SINGLE INDEX MODELS FOR NONPARAMETRIC CONDITIONAL FRONTIERS

Catherine Cazals, Jean-Pierre Florens, Léopold Simar

LIDAM Discussion Paper ISBA
2025 / 22



UCLouvain

ISBA | LIDAM
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

Single Index Models for Nonparametric Conditional Frontiers

Catherine CAZALS* Jean-Pierre FLORENS*
Léopold SIMAR*,§

November 2025

Abstract

In production theory, a lot of attention has been paid in the literature to the analysis of the effect of environmental variables on the efficiency of firms. The usual and natural way to investigate this issue is to consider conditional frontier models. For nonparametric approaches, this can create serious problems if the number of these potential environmental factors increases, exacerbating the curse of dimensionality characteristic of nonparametric models. In this paper, to address this issue, we investigate whether Single Index Models (SIM) could be used for modeling the effect of these variables on the production process. We propose a test for the SIM hypothesis and analyse the asymptotic properties. If the SIM model is not rejected, we obtain better rates of convergence of the conditional efficiency estimates. The paper investigates, through some Monte-Carlo experiments, the finite sample properties of the proposed test and the properties of the resulting estimates of the SIM when it is not rejected. We illustrate the method with a real data set from the French national postal operator in charge of universal service.[1]

Key Words: Nonparametric conditional frontier, Single-Index, Robust frontier, Environmental variables.

JEL Classification: C10, C14, C51, D22.

*Toulouse School of Economics (TSE), Toulouse, France, catherine.cazals@tse-fr.eu, and jean-pierre.florens@tse-fr.eu. J.P. Florens acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissement d'Avenir), grant ANR-17-EURE-0010.

§Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), LIDAM, UCLouvain, Belgium, leopold.simar@uclouvain.be

[1]The authors acknowledge "La Poste", the French national postal operator for providing the anonymized data that illustrate our methodology.

1 Introduction

1.1 The setup and basic notations

In production theory and efficiency analysis, firms produce a set of outputs $Y \in \mathbb{R}^q_+$ by using a set of inputs $X \in \mathbb{R}^p_+$ and the efficiency of a firm operating at the level (x, y) is measured relative to the distance of this point to the efficient boundary of the production set. To formalize we need to define a model that characterizes the way inputs an outputs are generated. We define first the attainable (or production) set Ψ as

$$\Psi = \{(x, y) \mid x \text{ can produce } y\}. \tag{1.1}$$

Under the free disposability assumption,[1] this set is the support of the joint r-variate variable (X, Y), with $r = p + q$. The Data Generating Process (DGP) is then fully determined by the joint distribution of (X, Y). One natural way to characterize this distribution is

$$H_{XY}(x, y) = \text{Prob}(X \leq x, Y \geq y), \tag{1.2}$$

which is the probability for a production plan (x, y) to be dominated. Then

$$\Psi = \{(x, y) \mid H_{XY}(x, y) > 0\}. \tag{1.3}$$

The efficient boundary of Ψ, that is the technology, can be defined as

$$\Psi^\partial = \{(x, y) \in \Psi \mid (\gamma^{-1}x, \gamma y) \notin \Psi, \text{ for any } \gamma > 1\}. \tag{1.4}$$

The technical efficiency of the points can be measured according various directions. To save space we only consider here the Debreu (1951) and Farrell (1957) input measure of technical efficiency $\theta(x, y)$ as the minimal radial contraction of the inputs to project the point (x, y) on the efficient boundary:[2]

$$\theta(x, y) = \inf\{\theta \mid (\theta x, y) \in \Psi\} = \inf\{\theta \mid H_{XY}(\theta x, y) > 0\}. \tag{1.5}$$

Since $H_{XY}(x, y) = F_{X|Y}(x|Y \geq y)S_Y(y)$ where $S_Y(y) = \text{Prob}(Y \geq y)$ we can also define, for all y is in the support of Y, the input oriented efficiency score as

$$\theta(x, y) = \inf\{\theta \mid F_{X|Y}(\theta x|Y \geq y) > 0\}. \tag{1.6}$$

[1] The free disposability means that if $(x_1, y_1) \in \Psi$, then $(x_2, y_2) \in \Psi$ for any (x_2, y_2) such that $x_2 \geq x_1$ and $y_2 \leq y_1$. This is a minimal assumption of the production process, which assumes that it is always possible to waste resources (even if not economically sounded) and implies monotonicity of the technology.

[2] Variations are the output orientation, the hyperbolic distances or the directional distances, see e.g. Simar and Wilson (2015) for a detailed survey.

So, in a sense, choosing the input orientation assumes implicitly to consider Y as exogenous.

The coordinates of the projection of (x, y) on the efficient boundary are given by $(x^{\partial}(y), y)$, where $x^{\partial}(y) = \theta(x, y)x$. In the particular case (but very often used in applied works where X is a cost) of $p = 1$, we can define the minimal cost function

$$x^{\partial}(y) = \varphi(y) = \inf\{x \mid F_{X|Y}(x|Y \geq y) > 0\}. \tag{1.7}$$

In production theory a lot of attention has been paid in the literature on the analysis of the effect of environmental factors on the efficiency of the firms. These factors, denoted by $Z \in \mathcal{Z} \subset \mathbb{R}^d$ are neither inputs nor outputs and are typically not under the control of the firm, but they may influence the production process. The may affect the shape of the technology Ψ^{∂}, or they may affect the distribution of the inefficiencies inside Ψ or they may affect both. Therefore the DGP described above has to be augmented to take into account for these factors. A natural way suggested by Cazals et al. (2002) and Daraio and Simar (2005) is to consider conditional DGP, i.e. condition all the concepts defined above to a given value of Z. Namely, the conditional distribution of (X, Y) conditional to $Z = z$ process is defined as

$$H_{XY|Z}(x, y|z) = \text{Prob}(X \leq x, Y \geq y \mid Z = z), \tag{1.8}$$

and the conditional attainable set is now the support of $H_{XY|Z}(x, y|z)$ and is the set of attainable combinations of (x, y) when $Z = z$:

$$\Psi^z = \{(x, y) \mid H_{XY|Z}(x, y|z) > 0, \text{ i.e. } Z = z \text{ and } x \text{ can produce } y\}. \tag{1.9}$$

Clearly, for all $z \in \mathcal{Z}$, $\Psi^z \subseteq \Psi$ and $\Psi = \cup_{z \in \mathcal{Z}} \Psi^z$ and the input oriented conditional measure of efficiency is given by

$$\theta(x, y|z) = \inf\{\theta \mid (\theta x, y) \in \Psi^z\} = \inf\{\theta \mid H_{XY|Z}(\theta x, y|z) > 0\}. \tag{1.10}$$

Again, for the input orientation, for y in the conditional support of Y, i.e. $S_{Y|Z}(y|z) = \text{Prob}(Y \geq y|Z = z) > 0$ we have

$$\theta(x, y|z) = \inf\{\theta \mid F_{X|YZ}(\theta x|Y \geq y, Z = z) > 0\}, \tag{1.11}$$

so we consider Y as being exogenous, conditional on Z. The coordinates of the projection of a point (x, y) on the efficient boundary of Ψ^z are given by $(x^{\partial}(y, z), y)$ where $x^{\partial}(y, z) = \theta(x, y|z)x$. In the particular case of X being a univariate cost, we have a conditional minimal cost function

$$x^{\partial}(y, z) = \varphi(y, z) = \inf\{x \mid F_{X|YZ}(x|Y \geq y, Z = z) > 0\}. \tag{1.12}$$

Simar and Wilson (2007) discuss about the "separability" condition often assumed by researchers, which poses that for all $z \in \mathcal{Z}$, $\Psi^z = \Psi$, so that Z may only affect the distribution of inefficiencies inside Ψ. This is often an empirical question and Daraio et al. (2018) provide a way to test this restrictive assumption.

1.2 Nonparametric estimation

In practice, we do not know Ψ or Ψ^z and all the derived elements, like the efficiency scores, so we have to estimate these quantities from a random sample of input-output combinations, augmented by the environmental variables. This provides a set $\mathcal{X}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$, from which the above quantities have to be estimated. A large literature has been devoted to the nonparametric estimation of these quantities, they are mainly based on envelopment estimators (see again Simar and Wilson, 2015, for a survey). Since we do not want to impose the convexity assumption, we will focus here on the most general Free Disposal Hull (FDH) estimators, introduced by Deprins et al. (1984), which only requires the free disposability assumption. It has been shown that these estimators are obtained by replacing in all the above definition H_{XY} and $H_{XY|Z}$ by some nonparametric estimators. We have

$$\widehat{H}_{XY}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y) \tag{1.13}$$

$$\widehat{H}_{XY|Z}(x,y|z) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x, Y_i \geq y) K\left(\frac{Z_i - z}{h}\right)}{\sum_{i=1}^n K\left(\frac{Z_i - z}{h}\right)}, \tag{1.14}$$

with $h = (h_1, \ldots, h_d)$ being a vector of d bandwidths and $K(\cdot)$ is a product kernel. Optimal bandwidths can be obtained by Least Squares Cross Validation (LSCV) and the properties of the conditional estimator have been derived in Li et al. (2013). Typically these bandwidths are of order $O(n^{-1/(d+4)})$.

For the input orientation we need only $\widehat{F}_{X|Y}(\theta x|Y \geq y) = \widehat{H}_{XY}(x,y)/\widehat{H}_{XY}(0,y)$ and $\widehat{F}_{X|YZ}(\theta x|Y \geq y, Z = z) = \widehat{H}_{XY|Z}(x,y|z)/\widehat{H}_{XY|Z}(0,y|z)$. Hence the nonparametric efficiency estimators are given by

$$\widehat{\theta}(x,y) = \inf\{\theta \mid \widehat{F}_{X|Y}(\theta x|Y \geq y) > 0\} = \min_{Y_i \geq y} \max_{j=1,\ldots,p} \left(X_i^j / x^j\right) \tag{1.15}$$

$$\widehat{\theta}(x,y|z) = \inf\{\theta \mid \widehat{F}_{X|YZ}(\theta x|Y \geq y|, Z = z) > 0\} = \min_{Y_i \geq y, ||Z_i - z|| \leq h} \max_{j=1,\ldots,p} \left(X_i^j / x^j\right), \tag{1.16}$$

so that the conditional FDH can be viewed as a localized version of the FDH estimator for data points (X_i, Y_i) such that Z_i is in an h-neighborhood of z.

The properties of the FDH estimators are well known and established in Park et al. (2000) and in Jeong et al. (2010) for the conditional FDH. To summarize, under mild regularity

conditions,

$$n^{1/(p+q)}\big(\widehat{\theta}(x,y)-\theta(x,y)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\eta(x,y)) \tag{1.17}$$

$$(n\bar{h})^{1/(p+q)}\big(\widehat{\theta}(x,y|z)-\theta(x,y|z)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\eta_c(x,y,z)), \tag{1.18}$$

where $\bar{h}=\prod_{j=1}^{d} h_j$ and $\eta(x,y)$ and $\eta_c(x,y,z)$ are parameters that depend of the features of the DGP (curvature of the boundary, density of (x,y) near the boundary point). Bootstrap techniques using subsampling can be used for practical inference (see Jeong and Simar, 2006). We see clearly that the curse of the dimensionality impacts the rates of convergence, already for the FDH estimators when $r=p+q$ increases, but the situation deteriorates even worse for the conditional FDH when d increases. To see the latter, with optimal order for the bandwidths, $\bar{h}\propto n^{-d/(d+4)}$, we see that the value for n becomes $n^{4/(d+4)}$ when going from FDH rates to conditional to Z ones.

1.3 Partial frontier order-m

The FDH estimators fully envelop the sample of observations in the (X,Y) space. Consequently, these estimators are very sensitive to outliers or extreme data points. Cazals et al. (2002) introduce the concept of "partial" frontier that provides a less extreme benchmark than the full frontier of the support of (X,Y). For the input oriented case,[3] this leads to the concept of order-m efficiency scores defined as

$$\theta_m(x,y)=\int_0^\infty \big[1-F_{X|Y}(ux|Y\geq y)\big]^m\,du=\theta(x,y)+\int_{\theta(x,y)}^\infty \big[1-F_{X|Y}(ux|Y\geq y)\big]^m\,du, \tag{1.19}$$

where $F_{X|Y}(x|Y\geq y)=\text{Prob}(X\leq x|Y\geq y)$ is derived from $H_{XY}(x,y)$. Clearly when $m\to\infty$, $\theta_m(x,y)\to\theta(x,y)$. As explained in Cazals et al. (2002), the order-m efficiency score is the expectation of the FDH efficiency score obtained from a random sample build by m random draws of units generated according $F_{X|Y}(\cdot|Y\geq y)$.

In the particular case of a cost function ($p=1$) this defines the order-m cost frontier:

$$\varphi_m(y)=\mathbb{E}\left[\min(X_1,\ldots,X_m)\right]=\int_0^\infty \big[1-F_{X|Y}(x|Y\geq y)\big]^m\,dx, \tag{1.20}$$

i.e. the expected minimum of the costs of m units drawn at random from the population of units producing more output than y.

A natural nonparametric estimator of $\theta_m(x,y)$ is obtained by plugging the empirical version of $F_{X|Y}(\cdot|Y\geq y)$ in the equations above. The nice feature is that this estimator

[3] Cazals et al. (2002) present also the output oriented case, see Wilson (2011), for the hyperbolic distances and Simar and Vanhems (2012) for the directional distance cases.

preserves the parametric $\sqrt{n}$-rate of convergence (we avoid the curse of the dimensionality due to r) and we have a Gaussian limiting distribution. For instance we have

$$\sqrt{n}\big(\widehat{\theta}_m(x,y) - \theta_m(x,y)\big) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{m,x,y}) \tag{1.21}$$

where an expression for $\sigma^2_{m,x,y}$ is given in Cazals et al. (2002).

The conditional to Z version of the order-m efficiency score follows the same scheme. We have

$$\begin{aligned}\theta_m(x,y|z) &= \int_0^\infty \big[1 - F_{X|YZ}(ux|Y \geq y, Z = z)\big]^m \, du \\ &= \theta(x,y|z) + \int_{\theta(x,y|z)}^\infty \big[1 - F_{X|YZ}(ux|Y \geq y, Z = z)\big]^m \, du, \end{aligned} \tag{1.22}$$

where $F_{X|YZ}(x|Y \geq y, Z = z) = \text{Prob}(X \leq x|Y \geq y, Z = z)$ is derived from $H_{XY|Z}(x,y|z)$. Again we have, as $m \to \infty$, $\theta_m(x,y|z) \to \theta(x,y|z)$. Note again that for a cost function ($p = 1$) the conditional order-m cost frontier is

$$\varphi_m(y,z) = \mathbb{E}\left[\min(X_1, \ldots, X_m \mid Z = z\right] = \int_0^\infty \big[1 - F_{X|YZ}(x|Y \geq y, Z = z)\big]^m \, dx, \tag{1.23}$$

where $X_1, \ldots, X_m$ is a random sample of size m from the population of units producing more output than y, and facing the environmental conditions z.

Nonparametric estimators are provided by plugging the nonparametric estimator $\widehat{H}_{XY|Z}(x,y|z)$ defined in (1.14) and $\widehat{F}_{X|YZ}(x|Y \geq y, Z = z) = \widehat{H}_{XY|Z}(x,y|z)/\widehat{H}_{XY|Z}(0,y|z)$. Here, as shown in Cazals et al. (2002), we keep the asymptotic normality as in (1.21) but the rate of convergence is now $\sqrt{n\bar{h}^d}$. When optimal bandwidths are used this provides a rate $\sqrt{n^{4/(d+4)}}$, which deteriorates quickly if d increases.

1.4 How to reduce the impact of the dimension of Z.

The aim of this paper is to provide a tool for reducing the curse of the dimensionality due to the dimension d of Z. We will investigate if we can use a Single Index Models (SIM) to reduce the impact of Z in only one dimension in place of d. In applied works in productivity analysis, it is often the case that the number of inputs and of outputs can be limited, or even when we observe high collinearity, we can reduce the dimension of the input/output space (see Daraio and Simar, 2007 and Wilson, 2018). For the environmental factors, the practitioner can easily think about many potential variables Z, but at a cost of losing precision in the estimation process.

SIM are a natural way to try to reduce the dimension, but we have to check first is this dimension reduction is valid. Since the main object of interest, for the input orientation, is

$F_{X|YZ}(x|Y \geq y, Z = z)$, the first idea of this paper is to provide a test of the SIM hypothesis for this conditional cdf, i.e. we want to test the null hypothesis

$$H_0 : \text{ for all } (x,y), F_{X|YZ}(x|Y \geq y, Z) = F(x|Y \geq y, \beta' Z), \text{ for some } \beta \in \mathbb{R}^d, a.s., \quad (1.24)$$

where $F(\cdot|Y \geq y, \beta' Z)$ denotes the conditional cdf of X conditional on $Y \geq y$ and $\beta' Z$. This is in fact a test of conditional independence

$$H_0 : \; X \underline{\perp\!\!\!\perp} Z \mid Y \geq y, \beta' Z, \text{ for some } \beta. \quad (1.25)$$

As usually the case for SIM, we have to normalize β for identification purpose. We will impose $||\beta|| = 1$, with $\beta_1 \geq 0$ since the sign of β is undetermined. We will adapt to our setup the approach described in Delgado and Gonzales Manteiga (2001) (hereafter DGM) and derive a test statistic. As in DGM, we will use the bootstrap for practical inference. The methodological originality of our paper is that it addresses the specification test when conditioning on inequalities ($Y \geq y$) and not on equaities ($Y = y$) as in DGM.

The paper is organized as follows Section 2 introduces the proposed test statistic and analyze its asymptotic properties, including the asymptotic properties of the estimator of the SIM when the null is true. It introduce also a bootstrap algorithm to approximate these asymptotic distributions. Section 3 investigates the finite sample properties by Monte-Carlo experiments, showing that as expected, the tests have appropriate size and good power, with increasing quality when the sample size increases. We observe also that, when the SIM is true, the estimators of its parameters behave as expected by the theory. We illustrate also the method wit a real data set from the French Post. Section 4 concludes. In Appendix A we explain how to simulate, in our setup, data when imposing a structure on a conditional distribution relative to an inequality condition.

2 Testing the SIM for the Conditional Production Process

2.1 Basic Statistic

The null hypothesis (1.24) can equivalently be expressed for any (x, y) in terms of the "tested" residual $\varepsilon(X, Z; x, y, \beta)$ defined as

$$\varepsilon(X, Z; x, y, \beta) = \mathbb{1}(X \leq x) - \mathbb{E}\left[\mathbb{1}(X \leq x)|Y \geq y, \beta' Z)\right] \quad (2.1)$$

$$= \mathbb{1}(X \leq x) - F(x|Y \geq y, \beta' Z). \quad (2.2)$$

Hence, we can rewrite H_0 as

$$H_0 : \forall (x,y), \; \mathbb{E}\left[\varepsilon(X, Z; x, y, \beta)|Y \geq y, Z\right] = 0, \text{ for some } \beta \in \mathbb{R}^d, a.s. \quad (2.3)$$

where the expectation is over the values of X. We can say that under H_0, $\varepsilon(X, Z; x, y, \beta)$ is mean-independent of $Y \geq y, Z$, for some β and $\forall(x, y)$. Note that the hypothesis H_0 is equivalent to[4]

$$H_0 : \forall(x, y, t),\ \mathbb{E}\left[\varepsilon(X, Z; x, y, \beta)\, \mathbb{1}(Y \geq y) e^{\iota t' Z}\right] = 0,\ \text{for some } \beta \in \mathbb{R}^d, \tag{2.4}$$

where ι is the imaginary unit, i.e. $\iota^2 = -1$ and now the expectation is over the values of (X, Y, Z). Under the assumption that for all y in the support of Y (i.e. $S_Y(y) > 0$), $f(\beta' Z | Y \geq y) > 0$, *a.s.*, where f is the conditional pdf of $\beta' Z$ given $Y \geq y$, we can weight the residual by $f(\beta' Z | Y \geq y) S_Y(y)$ and define the random variable

$$U(x, y, t, \beta) = \mathbb{E}\left[f(\beta' Z | Y \geq y) S_Y(y) \varepsilon(X, Z; x, y, \beta)\, \mathbb{1}(Y \geq y) e^{\iota t' Z}\right]. \tag{2.5}$$

Weighting the residual is for practical reasons, in order to avoid below random denominator is some conditional expectation. Now the null hypothesis can be rewritten as

$$H_0 : \forall(x, y, t),\ U(x, y, t, \beta) = 0,\ \text{for some } \beta \in \mathbb{R}^d \tag{2.6}$$

The test statistic below will be defined as suitable functions of an estimator of U.

The empirical version of U from a sample of observations $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ can be obtained as follows. An estimate of the residual $\varepsilon(X, Z; x, y, \beta)$ for the observation (X_i, Z_i) is given by

$$\widehat{\varepsilon}(X_i, Z_i; x, y, \beta) = \mathbb{1}(X_i \leq x) - \widehat{F}(x | Y \geq y, \beta' Z_i) \tag{2.7}$$

where

$$\widehat{F}(x | Y \geq y, \beta' Z_i) = \frac{\sum_{j=1}^n \mathbb{1}(X_j \leq x)\, \mathbb{1}(Y_j \geq y) K_{ij}^{\beta,h}}{\sum_{j=1}^n \mathbb{1}(Y_j \geq y) K_{ij}^{\beta,h}} \tag{2.8}$$

$$K_{ij}^{\beta,h} = K\big(\beta'(Z_i - Z_j)/h\big), \tag{2.9}$$

with $K(\cdot)$ being an univariate (symmetric) kernel and $h \in \mathbb{R}_+$ is a bandwidth. This residual estimate will be weighted by the estimate $\widehat{f}(\beta' Z_i | Y \geq y)$ given by

$$\widehat{f}(\beta' Z_i | Y \geq y) = \frac{(nh)^{-1} \sum_{j=1}^n \mathbb{1}(Y_j \geq y) K_{ij}^{\beta,h}}{n^{-1} \sum_{j=1}^n \mathbb{1}(Y_j \geq y)}, \tag{2.10}$$

and finally, $\widehat{S}_Y(y) = n^{-1} \sum_{j=1}^n \mathbb{1}(Y_j \geq y)$. So that the empirical version of the basic quantity, $U(x, y, t, \beta)$, is given by

$$\widehat{U}_n(x, y, t, \beta) = \frac{1}{n} \sum_{i=1}^n \left\{\widehat{f}(\beta' Z_i | Y \geq y) \widehat{S}_Y(y) \widehat{\varepsilon}(X_i, Z_i, x, y, \beta)\, \mathbb{1}(Y_i \geq y) e^{\iota t' Z_i}\right\}. \tag{2.11}$$

[4]This equivalence comes from the fact that for any two random variables A and B, $\mathbb{E}(A|B) = 0$ if and only if for any measurable function $g(B)$, $\mathbb{E}(g(B)A) = 0$.

By plugging the estimators defined above we have,

$$\widehat{U}_n(x,y,t,\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{nh}\sum_{j=1}^{n}\left[\mathbb{1}(Y_j \geq y)K_{ij}^{\beta,h}\Big(\mathbb{1}(X_i \leq x) - \frac{\sum_{j=1}^{n}\mathbb{1}(X_j \leq x)\mathbb{1}(Y_j \geq y)K_{ij}^{\beta,h}}{\sum_{j=1}^{n}\mathbb{1}(Y_j \geq y)K_{ij}^{\beta,h}}\Big)\right] \times \mathbb{1}(Y_i \geq y)e^{\iota t' Z_i}\right\}. \tag{2.12}$$

After some algebraic simplifications this becomes

$$\widehat{U}_n(x,y,t,\beta) = \frac{1}{n^2 h}\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}K_{ij}^{\beta,h}\,\mathbb{1}(Y_j \geq y)\,\mathbb{1}(Y_i \geq y)e^{\iota t' Z_i}\left[\mathbb{1}(X_i \leq x) - \mathbb{1}(X_j \leq x)\right]\right\}. \tag{2.13}$$

2.2 Properties of $\widehat{U}_n$

We can prove that at an order $O(h^2)$, $\widehat{U}_n$ is a nondegenerate U-statistic. We can indeed rewrite (2.13) as $\widehat{U}_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi(W_i,W_j)$, where $W = (X,Y,Z)$ and $\phi(W_i,W_j) = h^{-1}K_{ij}^{\beta,h}\,\mathbb{1}(Y_j \geq y)\,\mathbb{1}(Y_i \geq y)e^{\iota t' Z_i}\left[\mathbb{1}(X_i \leq x) - \mathbb{1}(X_j \leq x)\right]$. We note that $\phi(W_i,W_i) = 0$. Now we symmetrize the function ϕ by defining

$$\begin{aligned}\phi_s(W_i,W_j) =& \frac{1}{2}\left[\phi(W_i,W_j) + \phi(W_j,W_i)\right] \\ =& \frac{1}{2h}K_{ij}^{\beta,h}\,\mathbb{1}(Y_j \geq y)\,\mathbb{1}(Y_i \geq y)\left[\mathbb{1}(X_i \leq x) - \mathbb{1}(X_j \leq x)\right]\left(e^{\iota t' Z_i} - e^{\iota t' Z_j}\right).\end{aligned} \tag{2.14}$$

Now we have

$$\widehat{U}_n(x,y,t,\beta) = \frac{2}{n(n-1)}\sum_{i<j}\phi_s(W_i,W_j). \tag{2.15}$$

Note that the functions ϕ and ϕ_s are indexed by the value of (x,y,t,β) but we don't explicit this in the notation for clarity. We will now prove the following properties.

Property 2.1. *Under the null Hypothesis* (2.6) *we have*

$$(i)\ \mathbb{E}(\phi_s(W_i,W_j)) = \mathbb{E}(\phi(W_i,W_j)) = O(h^2) \tag{2.16}$$

$$(ii)\ \mathbb{E}(\phi_s(w_i,W_j)) = \phi_1(w_i) \neq 0. \tag{2.17}$$

Proof. We prove (i) by computing to $\mathbb{E}(\phi(W_i,W_j))$. We first integrate on (X_i,X_j) conditionally on $(Y_i \geq y, Y_j \geq y, Z_i, Z_j)$ (for $Y < y$, $\phi = 0$). Under the null H_0 and due to the independence between W_i and W_j, we have

$$\begin{aligned}\mathbb{E}_{X_i,X_j|Y_i\geq y,Y_j\geq y,Z_i,Z_j}(\phi(W_i,W_j)) =& h^{-1}K_{ij}^{\beta,h}\,\mathbb{1}(Y_j \geq y)\,\mathbb{1}(Y_i \geq y)e^{\iota t' Z_i} \\ & \times \left[F(x|Y \geq y, \beta' Z_i) - F(x|Y \geq y, \beta' Z_j)\right]\end{aligned}$$

Now taking the expectation with respect to (Y_i, Y_j) conditionally on (Z_i, Z_j) we have

$$\begin{aligned}\mathbb{E}_{X_i,X_j,Y_i,Y_j|Z_i,Z_j}(\phi(W_i,W_j)) =& h^{-1}K_{ij}^{\beta,h}S_Y(y|Z_i)S_Y(y|Z_j)e^{\iota t'Z_i} \\ & \times \left[F(x|Y\geq y,\beta'Z_i) - F(x|Y\geq y,\beta'Z_j)\right].\end{aligned}$$

Now for computing the expectation w.r.t. the variables Z we will use the general property that $\mathbb{E}_Z = \mathbb{E}_{\beta'Z}\big[\mathbb{E}_{Z|\beta'Z}\big]$. So we first compute the expectation of the last expression w.r.t. (Z_i, Z_j) conditionally on $(\beta'Z_i, \beta'Z_j)$. Here we note that

$$\mathbb{E}_{Z_j|\beta'Z_j}\big[S_Y(y|Z_j)\big] = S_Y(y|\beta'Z_j) \tag{2.18}$$

and that

$$\begin{aligned}\mathbb{E}_{Z_i|\beta'Z_i}\big[S_Y(y|Z_i)e^{\iota t'Z_i}\big] =& \mathbb{E}_{Z_i|\beta'Z_i}\big[S_Y(y|Z_i)\big] \times \mathbb{E}_{Z_i|\beta'Z_i}\big[e^{\iota t'Z_i}|S_Y(y|Z_i)\big] &(2.19)\\ =& S_Y(y|\beta'Z_i) \times g(R_i,\beta'Z_i), &(2.20)\end{aligned}$$

where $Z_i = \beta'Z_i i_d + R_i \in \mathbb{R}^d$ and $g(R_i,\beta'Z_i) = \mathbb{E}_{Z_i|\beta'Z_i}\big[e^{\iota t'Z_i}|S_Y(y|Z_i)\big] \in \mathbb{R}$. So we obtain

$$\begin{aligned}\mathbb{E}_{W_i,W_j|\beta'Z_i,\beta'Z_j}(\phi(W_i,W_j)) =& h^{-1}K_{ij}^{\beta,h}g(R_i,\beta'Z_i)S_Y(y|\beta'Z_i)S_Y(y|\beta'Z_j) \\ & \times \left[F(x|Y\geq y,\beta'Z_i) - F(x|Y\geq y,\beta'Z_j)\right]. \end{aligned}\tag{2.21}$$

For the final integration in $(\beta'Z_i, \beta'Z_j)$, we first compute the expectation relative to $\beta'Z_j$ conditionally on $\beta'Z_i$. It is known that for symmetric kernel with finite second moment, under appropriate smoothness conditions, we have

$$\mathbb{E}_{\beta'Z_j}\left[G(\beta'Z_j)h^{-1}K\big((\beta'Z_i - \beta'Z_j)/h\big)\right] = f(\beta'Z_i)G(\beta'Z_i) + O(h^2), \tag{2.22}$$

where $f(\cdot)$ is the density of $\beta'Z$. Hence plugging this in the above last expression (noting that Z_j is independent of Z_i), we have

$$\mathbb{E}_{W_i,W_j|\beta'Z_i}(\phi(W_i,W_j)) = f(\beta'Z_i)g(R_i,\beta'Z_i)\big[S_Y(y|\beta'Z_i)\big]^2 \times 0 + O(h^2) = O(h^2).$$

The final integration in $\beta'Z_i$ does not change the result and we have proven (i).

The proof of (ii) follows a similar development. Here we fix $W_i = w_i = (x_i, y_i, z_i)$ and now the expectation is only on W_j. We have, under H_0,

$$\begin{aligned}\mathbb{E}_{X_j|Y_j\geq y,Z_j}(\phi_s(w_i,W_j)) =& (2h)^{-1}K\big(\beta'(z_i - Z_j)/h\big)\mathbb{1}(Y_j\geq y)\mathbb{1}(y_i\geq y) \\ & \times \Big[e^{\iota t'Z_j}\big(F(x|Y\geq y,\beta'Z_j) - \mathbb{1}(x_i\leq x)\big) + e^{\iota t'z_i}\big(\mathbb{1}(x_i\leq x) - F(x|Y\geq y,\beta'Z_j)\big)\Big].\end{aligned}$$

Then taking the expectation on Y_j conditionally on Z_j, we have

$$\begin{aligned}\mathbb{E}_{X_j,Y_j|Z_j}(\phi_s(w_i,W_j)) =& (2h)^{-1}K\big(\beta'(z_i - Z_j)/h\big)\mathbb{1}(y_i\geq y) \\ & \times \Big[S_Y(y|Z_j)e^{\iota t'Z_j}\big(F(x|Y\geq y,\beta'Z_j) - \mathbb{1}(x_i\leq x)\big) + \\ & \quad e^{\iota t'z_i}S_Y(y|Z_j)\big(\mathbb{1}(x_i\leq x) - F(x|Y\geq y,\beta'Z_j)\big)\Big],\end{aligned}$$

where we used the identity $\mathbb{E}\big(\mathbb{1}(Y_j \geq y)e^{\imath t' Z_j}|Z_j\big) = S_Y(y|Z_j)e^{\imath t' Z_j}$. As above noting that $\mathbb{E}_{Z_j} = \mathbb{E}_{\beta' Z_j}\big[\mathbb{E}_{Z_j|\beta' Z_j}\big]$ to compute the expectation reative to Z_j we first compute $\mathbb{E}_{Z_j|\beta' Z_j}$. We have

$$\begin{aligned}\mathbb{E}_{W_j|\beta' Z_j}(\phi_s(w_i, W_j)) = (2h)^{-1}K\big((\beta' z_i - \beta' Z_j)/h\big)\,\mathbb{1}(y_i \geq y)\\ \times \Big[\mathbb{E}\big(S_Y(y|Z_j)e^{\imath t' Z_j}|\beta' Z_j\big)\big(F(x|Y \geq y, \beta' Z_j) - \mathbb{1}(x_i \leq x)\big) +\\ e^{\imath t' z_i} S_Y(y|\beta' Z_j)\big(\mathbb{1}(x_i \leq x) - F(x|Y \geq y, \beta' Z_j)\big)\Big].\end{aligned}$$

For the final integration over $\beta' Z_j$ we use again the property of kernels given in (2.22) and we obtain

$$\begin{aligned}\mathbb{E}_{W_j}(\phi_s(w_i, W_j)) = (2)^{-1} f(\beta' z_i)\,\mathbb{1}(y_i \geq y)\\ \times \Big[\mathbb{E}\big(S_Y(y|Z_j)e^{\imath t' Z_j}|\beta' Z_j = \beta' z_i\big)\big(F(x|Y \geq y, \beta' z_j) - \mathbb{1}(x_i \leq x)\big) +\\ e^{\imath t' z_i} S_Y(y|\beta' z_i)\big(\mathbb{1}(x_i \leq x) - F(x|Y \geq y, \beta' z_i)\big)\Big] + O(h^2).\end{aligned}$$

This can be rewritten as

$$\begin{aligned}\mathbb{E}_{W_j}(\phi_s(w_i, W_j)) = (2)^{-1} f(\beta' z_i)\,\mathbb{1}(y_i \geq y)\,[\mathbb{1}(x_i \leq x) - F(x|Y \geq y, \beta' z_i)]\\ \times \Big[e^{\imath t' z_i} S_Y(y|\beta' z_i) - \mathbb{E}\big(S_Y(y|Z_j)e^{\imath t' Z_j}|\beta' Z_j = \beta' z_i\big)\Big] + O(h^2). \qquad (2.23)\end{aligned}$$

The factor at the first line of (2.23) is clearly not equal to zero and the same is true for the second factor at the second line. Note that even if $S_Y(y|z) = S_Y(y|\beta' z)$, this second factor becomes $S_Y(y|\beta' z_i)\left[e^{\imath t' z_i} - \mathbb{E}\big(e^{\imath t' Z_j}|\beta' Z_j = \beta' z_i\big)\right]$ which cannot be equal to zero, if we assume that the covariances matrix of Z is of full rank. This completes the proof of (ii). □

Note that it is easy to verify that $\mathbb{E}(\phi_1(W_i)) = 0 + O(h^2)$ as it should by (2.16). Indeed the expectation of the first line of (2.23) is identically equal to zero.

Now from this results we know that $\widehat{U}_n$ is a non-degenerate centered (at the order $O(h^2)$ U-statistic, indexed by (x, y, β, t). So we have the following property (see e.g. Theorem 12.3 in van der Vaart, 1998)

Property 2.2. *Under the null Hypothesis* (2.6) *and if* $\sqrt{n}h^2 = o(1)$*, we have as* $n \to \infty$

$$\sqrt{n}\,\widehat{U}_n \xrightarrow{\mathcal{L}} N(0, \sigma^2), \qquad (2.24)$$

where $\sigma^2 = 4\mathbb{E}\left[\phi_1(W_i)^2\right]$ *is also indexed by* (x, y, β, t).

2.3 Deriving a Test Statistic

We know that under the null H_0, $U(x, y, t, \beta) = 0$ for some β and for all (x, y, t). Therefore, for any positive weight function (e.g. a d-variate density function) $\pi(\cdot)$ on $\mathbb{R}^d$, we can define the random variable

$$T_\beta(x, y) = \int_t \{U(x, y, t, \beta)\}^2 \pi(t)dt, \tag{2.25}$$

where the integral is d-dimensional. Hence we can rewrite the null hypothesis as

$$H_0 : \forall(x, y) \text{ and for some } \beta,\ T_\beta(x, y) = 0, \tag{2.26}$$

By simple algebraic manipulations, noting that the square of a complex number is the product of the number by its complex conjugate, the integral in H_0 can be written as

$$\int_t \Big(\int_z V(x, y, \beta, z)e^{\iota t'z} f(z)dz \Big) \Big(\int_{\tilde{z}} V(x, y, \beta, \tilde{z})e^{-\iota t'\tilde{z}} f(\tilde{z})d\tilde{z} \Big) \pi(t)dt,$$

where $f(\cdot)$ is the pdf of Z and where

$$V(x, y, \beta, z) = \mathbb{E}\left[f(\beta' Z|Y \geq y)S_Y(y)\varepsilon(X, Z; x, y, \beta)\, \mathbb{1}(Y \geq y)|Z = z \right]. \tag{2.27}$$

This is equivalent to

$$\int_z \int_{\tilde{z}} \int_t V(x, y, \beta, z)f(z)V(x, y, \beta, \tilde{z})f(\tilde{z})e^{\iota t'(z-\tilde{z})}\pi(t)dtdzd\tilde{z},$$

so that equation (2.26) can be written in terms of $T_\beta(x, y)$ defined as

$$T_\beta(x, y) = \int_z \int_{\tilde{z}} V(x, y, \beta, z)f(z)V(x, y, \beta, \tilde{z})f(\tilde{z})\mathcal{F}_\pi(z - \tilde{z})dzd\tilde{z},$$

where $\mathcal{F}_\pi(\zeta)$ is the characteristic function of the density $\pi(t)$. For instance if $\pi(t)$ is the standard multivariate normal density (our choice in what follows), $\mathcal{F}_\pi(\zeta) = e^{-\zeta'\zeta/2}$. Finally we obtain

$$\begin{aligned} T_\beta(x, y) =& \mathbb{E}\left[f(\beta' Z|Y \geq y)S_Y(y)\varepsilon(X, Z; x, y, \beta)\, \mathbb{1}(Y \geq y) \right. \\ & \left. \times f(\beta' \widetilde{Z}|Y \geq y)S_Y(y)\varepsilon(\widetilde{X}, \widetilde{Z}; x, y, \beta)\, \mathbb{1}(\widetilde{Y} \geq y)\mathcal{F}_\pi(Z - \widetilde{Z}) \right], \end{aligned} \tag{2.28}$$

where the expectation is over $(X, Y, Z, \widetilde{X}, , \widetilde{Y}, \widetilde{Z})$. So that the null hypothesis in (2.26) can now equivalently be written as

$$H_0 : \forall(x, y),\ T_\beta(x, y) = 0,\ \text{for some } \beta. \tag{2.29}$$

The test statistic below will be an appropriate function of an estimate of T_β obtained by replacing U by $\widehat{U}_n$. We need indeed a weighted version of $T_\beta(x,y)$ to integrate out the variables (x,y). So we can define

$$L_\beta = \int_{x,y} T_\beta(x,y)\rho(x,y)dx\,dy, \tag{2.30}$$

where $\rho(\cdot,\cdot)$ is some positive weight function. For instance we may choose $\rho(x,y) = f_{XY}(x,y)$, i.e. the pdf of (X,Y). We should reject the null if L_β is too large for all β (remember that by construction $L_\beta \geq 0$ with equality to 0 if and only if the null is true), so that in practice we will reject the null if $L_{\tilde{\beta}}$ is too large where $\tilde{\beta} = \arg\min_\beta L_\beta$.

Estimators of the components of $T_\beta(x,y)$ from a sample of observations $\{(X_i,Y_i,Z_i)\}_{i=1}^n$ can be obtained by plugging $\widehat{U}_n(x,y,t,\beta)$ at the place of $U(x,y,t,\beta)$ in the developments above. By using the multivariate standard normal as weight function $\pi(t)$, this leads to the following empirical analog of (2.28)

$$\begin{aligned}\widehat{T}_\beta(x,y) =& n^{-2}\sum_{i=1}^n\sum_{k=1}^n \Big[\widehat{f}(\beta' Z_i|Y\geq y)\widehat{S}_Y(y)\widehat{\varepsilon}(X_i,Z_i;x,y,\beta)\,\mathbb{1}(Y_i\geq y)\\ &\times\widehat{f}(\beta' Z_k|Y\geq y)\widehat{S}_Y(y)\widehat{\varepsilon}(X_k,Z_k;x,y,\beta)\,\mathbb{1}(Y_k\geq y)\mathcal{F}_\pi(Z_i-Z_k)\Big],\end{aligned} \tag{2.31}$$

which simplifies (see the derivations from (2.11) to (2.13)), with our choice of the weight function $\pi(\cdot)$, to

$$\begin{aligned}\widehat{T}_\beta(x,y) = &\frac{1}{(n^2h)^2}\sum_{i=1}^n\sum_{k=1}^n\left\{\sum_{j=1}^n K_{ij}^{\beta,h}\,\mathbb{1}(Y_j\geq y)\,\mathbb{1}(Y_i\geq y)\,[\mathbb{1}(X_i\leq x)-\mathbb{1}(X_j\leq x)]\right\}\\ &\times\left\{\sum_{j=1}^n K_{kj}^{\beta,h}\,\mathbb{1}(Y_j\geq y)\,\mathbb{1}(Y_k\geq y)\,[\mathbb{1}(X_k\leq x)-\mathbb{1}(X_j\leq x)]\right\}e^{-(Z_i-Z_k)'(Z_i-Z_k)/2}.\end{aligned} \tag{2.32}$$

Finally, to integrate out the values of (x,y) we choose the empirical density of the (X,Y)s and we have as practical test statistics

$$\widehat{L}_\beta = \frac{1}{n}\sum_{\ell=1}^n \widehat{T}_\beta(X_\ell,Y_\ell), \tag{2.33}$$

and we should reject H_0 if $\widehat{L}_\beta$ is too large for all β, so we will reject if $\min_\beta \widehat{L}_\beta$ is too large.

We will also see in the next section that β can be estimated by

$$\widehat{\beta} = \arg\min_\beta \widehat{L}_\beta. \tag{2.34}$$

So in practice we will reject H_0 if $\widehat{L}_{\widehat{\beta}}$ is too large, and we will use the bootstrap to approximate the p-value of the observed $\widehat{L}_{\widehat{\beta}}$.

2.4 Properties of the Estimation of β

We summarize below the main steps for analyzing the asymptotic properties of $\widehat{\beta}$. We will give a sketch of the proof of the following property.

Property 2.3. *Under the null Hypothesis* (2.5) *and if* $\sqrt{n}h^2 = o(1)$*, we have as* $n \to \infty$

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \mathcal{A}_\beta), \tag{2.35}$$

where $\mathcal{A}_\beta$ *is a* $d \times d$ *covariances matrix, whose structure is described in the proof.*

The main arguments are similar to those used in Delgado and Gonzales Manteiga (2001) and in Härdle and al. (1993). We first consider the space $\mathcal{L}^2_\omega$, where ω is a density on $\xi = (x, y, t)$ (so according to the notations above, $\omega(x, y, t) = \rho(x, y)\pi(t)$), defined as $\mathcal{L}^2_\omega = \{g(\xi) \mid \int g^2(\xi)\omega(\xi)\, d\xi < \infty\}$. We assume that for all β, $\widehat{U}_n(\xi, \beta) \in \mathcal{L}^2_\omega$ and we have shown above (see Property 2.2) that $\sqrt{n}\,\widehat{U}_n(\xi, \beta)$ has a limiting distribution $N(0, \sigma^2(\xi, \beta))$.

Following the same arguments we can prove that for any k, as $n \to \infty$ with $\sqrt{n}h^2 = o(1)$,

$$\sqrt{n} \begin{pmatrix} \widehat{U}_n(\xi_1, \beta) \\ \vdots \\ \widehat{U}_n(\xi_k, \beta) \end{pmatrix} \xrightarrow{\mathcal{L}} N(0, \Sigma^{k,k}_\beta), \tag{2.36}$$

where $\Sigma^{k,k}_\beta$ is a $k \times k$ matrix with elements $c_\beta(\xi_j, \xi_\ell)$ with $c_\beta(\xi_j, \xi_j) = \sigma^2(\xi_j, \beta)$. More generally, considering the operator Σ_β defined as

$$\text{For all } g \in \mathcal{L}^2_\omega,\ \Sigma_\beta g = \int g(\tilde{\xi}) c_\beta(\xi, \tilde{\xi}) \omega(\tilde{\xi})\, d\tilde{\xi}. \tag{2.37}$$

In fact, Σ_β is the asymptotic covariance operator of $\sqrt{n}\,\widehat{U}_n$.

We consider now an element $g \in \mathcal{L}^2_\omega$ and define the random variable

$$< g, \sqrt{n}\,\widehat{U}_n >= \sqrt{n} \int g(\xi)\widehat{U}_n(\xi, \beta)\omega(\xi)\, d\xi. \tag{2.38}$$

By using similar argument as in Section 2.2, we can prove that, as $n \to \infty$ with $\sqrt{n}h^2 = o(1)$,

$$< g, \sqrt{n}\,\widehat{U}_n > \xrightarrow{\mathcal{L}} N(0, < \Sigma_\beta g, g >). \tag{2.39}$$

More generally when considering a vector $G = (g_1, \ldots, g_k)'$ of k functions in $\mathcal{L}^2_\omega$, we have, as $n \to \infty$ with $\sqrt{n}h^2 = o(1)$,

$$< G, \sqrt{n}\,\widehat{U}_n > \xrightarrow{\mathcal{L}} N(0, < \Sigma_\beta G, G' >), \tag{2.40}$$

where $< \Sigma_\beta G, G' >$ denotes the $(k \times k)$ covariances matrix with elements $< \Sigma_\beta g_j, g_\ell >$.

The estimation of β is obtained in (2.34) by minimizing $\int \widehat{U}_n^2(\xi,\beta)\omega(\xi)\,d\xi$. Following Härdle et al. (1993), this minimization has to be analyzed in a neighborhood of the true value of β with radius $cn^{-1/2}$. We assume that there exist a unique minimum $\widehat{\beta}$ in this neighborhood with $||\widehat{\beta}-\beta|| \leq cn^{-1/2}$.

The first order condition can be written as

$$\int \frac{\partial}{\partial\beta}\widehat{U}_n(\xi,\widehat{\beta})\widehat{U}_n(\xi,\widehat{\beta})\omega(\xi)\,d\xi = 0. \tag{2.41}$$

Assuming that $\widehat{U}_n(\xi,\beta)$ is continuously differentiable in β up to the order two, we can write the Taylor expansion

$$\widehat{U}_n(\xi,\widehat{\beta}) = \widehat{U}_n(\xi,\beta) + \frac{\partial}{\partial\beta'}\widehat{U}_n(\xi,\beta)(\widehat{\beta}-\beta) + R_n(\xi,\beta), \tag{2.42}$$

where due to assumptions relatives to the choice of $\widehat{\beta}$ and to the differentiability of $\widehat{U}_n(\xi,\beta)$, we have $\sqrt{n}R_n(\xi,\beta) = o_p(1)$. Therefore, plugging this in (2.41), we obtain

$$\int \frac{\partial}{\partial\beta}\widehat{U}_n(\xi,\widehat{\beta})\left[\sqrt{n}\,\widehat{U}_n(\xi,\beta)\right]\omega(\xi)\,d\xi = -\left[\int \frac{\partial}{\partial\beta}\widehat{U}_n(\xi,\widehat{\beta})\frac{\partial}{\partial\beta'}\widehat{U}_n(\xi,\beta)\omega(\xi)\,d\xi\right]\sqrt{n}(\widehat{\beta}-\beta)+o_p(1). \tag{2.43}$$

By similar developments as in Section 2.2, we could prove that, as $n\to\infty$,

$$\frac{\partial}{\partial\beta}\widehat{U}_n(\xi,\widehat{\beta}) \xrightarrow{p} G(\beta,\xi), \tag{2.44}$$

where $G(\beta,\xi)$ is a d-vector with elements $g_j(\beta,\xi) = \frac{\partial}{\partial\beta_j}U(\xi,\beta) \in \mathcal{L}^2_\omega$ for $j=1,\ldots,d$. Hence the left part of equation (2.43) converges to a centered normal distribution with the $d\times d$ covariances matrix $<\Sigma_\beta G, G'>$.

In the right part of equation (2.43), we have

$$\left[\int \frac{\partial}{\partial\beta}\widehat{U}_n(\xi,\widehat{\beta})\frac{\partial}{\partial\beta'}\widehat{U}_n(\xi,\beta)\omega(\xi)\,d\xi\right] \xrightarrow{p} \mathcal{M}_\beta, \tag{2.45}$$

where $\mathcal{M}_\beta$ is a $d\times d$ matrix of rank $d-1$ due to the normalization constraint on β. So, at the end we obtain, as $n\to\infty$ with $\sqrt{n}h^2 = o(1)$,

$$\sqrt{n}(\widehat{\beta}-\beta) \xrightarrow{\mathcal{L}} N(0, \mathcal{M}_\beta^+ <\Sigma_\beta G, G'> \mathcal{M}_\beta^+), \tag{2.46}$$

where $\mathcal{M}_\beta^+$ is the generalized inverse of $\mathcal{M}_\beta$.

In practice, we will not use this result, but rather use bootstrap techniques to approximate the distribution of $\widehat{\beta}$.

2.5 The Bootstrap

The bootstrap has to mimic the way the sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is generated, but here the data (X, Y) take their values inside a conditional attainable set Ψ^z. In addition, we must generate bootstrap sample under the null hypothesis. One way to handle the bootstrap in a frontier setup has been investigated by many, see e.g. Simar and Wilson (1998, 2000) and Kneip et al. (2008, 2011). The procedure depends on the DGP, including, as described in Section 1.1, the chosen orientation to characterize distances from the efficiency boundary. Here we follow the input orientation but the procedure can easily be adapted to any orientation.

To generate, under the null hypothesis H_0, a bootstrap observation X_i^* conditional to the output level Y_i and the environmental conditions Z_i we have to generate X_i^* according $\widehat{F}(x|Y \geq Y_i, \widehat{\beta}'Z_i)$ by using (2.8). Since X may be multivariate we proceed as follows:

[1] We first generate a random ray in the X-space, conditionally on $Y \geq Y_i, \beta'Z = \widehat{\beta}'Z_i$, this may be obtained by drawing $\widetilde{X}_i$ a random observation in the restricted sample of observations X_j satisfying this condition, i.e. $Y_j \geq Y_i$ and $|\widehat{\beta}'Z_j - \widehat{\beta}'Z_i| \leq h_{\beta z}$[5] .

[2] Then we project this point $\widetilde{X}_i$ on the estimated frontier, conditionally on $Y \geq Y_i, \beta'Z = \widehat{\beta}'Z_i$ by computing its input efficiency under the null: $\widehat{\theta}(\widetilde{X}_i, Y_i|\widehat{\beta}'Z_i)$. So we have

$$\widetilde{X}_i^{\partial}(Y_i, Z_i) = \widehat{\theta}(\widetilde{X}_i, Y_i|\widehat{\beta}'Z_i)\, \widetilde{X}_i. \tag{2.47}$$

[3] We generate a bootstrap value of the conditional efficiency as follows:

[3.1] To simplify the notations, we denote $\widehat{\theta}_i = \widehat{\theta}(X_i, Y_i|\widehat{\beta}'Z_i)$ the n estimated conditional efficiencies computed under the null. From the sample $\left\{(\widehat{\theta}_j, Y_j, \widehat{\beta}'Z_j)\right\}_{j=1}^n$ we compute a smooth estimator of the conditional cumulative distribution $F_\theta(\theta|Y \geq Y_i, \beta'Z = \widehat{\beta}'Z_i)$ as follows

$$\widehat{F}_\theta(\theta \mid Y \geq Y_i, \widehat{\beta}'Z = \widehat{\beta}'Z_i) = \frac{\sum_{j=1}^n G((\theta - \widehat{\theta}_j)/h_\theta)\, \mathbb{1}(Y_j \geq Y_i) K(\widehat{\beta}'(Z_j - Z_i)/h_{\beta z})}{\sum_{j=1}^n \mathbb{1}(Y_j \geq Y_i) K(\widehat{\beta}'(Z_j - Z_i)/h_{\beta z})}, \tag{2.48}$$

where h_θ is an univariate bandwidth of order $n^{-1/5}$ and $G(\cdot)$ is an integrated kernel function defined as $G(t) = \int_{-\infty}^t g(s)ds$ where $g(\cdot)$ is a kernel density (see e.g. Li et al., 2013 for details).[6]

[5] The univariate bandwidth $h_{\beta z}$ may be obtained by least squares cross validation, as usually done in conditional frontier models, see e.g. details in Appendix A in Simar et al. (2016).

[6] In practice, we select h_θ by the usual rule of thumb, i.e. $h_\theta = 1.06 std(\widehat{\theta}) n^{-1/5}$, where $std(\widehat{\theta})$ is the standard deviation of the n estimates $\widehat{\theta}_j$.

[3.2] Draw a random value $\epsilon_i^* \sim \text{Unif}(0,1)$.

[3.3] Define $\theta_i^* = \theta^*(X_i, Y_i|\widehat{\beta}'Z_i)$ as the ϵ_i^*-quantile of $\widehat{F}_\theta(\theta \mid Y \geq Y_i, \widehat{\beta}'Z = \widehat{\beta}'Z_i)$ i.e.

$$\theta_i^* = \widehat{F}_\theta^{-1}(\epsilon_i^* \mid Y \geq Y_i, \widehat{\beta}'Z = \widehat{\beta}'Z_i). \tag{2.49}$$

[4] The i^{th} bootstrap data is thus given by (X_i^*, Y_i, Z_i), where $X_i^* = \widetilde{X}_i^{\partial}(Y_i, Z_i)/\theta_i^*$.

The bootstrap sample for $i = 1, \ldots, n$ leads to

$$\widehat{L}_\beta^* = \frac{1}{n}\sum_{\ell=1}^{n} \widehat{T}_\beta^*(X_\ell^*, Y_\ell). \tag{2.50}$$

and the bootstrap version of $\widehat{\beta}$ is given by $\widehat{\beta}^* = \arg\min_\beta \widehat{L}_\beta^*$, so that the bootstrap value of the test statistic is $\widehat{L}_{\widehat{\beta}^*}^*$. Here we use the same bandwidth h obtained in the original sample.

Repeating the bootstrap a large number B of time will produce a set of values $\widehat{L}_{\widehat{\beta}^{*,b}}^{*,b}$ for $b = 1, \ldots, B$. The p-value is then approximated by

$$\widehat{p}-\text{value} = \frac{\sum_{b=1}^{B} \mathbb{1}\left[\widehat{L}_{\widehat{\beta}^{*,b}}^{*,b} \geq \widehat{L}_{\widehat{\beta}}\right]}{B}. \tag{2.51}$$

We will see in our Monte-Carlo experiments below that this testing procedure works reasonably well even with moderate sample sizes, with as expected, better behavior as n increases. We will also see that the sampling distribution of the estimated β's when the SIM is true behaves also very nicely and as expected by the theory.

3 Numerical examples

3.1 Simulated data and Monte-Carlo experiments

Simulating data that satisfy some hypothesis on the conditional distribution $F_{X|YZ}(x \mid Y \geq y, Z = z)$ (like the SIM assumption) is not easy. One way to achieve this is described in Appendix A. Here, as explained in Remark A.3 of the same appendix, we simplify the procedure.

We consider a cost function with one input X (the cost) and one output Y with three environmental variables $Z \in \mathbb{R}^3$. So we have $p = q = 1$ and $d = 3$. For the SIM, we select the three coefficients $(1, -0.5, 2)'$ which after normalization provide the value $\beta = (0.436436, -0.218218, 0.872872)'$ which will be the true value of β when the null hypothesis is true.

The effect of the Z variables on the production process will be defined through a weighted combination of the SIM model (under the null hypothesis) and a nonlinear part. We define

$$g(z) = w_0 \times (\beta' z) + (1 - w_0) \times \left(2 \sin z_1 + (z_2 - 2)^2\right) \tag{3.1}$$

where $w_0 \in [0, 1]$ tunes departures from the null hypothesis which corresponds to the case $w_0 = 1$. When w_0 decreases, the nonlinear part of $g(z)$ becomes more important. We will consider in our Monte-Carlo experiments the 5 cases $w_0 = 1, 0.75, 0.50, 0.25$ and 0. For the latter case, the linear part of the SIM has fully disappeared. However, for instance, when $w_0 = 0.75$ we have a small departure of the null since the nonlinear part of $g(z)$ has only a weight of 0.25. The three components of Z are simulated as independent uniform: $Z^j \sim \text{Unif}(0, 4)$, $j = 1, \ldots, 3$.

The conditional cost function (see (1.12)) is defined as

$$\varphi(y, z) = (1 + y^2) \exp\left[-(g(z) - 2)/3)\right], \tag{3.2}$$

where the values of Y will be simulated as $Y \sim \text{Unif}(0, 3)$.

So we may obtain a sample of size n of frontier points $\varphi(Y_i, Z_i))$, $i = 1, \ldots, n$. Finally we simulate the observations of the costs X_i by introducing inefficiencies:

$$X_i = \varphi(Y_i, Z_i)) \exp(V_i), \tag{3.3}$$

where the distribution of V is also dependent of $g(Z)$, we define for $i = 1, \ldots, n$

$$U_i | Z_i \sim N^+(0, \sigma_V^2(Z_i)), \text{ with } \sigma_V(Z_i)) = 0.2 + g(Z_i)/10, \tag{3.4}$$

where $N^+(0, \sigma^2)$ denotes the half-normal distribution, i.e. the positive part of a $N(0, \sigma^2)$.

So we see that both the frontier level and the distribution of the inefficiencies are affected by the function $g(z)$, but the null hypothesis (the SIM model) is only true when $w_0 = 1$. In our experiments we will choose the 4 values $n = 100, 200, 400$ and 800 and we perform $MC = 500$ Monte-Carlo repetitions in each case, i.e. for each of the 20 selected pair (n, w_0). For the Monte-Carlo evaluation of the bootstrap we use the warp-speed method, introduced by Giacomini et al. (2013). The idea is to perform from each Monte-Carlo sample only one bootstrap drawn, in the way described above. Then the MC values, of this single bootstrap sample, approximate the desired bootstrap distribution.

3.1.1 Testing the SIM

The test statistics in (2.33) depends on an univariate bandwidth h. By Property 2.2 , we need $\sqrt{n}h^2 = o(1)$ as $n \to \infty$, so we select $h = Cn^{-1/3}s_{\beta' Z}$ for some constant $C > 0$ and where $s_{\beta' Z}$ is the sample standard deviation of the values $\beta' Z$ computed with the estimated value of β.[7] All the results below are obtained with $C = 1$, but a few pilot Monte-Carlo experiments have shown a great robustness with respect to C. We obtained qualitatively similar results with $C = 0.5$ and 2.

The results of our Monte-Carlo experiments are summarized in Table 1. We observe that the size of the test (when the null is true: $w_0 = 1$) seems rather good even with few observations ($n = 100$) and from $n = 200$ the size of the test is almost correct for the 3 suggested size α. As expected is this pure nonparametric approach, the power is not very good for small sample sizes but note that in the case $w_0 = 0.75$, the model is not far from the SIM (only 25% weight for the nonlinear part), so it is hard to reject the null. However the power increases quickly when n increases and of course, as expected, when w_0 decreases. With $n = 800$ the rejection rates are almost perfect.

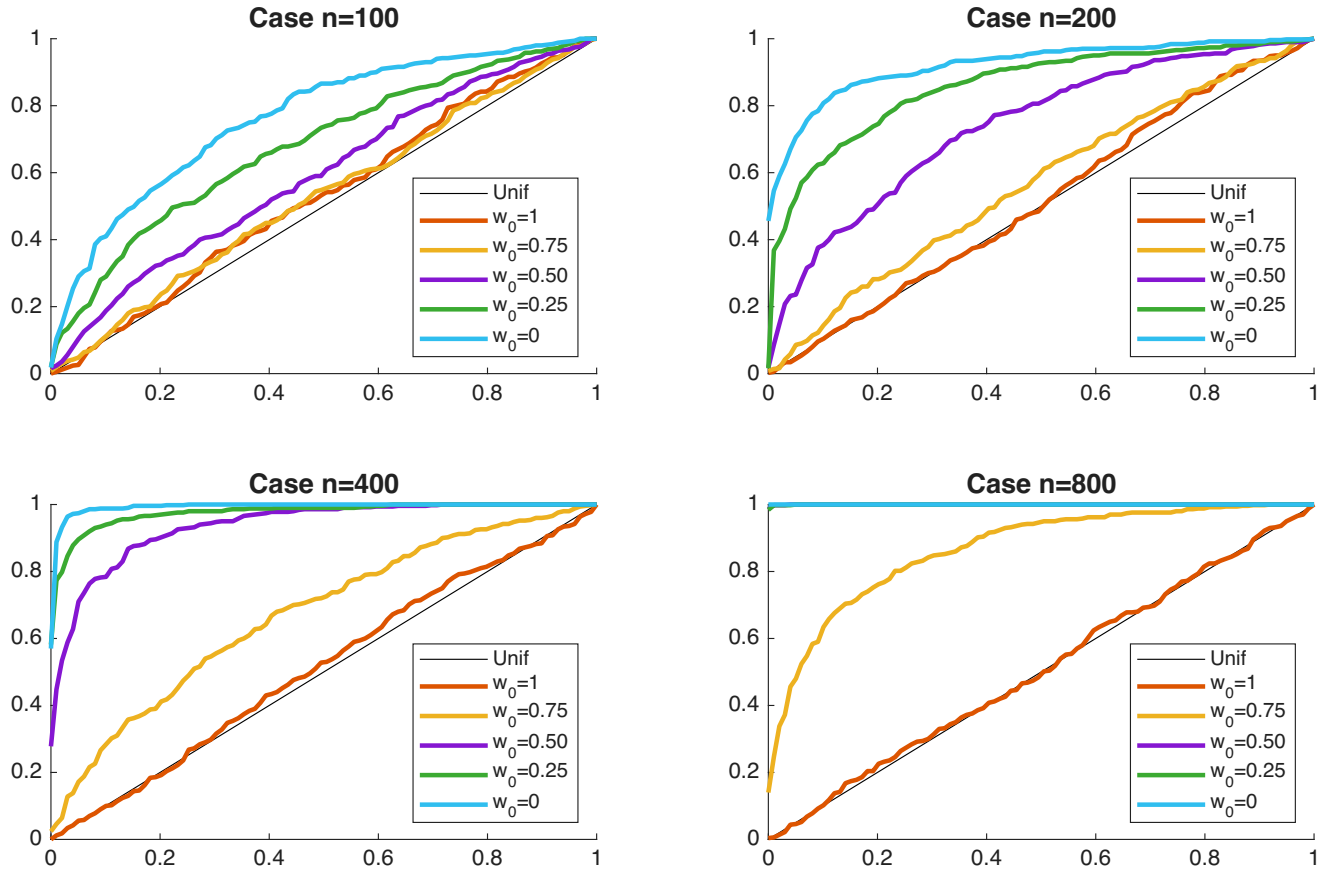
All these nice properties of our testing procedure appear clearly in Figure 1 where we display the accuracy plots, i.e. the Monte-Carlo distributions of the achieved p-values. Under the null, this distribution should be the uniform. The picture present the various sample sizes and the various distances from the null (increasing when w_0 decreases). We see that even with $n = 100$, the distribution under the null is not so far from the uniform, but we confirm the rather low power. Things improve quickly when n increases, in particular from $n \geq 400$.

[7] In order to compute $\widehat{\beta}$ by solving (2.34), we first use a rough mesure of $s_{\beta' Z}$ by choosing the square root of the largest eigenvalue of S_Z, the sample covariance matrix of the observations Z_i. Once this $\widehat{\beta}$ is obtained, we correct the bandwidth by using rather the standard deviation of the $\widehat{\beta}' Z$. Then we recompute the corresponding value of the test statistics $\widehat{L}_{\widehat{\beta}}$ as explained in the text below equation (2.34).

Table 1: Percentages of rejection on 500 Monte-Carlo experiments for tests of level α. The null is true when $w_0 = 1$ and the null is false whenever $w_0 < 1$ (w_0 is the weight of the linear component in the function of Z acting on the conditional $F_{X|Y,Z}(x|Y \geq y, Z)$). When $w_0 = 0$, there is no more linear component.

$n = 100$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$w_0 = 1$	0.0060	0.0260	0.1100
$w_0 = 0.75$	0.0200	0.0480	0.1120
$w_0 = 0.50$	0.0240	0.1040	0.1880
$w_0 = 0.25$	0.0880	0.1800	0.2900
$w_0 = 0.00$	0.1040	0.2900	0.4100
$n = 200$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$w_0 = 1$	0.0060	0.0440	0.1040
$w_0 = 0.75$	0.0140	0.0860	0.1440
$w_0 = 0.50$	0.0860	0.2360	0.3840
$w_0 = 0.25$	0.3680	0.5240	0.6280
$w_0 = 0.00$	0.5460	0.7060	0.8080
$n = 400$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$w_0 = 1$	0.0120	0.0560	0.1000
$w_0 = 0.75$	0.0460	0.1720	0.2840
$w_0 = 0.50$	0.4480	0.7100	0.7840
$w_0 = 0.25$	0.7740	0.8960	0.9400
$w_0 = 0.00$	0.8880	0.9740	0.9880
$n = 800$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$w_0 = 1$	0.0060	0.0460	0.1020
$w_0 = 0.75$	0.2460	0.4800	0.6340
$w_0 = 0.50$	0.9980	1.0000	1.0000
$w_0 = 0.25$	0.9980	1.0000	1.0000
$w_0 = 0.00$	1.0000	1.0000	1.0000

Figure 1: *Accuracy plots: Monte-Carlo distributions of the p-values, with* $MC = 500$. *If the null is true, it should be like the uniform cdf (the black line). The null is true when* $w_0 = 1$ *(red curves) and the null is false when* $w_0 < 1$ *(*w_0 *is the weight of the linear component in the function of* Z *acting on the conditional* $F_{X|Y,Z}(x|Y \geq y, Z)$*). When* $w_0 = 0$*, there is no more linear component.*

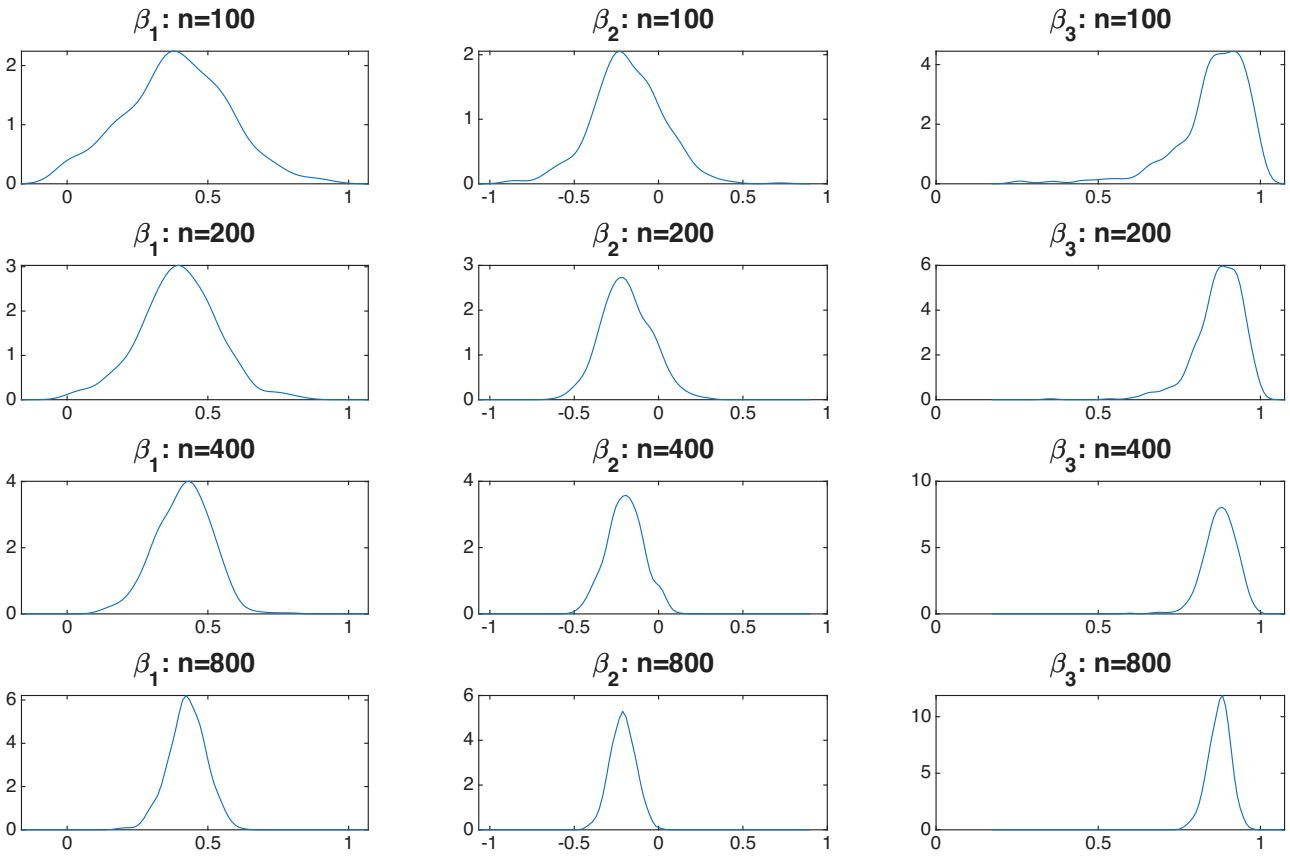


3.1.2 The estimation of the β when the null is true

In this case also we observe through our Monte-Carlo experiments that if the null hypothesis is true (the SIM is satisfied) a rather good behavior of the resulting estimators of β given in (2.34). Figure 2 dispalys the Monte-Carlo distributions of the 3 components of $\widehat{\beta}$.

Even with only $n = 100$ observations, the estimators are not far from being unbiased. Of course, even the bias disappears, as expected with n increasing. We also observe that these estimators seem to be consistent (as expected by the theory), since the Monte-Carlo variance reduces when n increases. Finally, the Monte-Carlo distributions are typically bell shaped confirming our theoretical results of asymptotic normality (see Property 2.3).

Figure 2: *Monte-Carlo distribution of the* $MC = 500$ *estimates of* β *obtained under the null. In each case, the true values of* $\beta_1, \beta_2, \beta_3$ *are* $0.436436, -0.218218$, *and* 0.872872 *respectively.*



3.2 Illustration with real data: the case of the French Post

The data are provided by "La Poste", the French national postal operator in charge of universal service, on delivery offices. They are related to the annual activity observed in 2013 in premises (the "delivery offices") where postal items are sorted one last time in the direction of the delivery round and from where postmen leave (and return) to make their last-mile delivery round. We used for the illustration a random sample of $n = 500$ such units. The variables are:

- X is the number of worked hours at this unit.
- Y is the total number of items delivered by this unit.
- Three variables are considered as environmental variables:
  - Z_1 is the quantity of items to be delivered per distribution point (i.e. mailbox);
  - Z_2 is the inverse of a measure of the geographical complexity of the delivery round of the postman. Z_2 is defined as the ratio of the area covered (in hm^2) by the postmen of the units, divided by the average length of the delivery route (in hm);
  - Z_3 is a measure of the density of the delivery round defined as the number of delivery addresses divided by the length of the delivery route.
  - The scale of the three Z being quite different (and measured in different units), all the three variables are scaled by their standard deviations.

Table 2 below gives some descriptive information on these data.

Table 2: Summary statistics on the data, $n = 500$.

variable	min	Q_{25}	Q_{50}	Q_{75}	max
X	115.7600	764.8200	1207.0450	2130.9500	6872.2700
Y	11426.7000	91147.6280	161443.4050	301336.2000	972856.9900
scaled Z_1	1.2192	3.8470	4.2841	4.8447	9.4461
scaled Z_2	0.0473	0.3335	0.5934	1.1093	9.4857
scaled Z_3	0.0898	1.0939	1.6238	2.3101	5.5677

3.2.1 Test of the SIM model.

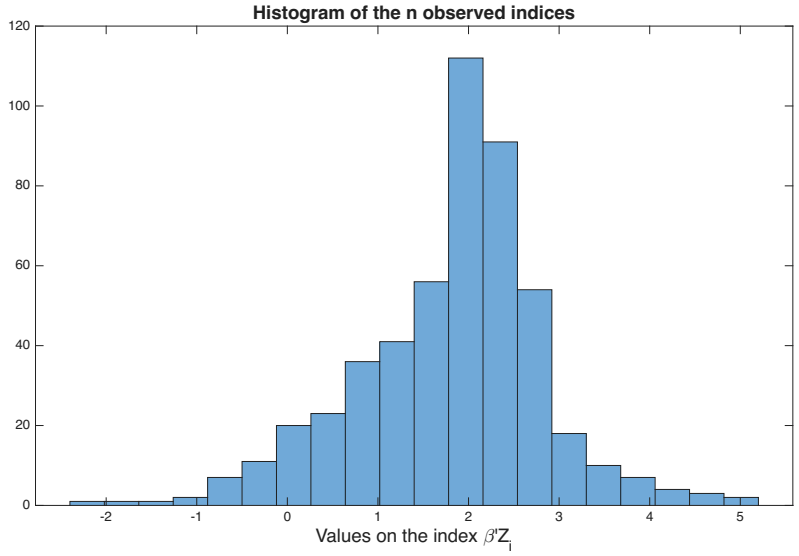
The bootstrap test was performed with $B = 1000$ bootstrap replications and we obtain a $p-$value $= 0.9960$, so the SIM model appears quite reasonable here. The estimation of β is $\widehat{\beta} = (0.57753, 0.50331, -0.64276)'$. We will come back below on the interpretation of these

β's. Table 3 and Figure 3 give an idea of the distribution of the resulting values of $\widehat{\beta}'Z_i$ in the sample. We will comment below the effect of these indices when analyzing the efficiencies of the units.

Table 3: *Summary statistics for the n estimated index* $\widehat{\beta}'Z_i$.

min	Q_{25}	Q_{50}	Q_{75}	max
-2.2347	1.2353	1.9552	2.4315	5.1822

Figure 3: *Distribution of the n estimated values of the index* $\beta'Z_i$.



3.2.2 Efficiency analysis.

Table 4 displays the data and the estimated efficiencies of 20 randomly chosen units in our sample. We provide the estimation of the conditional efficiencies $\widehat{\theta}(X_iY_i|\widehat{\beta}'Z_i)$ and their robust version $\widehat{\theta}_m(X_iY_i|\widehat{\beta}'Z_i)$. The value of m was selected as usual in this literature by selecting the value showing an "elbow effect" on the the percentage of points below the cost frontier as a function of m. Figure 4 displays the picture and the value $m = 150$ let approximatively 9% of observations below the order-m cost frontier (these points are considered as "extreme" or "outlying" data points). For the order-m efficiency measures, we can derive confidence intervals obtained by the 1000 bootstrap replications.

We can also provide a 3D picture of the estimated frontiers as a function of y and $\widehat{\beta}'z$, for the full frontier and for the order-m frontier. We see a slight favorable effect of $\beta'Z$ on

the frontier levels (for a fixed Y). Small values of $\widehat{\beta}'z$ seems to produce higher cost frontier. We see also very similar shape for full cost frontier and order-m, with $m = 150$.

Figure 4: *Choosing the value of* m*: percentages of data points below the order-*m *cost frontier as a function of* m*.*

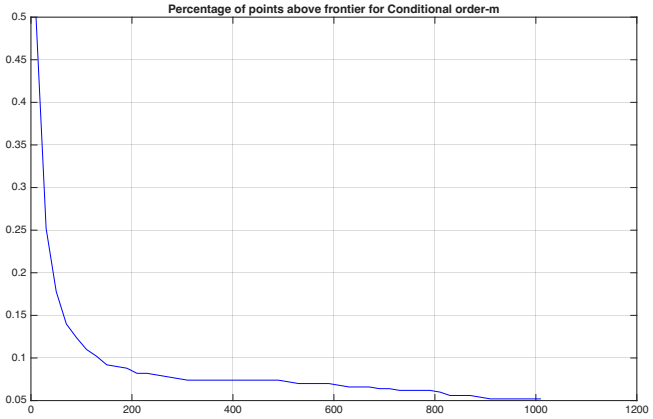
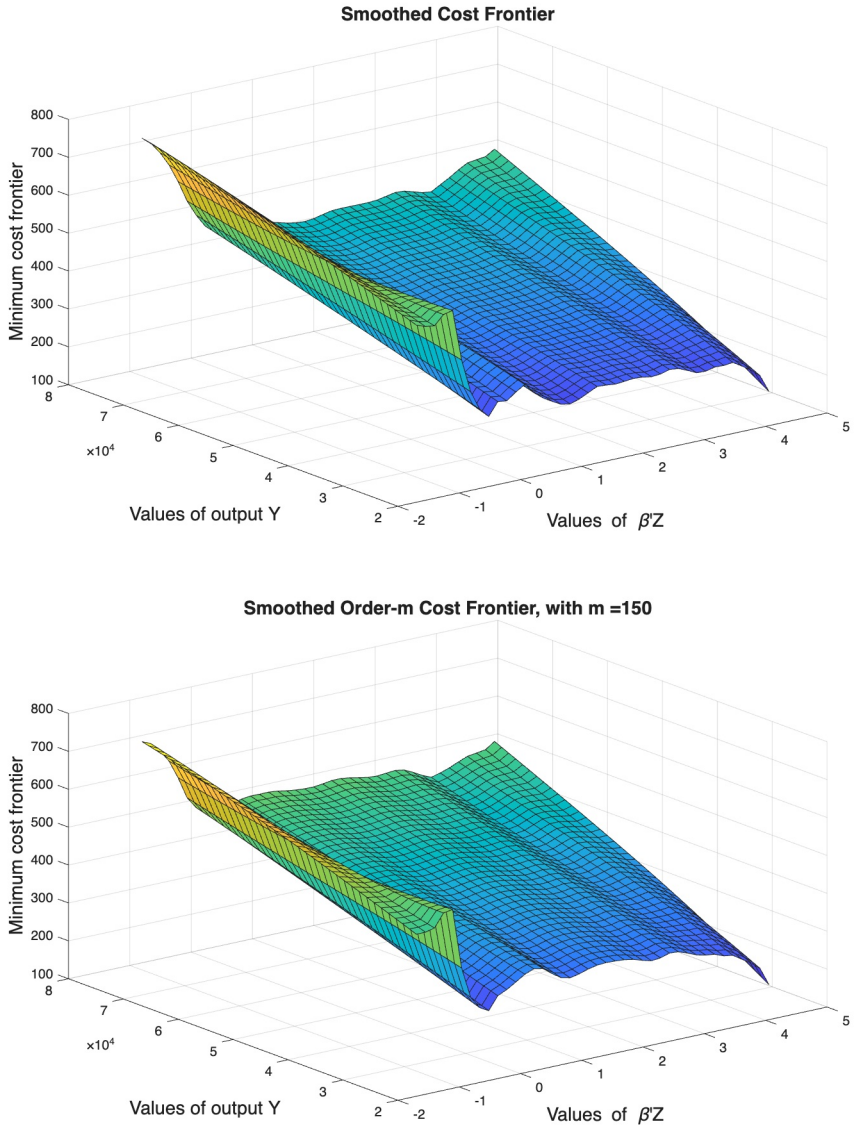


Table 4: *Data and Efficiency estimates for 20 randomly chosen units. The column "full" is for* $\widehat{\theta}(X_i, Y_i|\widehat{\beta}'Z_i)$ *and the column "order-m" is for* $\widehat{\theta}_m(X_i, Y_i|\widehat{\beta}'Z_i)$ *with* $m = 150$*. The columns "low" and "upper" are the limits of the 95% confidence intervals for the order-m inefficiencies, obtained through* $B = 1000$ *bootstrap replications.*

Unit i	X_i	Y_i	Z_{1i}	Z_{2i}	Z_{2i}	$\beta'Z$	full	order-m	low	upper
415	1116.05	154420.23	3.5058	0.4644	2.1601	0.8700	0.7243	0.7901	0.7500	0.8864
113	1206.04	200246.62	3.4613	0.1440	2.7262	0.3192	0.8836	0.8899	0.8496	0.9661
280	1321.79	300031.18	3.9415	0.0637	4.5364	-0.6074	1.0000	1.0001	0.7406	1.0604
410	1791.08	230081.58	4.2121	0.9534	1.5817	1.8958	0.6088	0.6660	0.6626	0.7588
231	774.94	86367.43	3.6299	0.2787	1.1301	1.5103	0.6180	0.7132	0.6873	0.8125
322	2060.57	375106.79	5.0495	0.3293	1.6022	2.0521	0.8434	0.8701	0.8190	0.9599
249	5157.56	715681.39	3.7152	0.0917	2.8426	0.3647	0.9171	0.9171	0.8763	1.0021
435	4591.56	673490.32	3.7980	0.2353	1.6486	1.2522	0.9085	0.9086	0.8078	0.9852
168	802.37	134363.48	5.3267	0.3303	2.8886	1.3859	0.9266	0.9659	0.8964	1.0243
94	374.99	34194.42	4.5266	7.0760	5.3981	2.7060	0.8626	0.8920	0.7655	0.9100
394	5146.82	745180.56	4.2418	0.3045	1.6225	1.5602	0.9190	0.9190	0.8629	1.0020
337	1065.63	224429.12	4.0717	0.0508	3.5791	0.0766	1.0000	1.0038	0.8907	1.0737
261	500.37	48991.78	2.9955	1.4102	0.8065	1.9214	0.7180	0.8488	0.8414	0.9819
457	1049.23	133224.16	4.4342	1.3514	1.8340	2.0622	0.6134	0.6800	0.5856	0.7255
65	1322.09	145777.58	4.0197	1.3117	1.6420	1.9263	0.6385	0.6782	0.6953	0.7675
340	1228.04	182796.67	6.8016	0.7269	1.6255	3.2492	0.6874	0.7093	0.4839	0.7476
41	534.34	54542.17	5.5860	1.6189	0.8667	3.4838	0.6723	0.7195	0.6243	0.7906
162	641.10	90101.48	4.2500	0.8917	3.3475	0.7516	0.7471	0.8886	0.8062	1.0114
154	2938.56	473123.30	6.3697	0.1518	3.7306	1.3573	0.6693	0.6724	0.6603	0.7457
360	1159.41	192059.63	5.5471	1.3424	1.8684	2.6784	1.0000	1.0077	1.0104	1.1349

Figure 5: *Cost frontiers: local-linear smoothed values of the cost frontiers: the full frontier (upper panel) and the order-m frontier (lower panel).*



Since we observe the favorable effect of $\beta' Z$ on the cost frontier, we can now come back to the estimated values of β, given by $\widehat{\beta} = (0.57753, 0.50331, -0.64276)'$ and interpret the effect of individual Z's. We see that, as expected, Z_1 and Z_2 have a positive effect on the cost frontier of the same order (in scaled units) and that Z_3 has a negative effect, of the same order (in scaled units). For instance if we focus on Z_3, we can say that, everything else being kept constant, the density of the delivery rounds managed by the delivery office makes the job more complicated, i.e. at more costly. Similar comments, but in the opposite direction could be said for Z_1 and Z_2.

4 Conclusions

In production theory, it is important for practitioners to analyze the possible effect of environmental variables on the efficiency of firms. The usual and natural way to investigate this issue is to consider conditional frontier models. As described in the literature, for nonparametric approaches, this can create serious problems if the number of these potential environmental factors increases, because the rates of convergence are deteriorated when the number of such variables increases.

In this paper, to address this issue, we investigate whether Single Index Models (SIM) could be used for modeling the effect of these variables on the production process. The SIM impose some structure on the conditional distribution $F_{X|YZ}(x \mid Y \geq y, Z = z)$, where the condition on Z is replaced by a condition on a linear combination $\beta' Z$, for some β. In terms of gain in dimension we condition only to one variable $\beta' Z$ in place of a multivariate Z. We show this is equivalent to an assumption of conditional independence, but with an inequality on Y in the condition and not an equality. This implies to develop an original methodology, adapting the Delgado et al. (2001) to this particular setup. We thus propose a test for the SIM in conditional frontiers and we analyze its asymptotic properties. In practice the bootstrap is proposed to approximate the p-value. If the SIM model is not rejected, we obtain better rates of convergence of the conditional efficiency estimates.

The paper investigates, through some Monte-Carlo experiments, the finite sample properties of the proposed test and the properties of the resulting estimates of the SIM when it is not rejected and the results confirm the test is working well in practice, with a size not far from the desired level and power when deviating from the null hypothesis.

We illustrate the method with a real data set from the French national postal operator, where the SIM is not rejected. We show how easy we can interpret the resulting coefficients β of the SIM and hence the individual effect of each variable Z on the cost frontiers.

Conflict of interest : None

A How to simulate a random sample from a given $F_{X|YZ}(x \mid Y \geq y, Z = z)$?

The question is how to simulate a random sample of (X_i^*, Y_i, Z_i) for $i = 1, \ldots, n$ from $\text{Prob}(X_i^* \leq x \mid Y \geq Y_i, Z = Z_i) = F_{X|YZ}(x \mid Y \geq Y_i, Z = Z_i)$. This is needed if we want to simulate data that satisfy some restriction (hypothesis) on the function $F_{X|YZ}(x \mid Y \geq y, Z = z)$ (e.g. a SIM for the dependence on Z).

From the econometric literature on frontier we know how to simulate a random sample (X_i, Y_i, Z_i) for $i = 1, \ldots, n$ from $F_{X|YZ}(x \mid Y = y, Z = z)$, with an equality when conditioning on $Y = y$. But we know that a property on the latter does not imply automatically the same property for the former (with the condition $Y \geq y$).

We have seen in (1.11) that

$$\theta(x, y|z) = \inf\{\theta \mid F_{X|YZ}(\theta x|Y \geq y, Z = z) > 0\} \in (0, 1],$$

and the coordinates of the radial projection of a point (x, y) on the efficient boundary of Ψ^z are given by $(x^\partial(y, z), y)$, where $x^\partial(y, z) = \theta(x, y|z)x$. This efficient boundary can be viewed as a surface defined in Ψ^z for each value of (x, y), given $Z = z$.

It is rather easy and standard in the literature on frontier models, to simulate a sample from a cdf on X, conditioned by equalities on Y and Z. We select a p-valued mathematical expression for $x^\partial(y, z)$ as a non-decreasing function of y and some given function of z. We generate a random sample of pairs (Y_i, Z_i), according to some selected scenario. Then we compute the random points $x^\partial(Y_i, Z_i)$ in the X-space. Now given $Y = Y_i, Z = Z_i$ we can simulate points X_i above the efficient surface as follows. Select a scenario for generating for $i = 1, \ldots, n$, $V_i|Y_i, Z_i \sim D_+(\eta(Y_i, Z_i))$, where $D_+(\cdot)$ is some defined scalar valued positive random variable (like exponential or half-normal, etc.). This provides a random sample $\mathcal{V}_n = \{(V_i, Y_i, Z_i)\}_{i=1}^n$. Then we define $X_i = x^\partial(Y_i, Z_i) * \exp(V_i)$. This provides a random sample $\mathcal{X}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ generated according to the $F_{X|YZ}(x \mid Y = y, Z = z)$, specified by the various components previously defined.

Now to simulate a sample according $F_{X|YZ}(x \mid Y \geq y, Z = z)$, we first draw a random ray $\widetilde{X}_i$ by drawing at random, in $\mathcal{X}_n$, an observation X_j among the restricted indices j satisfying $Y_j \geq Y_i$ and $|Z_j - Z_i| \leq h_{xz}$. Then we draw a random variable $\widetilde{V}_i$, by drawing at random, in $\mathcal{V}_n$, a value V_j among the restricted indices j satisfying $Y_j \geq Y_i$ and $|Z_j - Z_i| \leq h_{vz}$.[8] Finally we define $\theta_i^* = \exp(-\widetilde{V}_i) \in (0, 1]$. Hence, for $i = 1, \ldots, n$, the values of $X_i^* = x^\partial(Y_i, Z_i)/\theta_i^*$ provide the desired sample of values (X_i^*, Y_i, Z_i).

[8]The two bandwidths h_{xz} and h_{vz} require some smoothing for the Z variables. they can be determined, e.g., by least squares cross validation when estimating $F_{X|YZ}(x \mid Y \geq Y_i, Z = Z_i)$ with the sample $\mathcal{X}_n$ and, respectively, estimating $F_{V|YZ}(v \mid Y \geq Y_i, Z = Z_i)$ from the random sample $\mathcal{V}_n$.

Remark A.1. *Since $\widetilde{V}_i$ is univariate, a more sophisticated way to simulate it would be as follows: (i) generate a random value $\epsilon_i \sim \text{Unif}(0,1)$, (ii) define $\widetilde{V}_i$ as the ϵ_i-quantile of $\widehat{F}_{V|YZ}(v \mid Y \geq Y_i, Z = Z_i)$, i.e. $\widetilde{V}_i = \widehat{F}^{-1}_{V|YZ}(\epsilon_i \mid Y \geq Y_i, Z = Z_i)$.*

Remark A.2. *If we want to impose the hypothesis that , for all x, y, $F_{X|YZ}(x \mid Y \geq y, Z = z) = F_{X|YZ}(x \mid Y \geq y, g(Z) = g(z))$, for some given function $g : \mathbb{R}^d \mapsto \mathbb{R}$, all the procedure explained in this appendix can be followed by replacing Z by $g(Z)$ (and z by $g(z)$) everywhere. Note that in this case, the bandwidths h_{xz} and h_{vz} are univariate. The SIM hypothesis, analyzed in this paper, is the particular case where $g(Z) = \beta' Z$ for some β.*

Remark A.3. *In the DGP described in Section 3.1 the cost X is univariate ($p = 1$) which avoids to generate the random ray $\widetilde{X}$ described above (all the X's have he same ray), the g function is given in (3.1) and $x^\partial(y, z)$ is given by the real function $\varphi(y, z)$ in (3.2) and we choose $V \perp\!\!\!\perp Y | g(Z)$, which avoids to simulate the $\widetilde{V}$. Note also that $Y \perp\!\!\!\perp Z$. All these assumptions simplify the way to simulate our DGP.*

References

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

[2] 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.

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

[4] 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.

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

[6] Delgado. M. and W. Gonzales Manteiga (2001), Significance Testing in Nonparametric Regression Based on the Bootstrap, *The Annals of Statistics*, 29 (5), 1469–1507.

[7] 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.

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

[9] Giacomini, R., Politis, D.N. and H. White (2013), A Warp-Speed Method for Conducting Monte-Carlo Experiments involving Bootstrap Estimators, *Econometric Theory*, 29, (3), 567–589.

[10] Härdle, W.K., P. Hall and H. Ichimura (1993), Optimal Smoothing In Single-Index Models, *The Annals of Statistics*, 21, 1, 157–176.

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

[12] Jeong, S.O. and L. Simar (2006), Linearly interpolated FDH efficiency score for nonconvex frontiers, *Journal of Multivariate Analysis*, 97, 2141–2161.

[13] Kneip, A, L. Simar and P.W. Wilson (2008), Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models, *Econometric Theory*, 24, 1663–1697.

[14] Kneip, A., Simar, L. and P.W. Wilson (2011), A Computational Efficient, Consistent Bootstrap for Inference with Non-parametric DEA Estimators. *Computational Economics*. , 38,483–515.

[15] 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.

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

[17] Simar, L. and A. Vanhems (2012), Probabilist Characterization of Directional Distances and their Robust versions. *Journal of Econometrics*, 166, 342–354.

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

[19] Simar, L. and P. Wilson (1998), Sensitivity of efficiency scores : How to bootstrap in Nonparametric frontier models, *Management Sciences*, 44, 1, 49–61.

[20] Simar L. and P. Wilson (2000), A General Methodology for Bootstrapping in Nonparametric Frontier Models, *Journal of Applied Statistics*, Vol 27, 6, 779–802.

[21] 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.

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

[23] van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, New York.

[24] Wilson, P. W. (2011), Asymptotic properties of some non-parametric hyperbolic efficiency estimators. In *Exploring research frontier in contemporary statistics and econometrics*, eds. I. Van Keilegom and P. W. Wilson, pp 115–150. Berlin: Springer-Verlag.

[25] Wilson, P. W. (2018), Dimension reduction in nonparametric models of production, *European Journal of Operational Research*, 267, 349–367.