach (as confirmed by our Monte-Carlo experiments below), this may provide promising tools for the practitioner.*

Remark 2.2. *As pointed out above, the Florens et al. (2014) location-scale model* (1.16) *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 $U_\ell = \eta^{-1}_\ell(\varepsilon_\ell)$. 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 2.3. *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 (2.3)–(2.4). 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 $\|f\|_{1+\delta} \leq M$, 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,*

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

where

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

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

(*ii*) *We have* $\sup_{0\leq u_x,u_y\leq 1}|\widehat{S}_{U_x|U_y}(u_x|u_y) - S_{U_x|U_y}(u_x|u_y)| = O_P(n^{-1/2}) + O(h_x^2+h_y^2+h_z^2)$*, 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* $0\leq C_x, C_y, C_z < \infty$*, then the process* $n^{1/2}(\widehat{S}_{U_x|U_y}(u_x|u_y) - S_{U_x|U_y}(u_x|u_y))$*,* $0\leq u_x, u_y\leq 1$*, 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,*

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

where

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

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

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

and

$$b_x(y|z) = \int_0^1 \frac{B_x(F^{-1}_{X|Z}(t|z)|z)}{f_{X|Z}(F^{-1}_{X|Z}(t|z)|z)} dS^m_{U_x|U_y}(t|u_y)$$

$$b_y(y|z) = B_y(y|z) \int_0^1 F^{-1}_{X|Z}(t|z) \frac{\partial}{\partial u_y} \left[S^{m-1}_{U_x|U_y}(t|u_y) f_{U_x|U_y}(t|u_y) \right] dt$$

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

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* $0 \leq C_x, C_y, C_z < \infty$, *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 given by*

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

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,*

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

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

4 Numerical Illustration

The following is inspired from the DGPs proposed by Bădin et al. (2019) and Simar and Wilson (2011) for illustrating the various ways an external variable Z can influence the production process. We adapt the examples to an input oriented measure for the one input X case (X is a cost), with one output Y, first with one environmental variable Z. We will investigate four different cases.

4.1 Univariate Z

The basic "cost" frontier is given by the function $\tau(y) = 1 + y^2$ for $y \in (0, 3)$. We will consider an external factor $Z \in (0, 4)$. The role of Z can be introduced in various way. In all the cases below we will consider that the output Y depends on Z by the assumption

$$Y \mid Z = z \sim 3\text{Beta}(1 + z/4, 2 - z/4) \tag{4.1}$$

where $\text{Beta}(a, b)$ is the traditional beta distribution. So depending on the value of z, the shape of $f(y|Z = z)$ give more weights to values of Y near its lower bound for small z and near its upper bound for large values of z. For instance we have $\mathbb{E}(Y \mid Z = z) = 1 + z/4$, where $z \in (0, 4)$. We will generate Z as a uniform on $(0, 4)$

In all the cases, the inefficiency will be governed by a multiplicative factor $\exp(\xi) \geq 1$ to the frontier function, where $\xi \geq 0$. We will choose

$$\xi \mid Z = z \sim N^+(\mu_\xi(z), \sigma_\xi^2(z)), \tag{4.2}$$

where the dependence on Z will be adapted according to our cases, by selecting $\mu_\xi(z)$ and $\sigma_\xi(z)$ in an appropriate way.

In addition to this dependence, Z may also influence the shape of the frontier, or not. This will give the next four cases of DGPs we will consider in our Monte-Carlo experiments for generating the cost X.

- **Case A**. We assume that Z is independent of ξ and does not influence the shape of the frontier, i.e. the "separability" condition in the sense of Simar and Wilson (2007, 2011). Only Y depends on Z by (4.1). So we have

  $$X = \tau(Y) \times \exp(\xi), \tag{4.3}$$

  with $\mu_\xi(z) = 0$ and $\sigma_\xi(z) = 0.5$. So here $\tau(y, z) = \tau(y) = 1 + y^2$.

- **Case B**. We still have the "separability" condition as in Case A, but ξ depends on Z. We have here

  $$X = \tau(Y) \times \exp(\xi), \tag{4.4}$$

  with $\mu_\xi(z) = 1 - z$ and $\sigma_\xi(z) = (z + 3)/10$. Here again $\tau(y, z) = \tau(y) = 1 + y^2$.

- **Case C**. The same as Case A, but Z influences the shape of the cost frontier. We have

  $$X = \tau(Y) \times \exp(-(Z - 2)/3) \times \exp(\xi), \tag{4.5}$$

  with $\mu_\xi(z) = 0$ and $\sigma_\xi(z) = 0.5$. Now $\tau(y, z) = (1 + y^2) \exp(-(z - 2)/3)$.

- **Case D**. This final case allows Z to play a role at all the possible levels, combining Case B and Case C, here ξ and the level of the cost frontier depend both on Z. We have

$$X = \tau(Y) \times \exp(-(Z-2)/3) \times \exp(\xi), \tag{4.6}$$

with $\mu_\xi(z) = 1 - z$ and $\sigma_\xi(z) = (z+3)/10$. Again, $\tau(y, z) = (1 + y^2)\exp(-(z-2)/3)$.

So, to generate a sample of size n, we simulate first $Z \sim \text{Unif}(0, 4)$, then for each Z_i we simulate Y_i as in (4.1) and we simulate ξ_i as in (4.2), where the latter is specified according one of the four cases described above. Finally X_i is generated as (4.3) to (4.6), depending on the chosen scenario (Case A to D).

To fix the ideas, we show some figures of simulated samples and frontier. First to see the role of Z on the distribution of ξ driving the inefficiency, Figure 1 gives $\mathbb{E}(\xi \mid Z = z)$ and $\sigma(\xi \mid Z = z)$ as function of z.

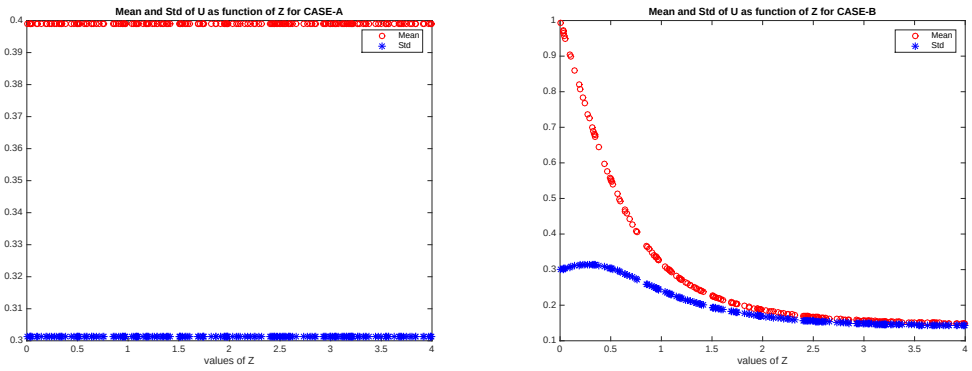


Figure 1: Values of $\mathbb{E}(\xi \mid Z = z)$ and $\sigma(\xi \mid Z = z)$ as function of $z = Z_i$. Left panel, Case A and Case C, right panel Case B and Case D.

For the impact of Z and Y on the frontier level, Figure 2 displays the different cases.

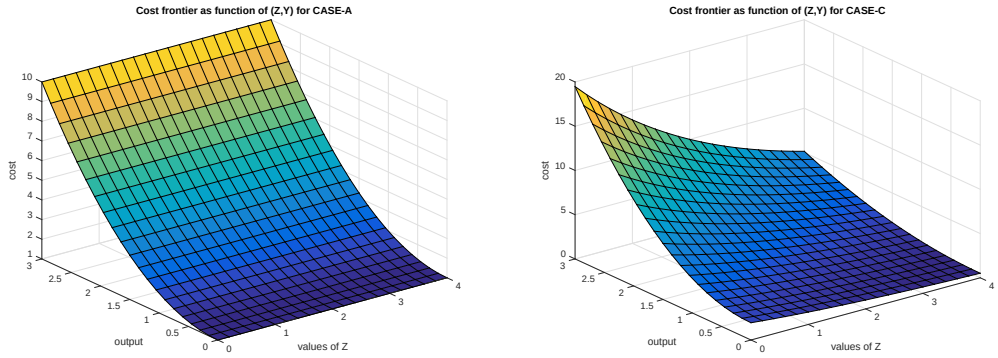


Figure 2: Shape of the minimal cost frontier as function of Y and Z. Left panel, Case A and Case B, right panel Case C and Case D.

Finally we may have an idea on how the values of X are distributed as function of Y (and Z) by displaying the simulated data points (X_i, Y_i) and their corresponding frontier points, for the four cases. This may be seen in Figure 3 for one sample of size $n = 200$ in the 4 scenarios.

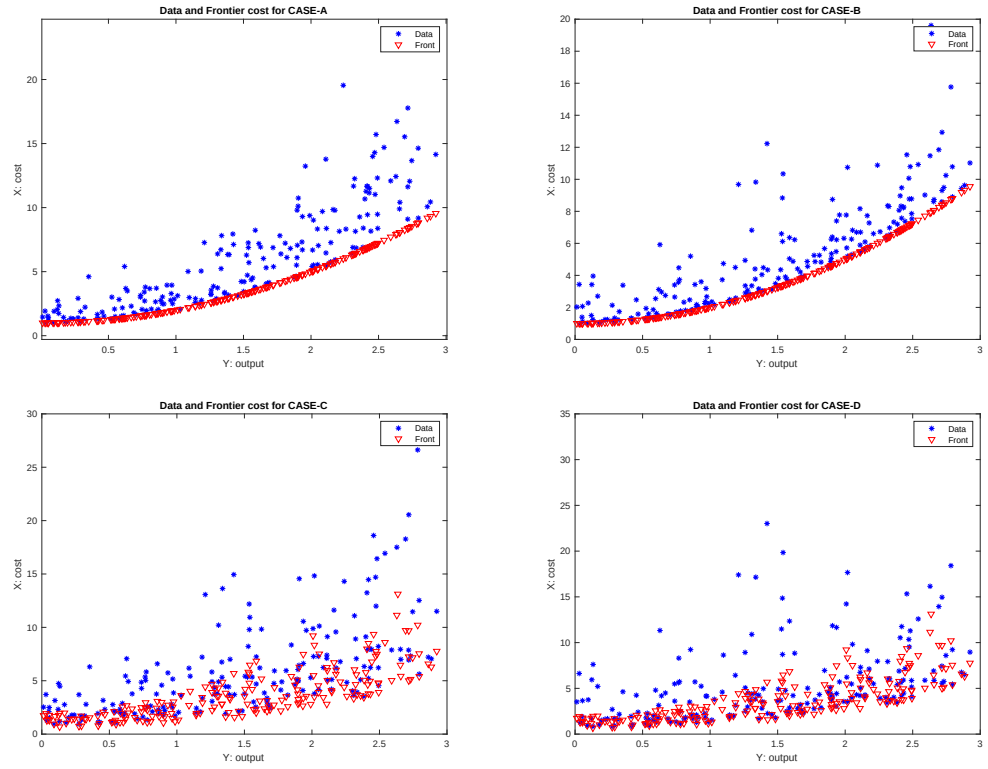


Figure 3: The data and their corresponding conditional frontier $\tau(Y_i, Z_i)$ for $n = 200$ for the four cases. Note that the vertical scale is different for each case.

Now the idea is to see if we can correctly estimate the true values $\tau(Y_i, Z_i)$ in each case. We will compare below the performance of our estimator compared with the traditional, direct one. Note that the true value is given by the chosen scenario, so we can look at the estimate of the ISE (Integrated Squared Error) given by

$$ISE = n^{-1} \sum_{i=1}^{n} (\widehat{\tau}(Y_i, Z_i) - \tau(Y_i, Z_i))^2 \tag{4.7}$$

We will do this for $n = 100, 200$ and 500. We estimate the Mean Integrated Squared Error ($MISE$) by doing 500 Monte-Carlo repetitions and averaging the ISE over the 500 trials. To check if some differences are significant we give also the estimate of the standard deviation of the Monte-Carlo estimator of the $MISE$. The results are summarized through Tables 1 to 4, where "Direct" are the results for the traditional method and "New Method" is for our approach. We consider the four different scenarios, Case A to D described above, regarding the situations on Z being separable or not and influencing the distribution of inefficiencies or not.

For the order-m partial frontiers, we will only focus for 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 integration on x for computing $\tau_m(y, z)$ in each scenario. 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 when we will introduce some outliers in the sample (see below).

Table 1: $MISE$ fo the simulated examples, Univariate Z, Case A: Z does not affect neither the frontier, nor the efficiency distribution.

CASE-A

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	3.0886	3.9684	3.3968	3.1670	3.1194	3.1023
(std)	(0.1103)	(0.1107)	(0.1103)	(0.1104)	(0.1104)	(0.1104)
New Method	1.2622	2.3661	1.6585	1.3455	1.2811	1.2649
(std)	(0.0393)	(0.0454)	(0.0409)	(0.0395)	(0.0393)	(0.0393)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.0347	3.0541	2.4064	2.1310	2.0715	2.0506
(std)	(0.0582)	(0.0575)	(0.0576)	(0.0581)	(0.0582)	(0.0582)
New Method	0.7280	1.9522	1.1975	0.8443	0.7608	0.7346
(std)	(0.0232)	(0.0271)	(0.0243)	(0.0233)	(0.0232)	(0.0232)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.2657	2.4602	1.7354	1.3981	1.3157	1.2855
(std)	(0.0347)	(0.0342)	(0.0342)	(0.0345)	(0.0347)	(0.0347)
New Method	0.3879	1.6504	0.8803	0.5142	0.4242	0.3940
(std)	(0.0095)	(0.0119)	(0.0099)	(0.0093)	(0.0093)	(0.0094)

Table 2: $MISE$ fo the simulated examples, Univariate Z, Case B: Z does not affect the frontier, but affects the efficiency distribution.

CASE-B

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.7255	2.2613	1.9210	1.7835	1.7522	1.7389
(std)	(0.0746)	(0.0749)	(0.0746)	(0.0747)	(0.0747)	(0.0746)
New Method	0.9951	1.6356	1.2084	1.0281	0.9992	0.9954
(std)	(0.0317)	(0.0367)	(0.0335)	(0.0321)	(0.0318)	(0.0317)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.3737	2.0135	1.6125	1.4408	1.4016	1.3862
(std)	(0.0544)	(0.0544)	(0.0543)	(0.0544)	(0.0544)	(0.0544)
New Method	0.7660	1.4707	1.0052	0.7910	0.7549	0.7576
(std)	(0.0197)	(0.0232)	(0.0207)	(0.0195)	(0.0194)	(0.0196)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	0.9414	1.7394	1.2597	1.0362	0.9802	0.9582
(std)	(0.0298)	(0.0302)	(0.0298)	(0.0298)	(0.0298)	(0.0298)
New Method	0.6562	1.3931	0.8849	0.6335	0.5855	0.5968
(std)	(0.0120)	(0.0155)	(0.0129)	(0.0114)	(0.0111)	(0.0112)

Table 3: $MISE$ fo the simulated examples, Univariate Z, Case C: Z does affect the frontier, but not the efficiency distribution.

CASE-C

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.6104	3.2930	2.8020	2.6357	2.6130	2.6088
(std)	(0.1166)	(0.1190)	(0.1179)	(0.1173)	(0.1170)	(0.1168)
New Method	1.4562	2.0993	1.6428	1.4895	1.4643	1.4575
(std)	(0.0451)	(0.0476)	(0.0453)	(0.0450)	(0.0450)	(0.0451)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.8486	2.7045	2.1238	1.8980	1.8580	1.8488
(std)	(0.0896)	(0.0899)	(0.0899)	(0.0899)	(0.0899)	(0.0898)
New Method	1.0669	1.7416	1.2455	1.0902	1.0727	1.0691
(std)	(0.0303)	(0.0306)	(0.0292)	(0.0297)	(0.0301)	(0.0302)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.1907	2.2842	1.5969	1.2924	1.2236	1.2012
(std)	(0.0550)	(0.0556)	(0.0553)	(0.0553)	(0.0552)	(0.0552)
New Method	0.8327	1.4378	0.9287	0.7934	0.8024	0.8211
(std)	(0.0188)	(0.0150)	(0.0152)	(0.0170)	(0.0180)	(0.0185)

Table 4: $MISE$ fo the simulated examples, Univariate Z, Case D: Z does affect both the frontier and the efficiency distribution.

CASE-D

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	3.7323	4.3977	3.9437	3.7810	3.7504	3.7398
(std)	(0.2404)	(0.2453)	(0.2425)	(0.2413)	(0.2409)	(0.2406)
New Method	1.7641	2.6995	2.1119	1.8421	1.7827	1.7667
(std)	(0.0786)	(0.0978)	(0.0862)	(0.0804)	(0.0790)	(0.0787)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	3.1806	4.1347	3.5125	3.2608	3.2093	3.1923
(std)	(0.1817)	(0.1851)	(0.1833)	(0.1824)	(0.1821)	(0.1819)
New Method	1.5243	2.8060	2.0703	1.6782	1.5698	1.5332
(std)	(0.0535)	(0.0720)	(0.0618)	(0.0558)	(0.0541)	(0.0536)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.2519	3.6598	2.8180	2.4208	2.3203	2.2809
(std)	(0.0946)	(0.0990)	(0.0964)	(0.0953)	(0.0949)	(0.0948)
New Method	1.2702	2.9757	2.0652	1.5341	1.3618	1.2895
(std)	(0.0327)	(0.0511)	(0.0417)	(0.0357)	(0.0338)	(0.0329)

The results are quite clear. First we see that, in univariate Z scenario, for the 4 cases and sample sizes $n = 100, 200$ and 500, our new method behaves, as expected, better than the traditional approach, with lower $MISE$ and lower standard deviation of the Monte-Carlo estimator of the $MISE$. We have the same result for the order-m estimator for all values of m. These differences between the $MISE$ for the two methods are significant (see *std*). In

some cases, our new method provides less than 50% of the $MISE$ of the traditional "direct" method.

For the 4 scenarios, the MISEs, as expected, decrease when n increases and as m increases (since our true benchmark value is the full frontier). Note that the performances of the order-m estimator converge to those of the full (FDH) estimator faster with the new method than the traditional one. Interestingly the *std* of the MISE is much higher for the direct method than for our new method (both for the full frontier and for the order-m frontier cases). As expected, the direct method shows, much more statistical instability (std) than the new method, due to the facts discussed above: in the direct approach, we estimate the boundary of a CDF (cumulative distribution function) and not a CDF itself. So the methods faces sparsity of the data near boundaries and the problem of selecting optimal bandwidths.

Note that in the CASE-A (Table 1) with separability and no influence of Z on the efficiency distribution, no conditioning is required for both full and order-m frontiers. The marginal FDH estimator should be used. But when conditioning on Z, our approach provides, for all sample sizes, lower $MISE$ and standard deviations than the direct approach. The same remark applies for CASE-B (table 2) with separability, but Z affecting the efficiency distribution, so here the conditioning is required for order-m frontier, and again our method provides better results.

4.2 Bivariate Z

We do now the same exercise as the preceding section but adding a second variables $Z_2 \sim$ Unif$(0, 4)$ independent of $Z_1 = Z$ as above also Unif$(0, 4)$. But here Z_2 only affects the inefficiency ξ in Case B and D as follows:

$$\mu_\xi(z) = 1 - (z_1 + z_2)/2 \tag{4.8}$$

$$\sigma_\xi(z) = ((z_1 + z_2)/2 + 3)/10. \tag{4.9}$$

All the other elements of the four Cases remain the same as above. The results are displayed in Tables 5–8. To summarize we can say that qualitatively we have the same comments as for the univariate Z case. Our new method outperforms significantly the traditional direct one in all the cases. The *std* of the $MISE$ confirms also the higher statistical stability across the various Monte-Carlo trials.

Table 5: $MISE$ fo the simulated examples, Bivariate Z, Case A: Z does not affect neither the frontier, nor the efficiency distribution.

CASE-A

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.9571	3.8840	3.3152	3.0684	3.0082	2.9825
(std)	(0.1182)	(0.1182)	(0.1183)	(0.1185)	(0.1184)	(0.1183)
New Method	1.4228	2.5087	1.8123	1.5043	1.4415	1.4256
(std)	(0.0651)	(0.0667)	(0.0653)	(0.0651)	(0.0651)	(0.0651)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.1431	3.2121	2.5681	2.2767	2.2023	2.1709
(std)	(0.0684)	(0.0680)	(0.0683)	(0.0687)	(0.0687)	(0.0686)
New Method	0.8099	2.0233	1.2757	0.9242	0.8417	0.8162
(std)	(0.0232)	(0.0281)	(0.0249)	(0.0236)	(0.0233)	(0.0232)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.3786	2.5897	1.8737	1.5346	1.4463	1.4097
(std)	(0.0418)	(0.0414)	(0.0415)	(0.0418)	(0.0419)	(0.0419)
New Method	0.4118	1.6824	0.9069	0.5363	0.4458	0.4164
(std)	(0.0098)	(0.0125)	(0.0105)	(0.0098)	(0.0097)	(0.0097)

Table 6: $MISE$ fo the simulated examples, Bivariate Z, Case B: Z does not affect the frontier, but affects the efficiency distribution.

CASE-B

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	0.8727	1.2475	1.0127	0.9149	0.8903	0.8802
(std)	(0.0487)	(0.0490)	(0.0490)	(0.0487)	(0.0486)	(0.0487)
New Method	0.5080	0.9487	0.6512	0.5310	0.5115	0.5083
(std)	(0.0198)	(0.0184)	(0.0183)	(0.0190)	(0.0195)	(0.0197)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	0.6842	1.1119	0.8438	0.7327	0.7061	0.6951
(std)	(0.0251)	(0.0249)	(0.0251)	(0.0252)	(0.0252)	(0.0252)
New Method	0.3350	0.8437	0.5104	0.3631	0.3355	0.3325
(std)	(0.0099)	(0.0115)	(0.0102)	(0.0097)	(0.0097)	(0.0098)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	0.4997	1.0093	0.6991	0.5633	0.5294	0.5149
(std)	(0.0156)	(0.0156)	(0.0157)	(0.0159)	(0.0159)	(0.0158)
New Method	0.2460	0.7771	0.4193	0.2547	0.2220	0.2221
(std)	(0.0056)	(0.0065)	(0.0055)	(0.0051)	(0.0051)	(0.0052)

Table 7: $MISE$ fo the simulated examples, Bivariate Z, Case C: Z_1 does affect the frontier, but not the efficiency distribution.

CASE-C

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.3531	2.6726	2.3238	2.2669	2.2927	2.3191
(std)	(0.1162)	(0.1201)	(0.1186)	(0.1174)	(0.1169)	(0.1166)
New Method	1.5627	2.1481	1.7166	1.5854	1.5680	1.5636
(std)	(0.0527)	(0.0536)	(0.0522)	(0.0526)	(0.0527)	(0.0527)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.7979	2.4533	1.9530	1.7833	1.7672	1.7745
(std)	(0.0796)	(0.0848)	(0.0830)	(0.0814)	(0.0806)	(0.0801)
New Method	1.1495	1.7502	1.2835	1.1541	1.1480	1.1496
(std)	(0.0299)	(0.0299)	(0.0286)	(0.0292)	(0.0297)	(0.0298)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.1515	2.0892	1.4612	1.2030	1.1538	1.1433
(std)	(0.0388)	(0.0427)	(0.0418)	(0.0407)	(0.0401)	(0.0396)
New Method	0.9341	1.4903	0.9909	0.8708	0.8907	0.9167
(std)	(0.0220)	(0.0196)	(0.0193)	(0.0205)	(0.0213)	(0.0217)

Table 8: $MISE$ fo the simulated examples, Bivariate Z, Case D: Z_1 does affect the frontier, and Z_1 and Z_2 affect both the efficiency distribution.

CASE-D

$n = 100$	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)
$n = 200$	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)
$n = 500$	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)

4.3 Bivariate Z with outliers

Finally, to illustrate the usefulness of the robust order-m partial frontier we investigate how the order-m behaves for various values of m when adding 3 outliers in the DGP. In fact we replaced the 3 last simulated values of (X_i, Y_i, Z_i) by fixing $Y_i = 1, 1.75$ and 2, keeping the

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 Tables 9–12.

Of course, we can see that the order-m frontiers are more resistant to these outliers when compared to the full frontier estimates, unless m becomes too large (and converges to the full frontier). This being said, we can confirm the qualitative results observed in the scenarios without outliers: our new method behaves much better than the direct ones, not only in the $MISE$ but again in the higher stability indicated by the smaller *std* across the various simulated samples.

Table 9: $MISE$ fo the simulated examples, Bivariate Z, Case A: Z does not affect neither the frontier, nor the efficiency distribution, with 3 outliers.

CASE-A

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	3.3763	3.4837	3.2392	3.2870	3.3350	3.3577
(std)	(0.1121)	(0.1175)	(0.1156)	(0.1134)	(0.1126)	(0.1123)
New Method	2.4727	1.7133	1.5665	2.0347	2.3341	2.4503
(std)	(0.0580)	(0.0574)	(0.0560)	(0.0569)	(0.0577)	(0.0580)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.5729	2.9524	2.5170	2.4751	2.5205	2.5486
(std)	(0.0644)	(0.0696)	(0.0688)	(0.0667)	(0.0655)	(0.0649)
New Method	2.1161	1.4672	0.9886	1.2414	1.6695	1.9766
(std)	(0.0312)	(0.0262)	(0.0238)	(0.0251)	(0.0281)	(0.0304)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.7054	2.4435	1.8226	1.6347	1.6461	1.6737
(std)	(0.0413)	(0.0440)	(0.0440)	(0.0433)	(0.0424)	(0.0418)
New Method	1.7338	1.4111	0.7043	0.5657	0.8143	1.2057
(std)	(0.0204)	(0.0122)	(0.0102)	(0.0107)	(0.0127)	(0.0159)

Table 10: $MISE$ fo the simulated examples, Bivariate Z, Case B: Z does not affect the frontier, but affects the efficiency distribution, with 3 outliers.

CASE-B

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.5475	1.2302	1.2532	1.4053	1.4796	1.5150
(std)	(0.0464)	(0.0475)	(0.0468)	(0.0463)	(0.0463)	(0.0464)
New Method	1.9523	0.7437	0.9229	1.5037	1.8138	1.9304
(std)	(0.0276)	(0.0173)	(0.0200)	(0.0243)	(0.0267)	(0.0275)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.2214	1.0757	0.9735	1.0685	1.1443	1.1859
(std)	(0.0239)	(0.0258)	(0.0253)	(0.0241)	(0.0238)	(0.0238)
New Method	1.9060	0.6219	0.5398	0.9845	1.4514	1.7649
(std)	(0.0256)	(0.0101)	(0.0119)	(0.0180)	(0.0222)	(0.0247)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	0.8499	0.9647	0.7333	0.7214	0.7689	0.8075
(std)	(0.0148)	(0.0163)	(0.0163)	(0.0156)	(0.0150)	(0.0148)
New Method	1.7575	0.6314	0.3390	0.4268	0.7664	1.2116
(std)	(0.0231)	(0.0061)	(0.0057)	(0.0093)	(0.0140)	(0.0183)

Table 11: $MISE$ fo the simulated examples, Bivariate Z, Case C: Z_1 does affect the frontier, but not the efficiency distribution, with 3 outliers.

CASE-C

$n = 100$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	3.4325	2.7190	2.7861	3.0897	3.2522	3.3418
(std)	(0.1135)	(0.1167)	(0.1145)	(0.1135)	(0.1135)	(0.1135)
New Method	2.9423	1.9016	2.0057	2.5330	2.8151	2.9215
(std)	(0.0564)	(0.0479)	(0.0496)	(0.0531)	(0.0553)	(0.0562)
$n = 200$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	2.5705	2.3285	2.1138	2.2587	2.3928	2.4780
(std)	(0.0746)	(0.0835)	(0.0804)	(0.0771)	(0.0757)	(0.0751)
New Method	2.5055	1.5216	1.3591	1.7470	2.1329	2.3889
(std)	(0.0369)	(0.0293)	(0.0297)	(0.0320)	(0.0345)	(0.0364)
$n = 500$	Full	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 200$
Direct	1.6552	1.9790	1.4895	1.4288	1.5016	1.5686
(std)	(0.0371)	(0.0436)	(0.0419)	(0.0398)	(0.0385)	(0.0378)
New Method	2.1864	1.3435	0.9598	1.0915	1.4012	1.7567
(std)	(0.0280)	(0.0197)	(0.0202)	(0.0223)	(0.0239)	(0.0256)

Table 12: $MISE$ fo 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.

CASE-D

$n = 100$	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)
$n = 200$	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)
$n = 500$	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)

As a conclusion, through these Monte-Carlo simulations, which cover various cases of dependency to Z, our new method outperforms the traditional direct method, i.e. larger stability and a better performance, mainly due to the reason that we are still in a nonparametric approach (no model restrictions) and we avoid the delicate problem of bandwidths selection in the direct method.

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) using the direct traditional approach and in Florens et al. (2014), hereafter FSVK, using the location-scale model to clean the input (a cost) and the outputs. The original dataset come 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) and a measure of the diversity of the services proposed by the banks Z_2. 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 4 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 selected $m = 30$ as in FSVK for ease of comparison). We can compare this figure with Figure 10 in FSVK. The scale is different because here the pure units are in $[0, 1]$, where in FSVK they are real, but clearly we see that the full frontier in Figure 4 envelops nicely the cloud of data in pure units, which is less the case in Figure 10 of FSVK, with the same qualitative remark for the order-m frontier. Clearly our approach is fully nonparametric, whereas FSVK is based on a semiparametric model, imposing the location-scale models for cleaning the data. We know also from our Monte-Carlo experiments that our approach is statistically and numerically very stable. The FSVK location-scale approach involves in a second step, for the scale model, a nonparametric smoothing of the squares of the residuals from the nonparametric location estimation, and this is more sensitive to some extreme values. FSVK provides a test of the location-model and the assumption is not rejected but we know that the test based on Einmahl and Van Keilegom (2008) is not very powerful.

Figure 5 displays the distribution of the resulting "pure" efficiency scores, i.e. $\delta_i = \hat{u}_{x_i} - \widehat{\phi}(\hat{u}_{y_i})$ and $\delta_{i,m} = \hat{u}_{x_i} - \widehat{\phi}_m(\hat{u}_{y_i})$, this can be compared with Figure 11 in FSVK, but the comparison is difficult since our pure units and the "pure" frontiers are somewhat different as explained above. Still our approach is more general and does not rely on some semiparametric assumptions.

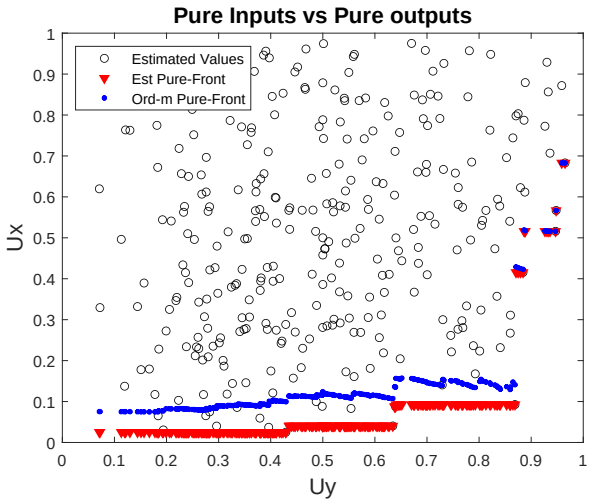


Figure 4: Bank example: Estimated "pure" inputs and outputs and the estimated efficient frontiers $\widehat{\phi}(\hat{u}_y)$ and $\widehat{\phi}_m(\hat{u}_y)$, here $m = 30$.

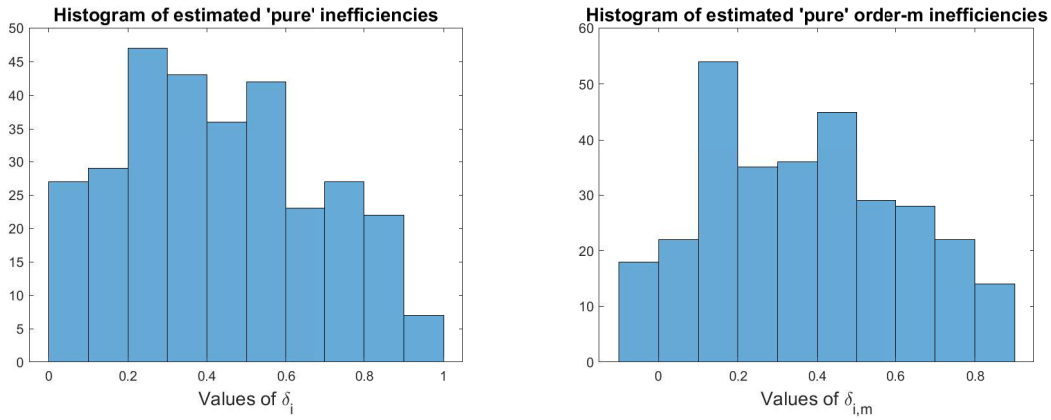


Figure 5: Bank example: distribution of the estimated "pure" inefficiencies, relative to the frontier estimates (full and order-m).

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. In Tables 13 and 14 we display the results for 15 randomly chosen banks. For ease of comparison with Tables 1 and 2 in FSVK, we select the same units and the efficiency scores, in Table 14, are computed in the same way as in FSVK (directional distance for the pure units and radial measure for the original units). Note that it is difficult to compare the column of results for the pure efficiency scores with the corresponding column in FSVK, since our pure inputs/outputs are measured on different scales. To save space, we only display our results for the order-m (as in FSVK). Note that the results for the full frontier case are qualitatively very similar.

It is also interesting to look at Figure 12 in FSVK that shows the conditional frontier (and its order-m version) in the original units. In our case we have Figure 6, where we see the data points (Y_i, X_i) and the estimates of $\widehat{\tau}(y, z)$ evaluated for each pair (Y_i, Z_i). As in FSVK, this figure is not easy to interpret since the estimated frontier points do not follow the shape of the lower boundary of the cloud of data points, because each bank is facing different environmental conditions.

Table 13: Results for 15 randomly selected banks, as in FSVK. Data values and estimates of the frontier levels at the data points. Full and order-m with $m = 30$: "old" is for the traditional direct method and "new" is for our new method.

Unit i	X_i	Y_i	Z_{1i}	Z_{2i}	$\widehat{\tau}_{old}(Y_i, Z_i)$	$\widehat{\tau}_{new}(Y_i, Z_i)$	$\widehat{\tau}_{m,old}(Y_i, Z_i)$	$\widehat{\tau}_{m,new}(Y_i, Z_i)$
259	7.2986	8.5127	1.1778	1.1388	7.2986	6.2393	7.2986	6.2432
237	0.3505	0.5186	0.9409	0.8371	0.3505	0.3505	0.3505	0.4047
258	0.1998	0.2958	0.8700	1.2650	0.1998	0.1604	0.1998	0.1830
1	1.1985	1.3999	1.0199	0.8539	1.1038	1.0134	1.1038	1.0565
241	0.8693	0.8173	0.9863	0.8737	0.8693	0.6903	0.8693	0.7254
66	0.3421	0.3546	0.9028	0.8407	0.2874	0.2396	0.2876	0.2592
164	1.8694	2.1868	1.0551	1.1069	1.8694	1.6086	1.8694	1.6930
274	0.4026	0.4185	0.9152	1.0590	0.3589	0.2715	0.3589	0.2990
303	0.2969	0.5053	0.9024	1.1611	0.2969	0.2625	0.2969	0.2725
199	2.6751	2.9132	1.0786	0.9124	2.6751	2.5152	2.6751	2.5158
216	7.2741	6.6494	1.1670	1.0387	7.2741	5.7449	7.2741	5.8366
125	1.0559	1.1407	1.0111	0.6974	1.0559	0.8559	1.0559	0.9117
239	1.3945	1.7164	1.0330	1.2376	1.2875	1.1799	1.2875	1.2690
170	2.9572	2.4389	1.0807	0.8205	2.4650	1.9072	2.4650	2.1360
242	1.8388	2.1842	1.0735	0.8725	1.8388	1.8138	1.8388	1.9190

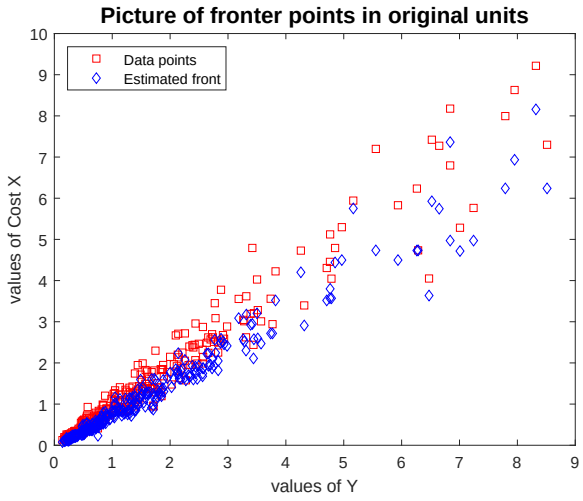


Figure 6: Bank example: data points in original units and estimated $\widehat{\tau}(Y_i, Z_i)$.

More interesting is to analyze the shape of the frontier in the (Y, X) space for fixed values of the environmental conditions. So as in FSVK, we selected 9 pairs $(QZ_{1k}, QZ_{2\ell})$, for $k, \ell = 1, 2, 3$, for each quartile of Z_1 and of Z_2. In Figure 7, to be compared with Figure 13 in FSVK, we see the case for Z_2 fixed at its median and the 3 different quartiles for Z_1 (as in FSVK, the conditional frontiers do not change so much when changing the value of Z_2). We see again that with our fully nonparametric approach the frontiers envelop nicely

Table 14: Pure and conditional efficiency scores of order-m, for the same 15 units, as computed in FSVK, i.e. $\widehat{\delta}_{m,i} = \widehat{U}_{x,i} - \widehat{\phi}_m(\widehat{U}_{y,i})$ and $\widehat{\theta}_m(Y_i, Z_i) = \widehat{\tau}(Y_i, Z_i)/X_i$. The ranks are computed relative to the pure order-m efficiency scores, $m = 30$.

Unit i	Rank$(\widehat{\delta}_{m,i})$	$\widehat{\delta}_{m,i}$	$\widehat{\theta}_{m,old}(Y_i, Z_i)$	$\widehat{\theta}_{m,new}(Y_i, Z_i)$
259	41	0.1017	1.0000	0.8554
237	1	-0.0745	1.0000	1.1548
258	37	0.0743	1.0000	0.9157
1	120	0.2848	0.9210	0.8816
241	150	0.3514	1.0000	0.8344
66	218	0.5225	0.8408	0.7577
164	104	0.2478	1.0000	0.9056
274	206	0.4826	0.8916	0.7428
303	53	0.1378	1.0000	0.9177
199	107	0.2558	1.0000	0.9404
216	271	0.7154	1.0000	0.8024
125	108	0.2562	1.0000	0.8634
239	145	0.3354	0.9233	0.9100
170	297	0.8274	0.8335	0.7223
242	7	-0.0438	1.0000	1.0436

the cloud of points, much better than in the location-scale model in FSVK, as can be seen in their Figure 13. Again, our model encompass the FSVK model and is more general. The picture confirms that the SIZE (i.e. Z_1) impacts the level of the frontier.

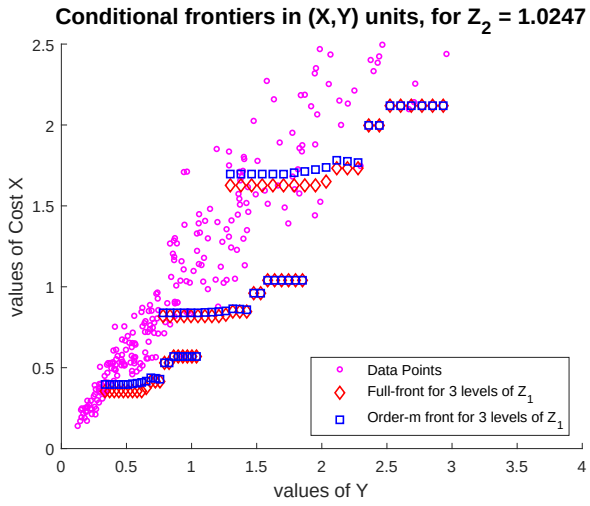


Figure 7: Bank example: data points in original units and frontier estimates when fixing the level of Z. Here Z_2 is fixed at its median value, and Z_1 is fixed at its 3 quartiles (from the left to the right).

A final point of interest is to investigate how the support of (Y, X) is impacted by the

values of Z. FSVK provide in Figure 14, 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 can be compared with our Figure 8. In FSVK only the surface of the ratios is shown for Y fixed at its median (since they claim that the 2 other cases are quite similar). 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 and also some difference with Figure 14 in FSVK. For the latter (Y equal its median) in our approach, the effect of the SIZE (Z_1) seems less important and the effect of DIVERSITY (Z_2) seems slightly more important for small value 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 favorable (smaller level for the conditional cost frontier compared to the marginal one), but this effect seems to disappear for banks having the largest output.

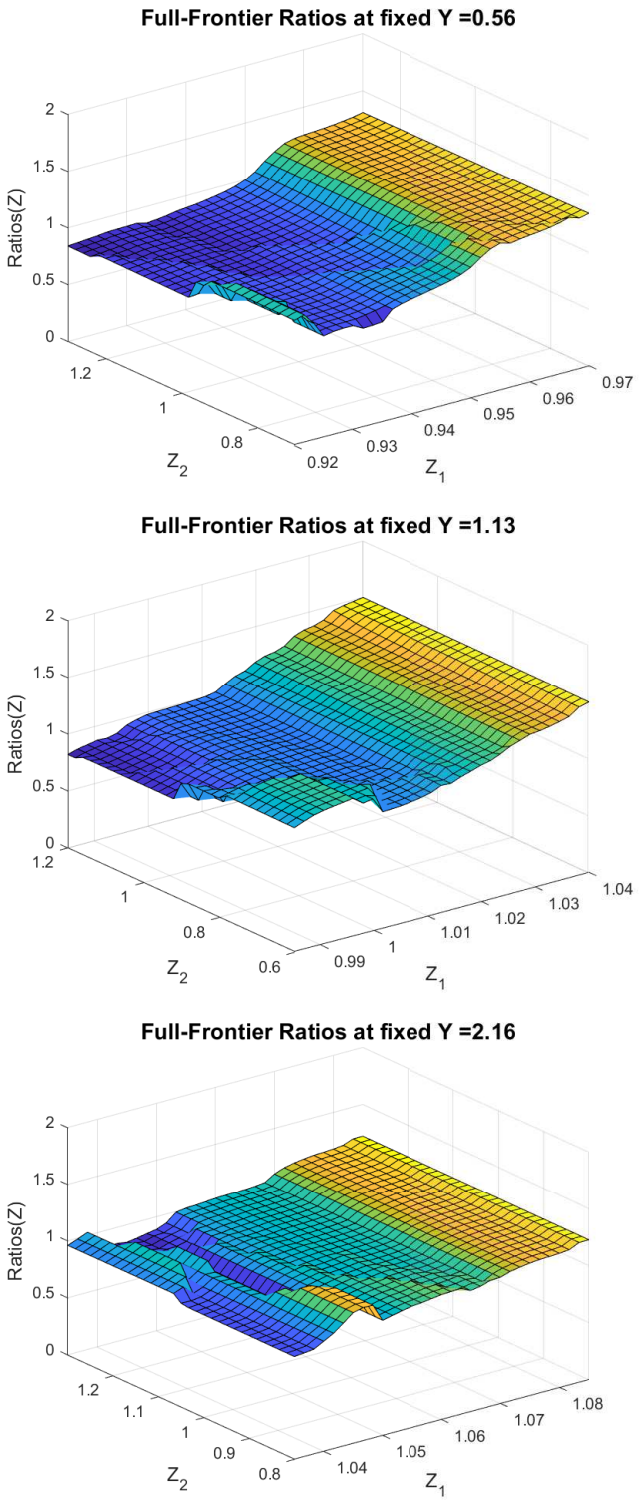


Figure 8: Bank example: Analysis of the ratios $\widehat{\tau}(y,z)/\widehat{\tau}(y)$ for fixed values of y. From top to bottom the 3 quartiles of Y.

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 prewhiten the variables (input and outputs) from the influence of these environmental factors in a first nonparametric stage by some control function type models. 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 semiparametric location-scale assumption.

Various Monte-Carlo experiments indicate that our procedure produce much more reliable estimators of the true frontier when compared to the direct approach. Finally we illustrate the practical use of our approach in a real data set, showing its great flexibility and usefulness.

References

[1] Aragon, Y. and Daouia, A. and Thomas-Agnan, C. (2005), Nonparametric Frontier Estimation: A Conditional Quantile-based Approach, *Econometric Theory*, 21, 358–389.

[2] Bădin, L., Daraio, C. 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.

[3] Bădin, L., Daraio, C. and L. Simar (2019), A Bootstrap Approach for Bandwidth Selection in Estimating Conditional Efficiency Measures, *European Journal of Operational Research*, 277, 784–797.

[4] Cazals, C. Florens, J.P. and L. Simar (2002), Nonparametric Frontier Estimation: a Robust Approach, *Journal of Econometrics*, 106, 1–25.

[5] Charnes, A., Cooper W.W. and E. Rhodes (1978), Measuring the inefficiency of decision making units, *European Journal of Operational Research* 2 (6), 429-444.

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

[7] Daouia, A. and A. Ruiz-Gazen (2006), Robust nonparametric frontier estimators: qualitative robustness and influence function, *Statitstica Sinica*, 16, 1233–1253.

[8] Daouia, A. and L. Simar (2007), Nonparametric efficiency analysis: a multivariate conditional quantile approach, *Journal of Econometrics*, 140, 375–400.

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

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

[11] Daraio, C., Simar, L. and P.W. 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.

[12] Debreu, G. (1951), The coefficient of resource utilization, *Econometrica*, 19:3, 273-292.

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

[14] Einmahl, J. and I. Van Keilegom (2008), Specification tests in nonparametric regression. *Journal of Econometrics*, 143, 88–102.

[15] Farrell, M.J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society*, A(120), 253-281.

[16] Florens, J.P., Simar, L. and I. Van Keilegom (2014), Frontier Estimation in Nonparametric Location-Scale Models, *Journal of Econometrics*, 178, 456–470.

[17] 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.

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

[19] Kneip, A., Simar, L. and P. W. Wilson (2015), When bias kills the variance: Central Limit Theorems for DEA and FDH efficiency scores, *Econometric Theory*, 31, 394–422.

[20] Li, Q., Lin, J. and J.S. Racine (2013), Optimal Bandwidth Selection for Nonparametric Conditional Distribution and Quantile Functions, *Journal of Business & Economic Statistics*, Vol 31 (1), 57–65.

[21] Park, B., Simar, L. and Ch. Weiner (2000), The FDH Estimator for Productivity Efficiency Scores: Asymptotic Properties, *Econometric Theory*, Vol 16, 855–877.

[22] Simar, L. (2003), Detecting Outliers in Frontiers Models: a Simple Approach, *Journal of Productivity Analysis*, 20, 391–424.

[23] Simar, L., Vanhems, A. and I. Van Keilegom (2016), Unobserved Heterogeneity and Endogeneity in Nonparametric Frontier Estimation, *Journal of Econometrics*, 190, 360–373.

[24] Simar, L. and P. Wilson (2007), Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes, *Journal of Econometrics*, vol 136, 1, 31–64.

[25] Simar, L. and P. W. Wilson (2015), Statistical Approaches for Nonparametric Frontier Models: A Guided Tour, *International Statistical Review*, 83, 1, 77–110.

[26] Van der Vaart, A.W. and J.A. and 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,*

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

and

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

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,*

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

and

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

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

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

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

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

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

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

for some $K, L < \infty$, provided $\sup_{y,z} f_{Y|Z}(y|z) < \infty$. This shows that $N_{[]}(\mathcal{F}, \epsilon, L_2(\mathbb{P})) \le M_\epsilon$, 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,

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

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

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

Finally we calculate

$$\begin{aligned}
&\mathrm{Var}\Big(\mathbb{1}(\widehat{F}_{Y|Z}(Y|Z)\ge u)-\mathbb{1}(F_{Y|Z}(Y|Z)\ge u)\Big)\\
&\le \mathbb{E}\Big(\mathbb{1}(\widehat{F}_{Y|Z}(Y|Z)\ge u)-\mathbb{1}(F_{Y|Z}(Y|Z)\ge u)\Big)^2\\
&=\int\Big[F_{Y|Z}(\widehat{F}^{-1}_{Y|Z}(u|z)|z)-F_{Y|Z}(\widehat{F}^{-1}_{Y|Z}(u|z)\wedge F^{-1}_{Y|Z}(u|z)|z)\Big]dF_Z(z)\\
&\quad+\int\Big[F_{Y|Z}(F^{-1}_{Y|Z}(u|z)|z)-F_{Y|Z}(\widehat{F}^{-1}_{Y|Z}(u|z)\wedge F^{-1}_{Y|Z}(u|z)|z)\Big]dF_Z(z)\\
&=\sup_{0\le u\le 1,z\in R_Z}|\widehat{F}^{-1}_{Y|Z}(u|z)-F^{-1}_{Y|Z}(u|z)|=o(1)\quad a.s.
\end{aligned}$$

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

Proof of Theorem 3.1.

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

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

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

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

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

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

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

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 $a^m - b^m = (a-b)(\sum_{k=0}^{m-1} c_k a^{m-1-k}b^k)$ 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

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

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:

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

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

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

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

Proof of Theorem 3.4. Let $i_1 = \operatorname{argmin}_j Z_j$, $i_2 = \operatorname{argmin}_{j: Z_j - Z_{i_1} > 2h}(Z_j - Z_{i_1})$, $i_3 = \operatorname{argmin}_{j: Z_j - Z_{i_2} > 2h}(Z_j - Z_{i_2})$, 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

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

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

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

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

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

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