PROPORTIONAL INCREMENTAL COST PROBABILITY FUNCTIONS AND THEIR FRONTIERS

Frédérique Fève, Jean-Pierre Florens, Léopold Simar

LIDAM Discussion Paper ISBA
2022 / 16

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.html

Proportional Incremental Cost Probability Functions and their Frontiers

Frédérique FÈVE* Jean-Pierre FLORENS*
Léopold SIMAR*,§

May, 2022

Abstract

The econometric analysis of cost functions is based on the analysis of the conditional distribution of the cost Y given the level of the outputs $X \in \mathbb{R}^p_+$ and given a set of environment variables $Z \in \mathbb{R}^d$. The model basically describes the conditional distribution of Y given $X \geq x$ and $Z = z$. In many applications, the dimension of Z is naturally large and a fully nonparametric specification of the model is limited by the curse of the dimensionality. Most of the approaches so far are based on two-stage estimations when the frontier level does not depend on the value of Z. But even in the case of separability of the frontier, the estimation procedure suffers from several problems, mainly due to the inherent bias of the estimated efficiency scores and the poor rates of convergence of the frontier estimates. In this paper we suggest an alternative semi-parametric model which avoids the drawbacks of the two-stage methods. It is based on a class of model called the Proportional Incremental Cost Functions (PICF), adapted to our setup from the Cox proportional hazard models extensively used in survival analysis for durations models. We define the PICF model, then we examine its properties and propose a semi-parametric estimation. By this way of modeling, we avoid the first stage nonparametric estimation of the frontier and avoid the curse of dimensionality keeping the parametric $\sqrt{n}$ rates of convergence for the parameters of interest. We are also able to derive $\sqrt{n}$-consistent estimator of the conditional order-m robust frontiers (which, by contrast to the full frontier, may depend on Z) and we prove the Gaussian asymptotic properties of the resulting estimators. We illustrate the flexibility and the power of the procedure by some simulated examples and also with some real data sets.

Key Words: Cost efficiency, Nonparametric robust frontier, Proportional hazard model, Environmental variables.

JEL Classification: C10, C14, C51, D22.

*Toulouse School of Economics (TSE), Toulouse, France, frederique.feve@tse-fr.eu, and jean-pierre.florens@tse-fr.eu. F. Fève and J.P. Florens acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissement dAvenir), grant ANR-17-EURE-0010.

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

1 Introduction and the Setup

The econometric analysis of cost functions is based on the analysis of the conditional distribution of the cost Y given the level of the outputs $X \in \mathbb{R}_+^p$ and given a set of environment variables $Z \in \mathbb{R}^d$. The model basically describes the conditional distribution of Y given $X \geq x$ and $Z = z$ e.g. through the survivor function

$$S(y|x,z) = \text{Prob}(Y \geq y \mid X \geq x, Z = z), \tag{1.1}$$

and the cost frontier is the lower bound of the support of this survivor (see Cazals et al., 2002):

$$\varphi(x,z) = \sup\{y \mid S(y|x,z) = 1\}, \tag{1.2}$$

the latter is often referenced as a "conditional to z" frontier, since it gives for a level of inputs x the minimal achievable cost for a firm facing environmental conditions z. We see that the basic object of interest is the conditional survivor function $S(y|x,z)$, where we note the peculiar form of the conditioning for X which guarantees the monotonicity of $\varphi(x,z)$ with respect of x. This specification does not cover the case where some X or Z are "endogenous".[1]

Nonparametric estimators have been proposed, based on envelopment estimators (like DEA-FDH type estimators), but they suffer from the curse of the dimensionality. See e.g. Cazals et al. (2002), Daraio and Simar (2005), Jeong et al. (2010) and Daraio et al. (2018). The basic idea is to plug in (1.2) a nonparametric estimator of $S(y|x,z)$, e.g. the kernel estimator

$$\widehat{S}_n(y|x,z) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i \geq y, X_i \geq x) K_h(Z_i - z)}{\sum_{i=1}^n \mathbb{1}(X_i \geq x) K_h(Z_i - z)}, \tag{1.3}$$

where $K_h(u) = h^{-d} \prod_{j=1}^d K(u_j/h)$ and $K(\cdot)$ is a univariate kernel with compact support. We simplify here the notation by choosing $h_j = h$, $j = 1, \ldots, d$. The resulting estimator of the conditional frontier is then the conditional FDH estimator

$$\widehat{\varphi}_n(x,z) = \min\{Y_i \mid X_i \geq x, |Z_i - z| \leq h\}, \tag{1.4}$$

whereas by convexifying the local condition set we obtain the conditional DEA estimator (for details, see e.g. Jeong et al., 2010).

To fix the ideas, it is shown in this literature that the conditional frontier estimator achieves the rates of convergence equal to $(nh^d)^\kappa$, where κ is the usual rate for the unconditional envelopment estimators, i.e. $\kappa = 1/(p+1)$ for the FDH and $2/(p+2)$ for the DEA

[1] Endogeneity means that the parameter of interest are not determined by the conditional distribution but by the joint distribution of Y and some variables. (see Cazals et al., 2016 or Simar et al., 2016).

estimator. This means that the errors of estimation have the order $O_p\big((nh^d)^{-\kappa}\big)$. When the bandwidths are driven by LSCV, we have $h \propto n^{-1/(d+4)}$, as a consequence, the nonparametric conditional frontier has then an order for the error given by $O_p\big(n^{-4\kappa/(d+4)}\big)$. So the usual nonparametric rate κ is still deteriorated by a factor $4/(d+4)$ by conditioning on $Z = z$. In many applications, the dimension d of Z is naturally large and thus a fully nonparametric specification of the model is jeopardized by the curse of the dimensionality.

Robust versions of the conditional frontiers have also been proposed, they share the nice properties of being more robust to extreme observations or outliers than the traditional envelopment estimators. We have the order-m frontier, introduced by Cazals et al. (2002), defined for a given integer $m \geq 1$ as

$$\varphi_m(x, z) = \mathbb{E}\left[\min(Y_1, \ldots, Y_m) \mid X \geq x, Z = z\right], \tag{1.5}$$

i.e., the expected minimum cost among m firms drawn randomly from the population of firms producing more outputs X than x and facing the environmental conditions z. Cazals et al. (2002) show that

$$\varphi_m(x, z) = \int_0^\infty S^m(y|x, z)\, dy = \varphi(x, z) + \int_{\varphi(x,z)}^\infty S^m(y|x, z)\, dy, \tag{1.6}$$

where it is clear from the second equation that $\varphi_m(x, z) \to \varphi(x, z)$ when $m \to \infty$.

An alternative robust version of frontiers is based on the order-α quantiles of $S(y|x, z)$, introduced by Aragon et al., (2005) and Daouia and Simar (2007). They define the order-α conditional frontier for some $\alpha \in]0, 1]$

$$\varphi_\alpha(x, z) = \sup\{y \mid S(y|x, z) \geq \alpha\} = S^{-1}(\alpha|x, z), \tag{1.7}$$

where $S^{-1}(\alpha|x, z)$ is the "α-quantile" of $S(y|x, z)$. It is clear that $\varphi_\alpha(x, z) \to \varphi(x, z)$ when $\alpha \to 1$. So there is a probability α of observing a firm producing more outputs X than x and facing the environmental conditions z having a cost greater or equal to $\varphi_\alpha(x, z)$. Since Y is univariate, it means the frontier allows to have a probability $1 - \alpha$ of these firms to have a lower cost than $\varphi_\alpha(x, z)$.

For finite m and $\alpha < 1$, the frontiers and their nonparametric estimators will not envelop all the data points, and so they are robust against extreme data points. These estimators are obtained by plugging $\widehat{S}_n(y|x, z)$ for $S(y|x, z)$ in the definition above. It is shown that the resulting estimators have an asymptotic normal distribution with an estimation error of the order $O_p\big((nh^d)^{-1/2}\big)$, where h is the bandwidth used to estimate $S(y|x, z)$. Again with the optimal bandwidth computed by LSCV, we have $h \propto n^{-1/(d+4)}$, resulting to an order for the estimation error given by $O_p\big(n^{-2/(d+4)}\big)$, which is better than the rates obtained for the full frontier estimators but still deteriorates quickly as d increases.

A large stream of the literature of efficiency analysis is devoted to investigate how the environmental factors Z may influence the production process and in particular the distribution of the inefficiencies. Most of the approaches so far have been based on two-stage estimations where the efficiency scores are non-parametrically estimated in a first stage and then regressed in a second stage on Z by some appropriate parametric regression model. As explained in Simar and Wilson (2007, 2011) this approach assumes a "separability" condition on the frontier, i.e. the frontier level does not depend on the value of Z, i.e.

$$\varphi_(x, z) = \varphi(x), \text{ for all } z. \tag{1.8}$$

This allows to interpret the resulting distance $y - \varphi(x)$ of a firm operating at the level (x, y) from the "marginal" frontier $\varphi(x)$, and facing environmental conditions z, as its real cost inefficiency. In practice most of these studies estimate in a first stage individual nonparametric efficiency scores, e.g. for the FDH case:

$$\widehat{\theta}(X_i, Y_i) = \inf\{\theta \mid \widehat{S}_n(\theta Y_i | X \geq X_i) < 1\}. \tag{1.9}$$

Then in a second stage, these estimated efficiencies $\widehat{\theta}(X_i, Y_i)$ are regressed on Z_i, like $\widehat{\theta}(X_i, Y_i) = g(\beta' Z_i) + \varepsilon_i$ with various appropriate specifications for g, ε_i (see Simar and Wilson, 2007 for details).

But even if the "separability" condition (1.8) holds, the estimation procedure suffers from several problems, as described in Kneip et al. (2015), mainly due to the inherent bias shared by the estimated efficiency scores in the first stage. The coefficients β, estimated in the second stage, allow to analyze the sensitivity of the efficiency to the environmental variable Z. However, the traditional tools of inference cannot be used (as explained in Simar and Wilson, 2007 and described in details in Section 5 of Kneip et al., 2015). In addition, in the best cases, after correcting for the bias and using subsampling techniques, i.e. using only $n^{2\kappa}$ observations in the second stage, the estimators of β achieve the rate κ of the chosen nonparametric estimators, which may be far from the parametric rate $1/2$ when p increases.

In this paper we suggest an alternative semi-parametric model which avoids the drawbacks of the two-stage semi-paremetric methods. It is based on a class of model called the Proportional Incremental Cost Functions (PICF), adapted to our setup from the Cox proportional hazard models extensively used in survival analysis for durations models. We define the PICF model in Section 2, and we examine its properties. Section 3 develops the semiparametric estimation. By this way of modeling, we respect the separability condition but we avoid the first stage nonparametric estimation of the frontier and so, we avoid the curse of dimensionality keeping the parametric $\sqrt{n}$ rates of convergence for the parameters of interest. As a byproduct, we are also able to derive $\sqrt{n}$-consistent estimator of the order-m

robust frontiers (which, by contrast to the full frontier, may depend on Z) and we prove the Gaussian asymptotic properties of the resulting estimators of the m-frontiers. Section 4 illustrates the flexibility of the procedure by a simulated example and with some real data sets.

2 The Model and Basic Concepts

2.1 The PICF model

If we assume that $S(y|x,z)$ has a derivative with respect to y, the density of the cost Y given $X \geq x$ and $Z = z$ is then given by

$$f(y|x,z) = -\frac{\partial}{\partial y} S(y|x,z). \tag{2.1}$$

It is well known that the survivor function is completely characterized by its "hazard" function, which in our setup here can be interpreted as an *incremental cost probability*, i.e.

$$h(y|x,z) = \lim_{\Delta \to 0^+} \text{Prob}\left(Y \in [y, y+\Delta] \mid Y \geq y, X \geq x, Z = z\right). \tag{2.2}$$

This function h represents the probability of a small increment of the cost for a firm having a cost $Y \geq y$ producing more outputs X than x and facing environmental conditions $Z = z$. Like in duration models, it is the risk function for a firm having a cost greater than y (given $X \geq x$ and $Z = z$) to face a cost Y in an infinitely small increment above y. In duration models (see e.g. Kalbfleisch and Prentice, 1980), Y represents a life time, the risk function is known as the (conditional) "instantaneous rate of failure" at the age y. Everything else being the same, we hope in general, this rate being as small as possible. In our setup of cost processes, on the contrary, if the firm encours a cost $Y \geq y$, we hope the (conditional) "instantaneous rate" at y will be as large as possible to be "cost efficient", i.e. we do not want to observe values Y bigger than the value y already achieved.

The fact that the function h characterize completely the cost process is seen by writing

$$h(y|x,z) = \frac{f(y|x,z)}{S(y|x,z)} = -\frac{\partial}{\partial y} \log S(y|x,z). \tag{2.3}$$

We may then define the integrated (or cumulative) incremental cost probability by

$$H(y|x,z) = \int_0^y h(u|x,z)\, du = -\log S(y|x,z). \tag{2.4}$$

So that we can recover the survivor finction and the density of Y given $X \geq x$ and $Z = z$:

$$S(y|x,z) = e^{-H(y|x,z)} \tag{2.5}$$

$$f(y|x,z) = h(y|x,z)\, e^{-H(y|x,z)}. \tag{2.6}$$

As in durations models, the specification of the function h has many interests. Any function h on $\mathbb{R}_+$ is a hazard function if and only if $\forall y \geq 0, h(y|x,z) \geq 0$ and $\int_0^\infty h(u|x,z)\,du = \lim_{y\to\infty} H(y|x,z) = \infty$. The latter condition ensures that the probability of an infinite cost is zero. Note also that the characterization through h is very convenient for the treatment of possibly censored data (only the minimum between y and some censoring value is observed). We will not consider this latter case in our setup.

We can now introduce the main definition of our PICF model, inspired from the Cox proportional hazard model, Cox (1972).

Definition 2.1. *A cost function satisfies the Proportional Incremental Cost Function (PICF) assumption if there exist (i), an unknown function $\beta : \mathbb{R}_+^p \to \Theta$ with $\Theta \subseteq \mathbb{R}^k$ for some $k > 0$, (ii), a known function $a : \mathbb{R}^d \times \Theta \to \mathbb{R}_+ \setminus \{0\}$ and (iii), a baseline survivor function $S_0(y|x) = \text{Prob}(Y \geq y \mid X \geq x)$ corresponding to some probability law on (Y, X), characterized equivalently either by $H_0(y|x) = -\log S_0(y|x)$ or by $h_0(y|x) = \frac{\partial}{\partial y} H_0(y|x)$, such that one of the 3 equivalent properties are satisfied:*

$$S(y|x,z) = S_0(y|x)^{a(z,\beta(x))} \tag{2.7}$$

$$H(y|x,z) = a(z,\beta(x))H_0(y|x) \tag{2.8}$$

$$h(y|x,z) = a(z,\beta(x))h_0(y|x). \tag{2.9}$$

The Cox proportional hazard model is very popular in survival analysis for duration models, due to its great flexibility. Here, the PICF assumption corresponds to a particular specification of the effect of the environmental variable. The cost process is viewed as a cost model including only the output level x (the baseline model) and an action of the environmental conditions which multiply by $a(z,\beta(x)) > 0$ the incremental cost probability. We note that the function a is a known parametric function of z, depending also of unknown parameters $\beta(x) \in \Theta$, these parameters may be x-dependent. This function links the baseline model (which does not depend on z) to the conditional to $Z = z$ model.

A particular case is to assume that $\beta(x) = \beta$ which is a common (implicit) assumption done in most of the two-stage approaches. Here we allow β to be x-dependent. Of central interest for the sensitivity analysis of the environmental variables Z on the cost process will be the analysis of this function $a(z,\beta(x))$. To see this, Figure 1 illustrates three cases, when $a(z,\beta(x)) = 1$, so that $S(y|x,z) = S_0(y|x)$ and Y is independent of Z given $X \geq x$, when $a(z,\beta(x)) > 1$, pushing the mass of $S(y|x,z)$ near the efficient boundary (so Z is favorable to efficiency) and the opposite case where $a(z,\beta(x)) < 1$ expanding $S(y|x,z)$ to the right, with more probability of being far from the efficient boundary (so Z is unfavorable to efficiency).

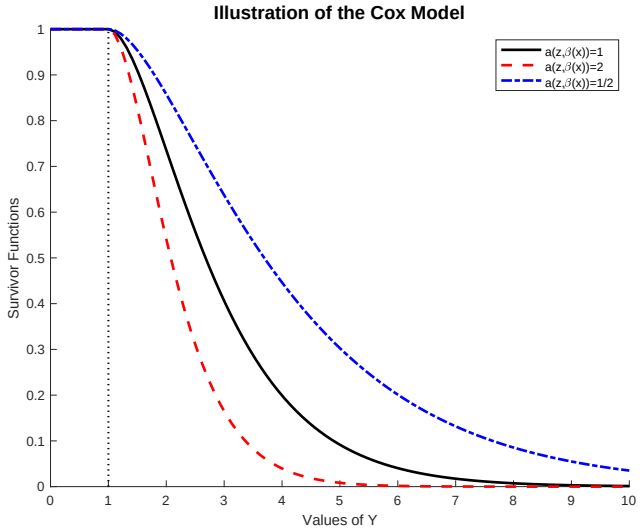

Figure 1: Survivor functions $S(y|x,z) = S_0(y|x)^{a(z,\beta(x))}$ for $a(z,\beta(x)) = 1$ (solid black), > 1 (favorable to efficiency: red dashed) and < 1 (unfavorable to efficiency: blue dot-dashed).

Note also that the baseline survivor function $S_0(y|x)$ is not the marginal survivor on Y given X of the general model. Indeed the latter is given by

$$\begin{aligned} S(y|x) &= \int_{z\in\mathbb{R}^d} S(y|x,z) f_{Z|x}(z|X \geq x)\, dz \\ &= \int_{z\in\mathbb{R}^d} S_0(y|x)^{a(z,\beta(x))} f_{Z|x}(z|X \geq x)\, dz, \end{aligned} \tag{2.10}$$

which in general is not equal to $S_0(y|x)$, unless $a(z,\beta(x)) = 1$ for all $z \in \mathbb{R}^d$, and so $S(y|x,z) = S_0(y|x)$, i.e. $Y|X \geq x$ is independent of Z.[2]

A common choice, very often used in durations models, is

$$a(z,\beta(x)) = e^{\beta'(x)z}, \text{ with } \Theta \subseteq \mathbb{R}^d. \tag{2.11}$$

Clearly, with this specification, if $\beta(x) = 0$ for all x, then $S(y|x,z) = S_0(y|x)$ for all (x,z), and $Y|X \geq x$ is independent of Z for all x. If for some j, $\beta_j(x) > 0$ indicates that an increase of Z_j will favor the efficiency at this value of the output x.

Our aim is to provide from a sample of observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ estimates of $\beta(x)$ and of $S_0(y|x)$ and these will provide estimators of the conditional survivor function $S(y|x,z)$.

Remark 2.1. *Our PICF model can also be formalized in terms of a semiparametric "non-separable" model, where "non-separable" is used in the general econometric sense (and not*

[2]If a is normalized such that for some (z_0, x_0), $a(z_0, \beta(x_0)) = 1$ the baseline model represents the cost process for this particular production unit.

in the "frontier" sense). Indeed, define $U = S(Y|x,z)$. *The variable* U *is uniform on* $(0,1)$ *and independent of* $X \geq x$ *and* Z. *Then* $H(Y|x,z) = -\log S(Y|x,z) = V$ *where* V *is an exponential:* $V \sim \text{Exp}(1)$ *independent of* $X \geq x$ *and* Z. *So we have the "non-separable" form of the model for* Y *given* $X \geq x$ *and* $Z = z$,

$$Y = H^{-1}(V|x,z) = \Lambda(V,x,z), \tag{2.12}$$

where $\Lambda(V,x,z)$ *is monotone in* V. *It is easy to show that our PICF model assumes that*

$$Y = \Lambda_0\left(\frac{V}{a(z,\beta(x))}, x\right), \tag{2.13}$$

where $\Lambda_0(\cdot,x) = H_0^{-1}(\cdot|x)$. *This model is a semiparametric "non-separable" model characterized by the two functions* $\Lambda_0 : \mathbb{R}_+ \times \mathbb{R}_+^p$ *monotone in its first argument and the function* $\beta : \mathbb{R}_+^p \to \Theta$. *The parametric part of the model is shared by the known function* a.

2.2 Cost Frontier and their Robust versions

The PICF model has some implications on the cost frontier and their robust versions. Since $a(z,\beta(x)) > 0$, $S(y|x,z) = S_0(y|x)^{a(z,\beta(x))} = 1$ if and only if $S_0(y|x) = 1$. From the definition of the full conditional cost frontier in (1.2) we have,

$$\varphi(x,z) = \sup\{y \mid S(y|x,z) = 1\} = \sup\{y \mid S_0(y|x) = 1\} = \varphi_0(x). \tag{2.14}$$

So our PICF model implies the separability condition of the frontier, in the terminology of Simar and Wilson (2007). In particular, we see also that the lower boundary of the support of the baseline survivor function $S_0(y|x)$, $\varphi_0(x)$, coincides with the marginal cost frontier $\varphi(x) = \sup\{y \mid S(y|x) = 1\}$. But as seen below, the main advantage of our approach is the fact that we do not need to estimate this frontier to make inference on the effect of Z on the cost process.

If the distribution of the conditional inefficiencies is of interest, one can analyze the distribution

$$\begin{aligned}\text{Prob}(Y - \varphi_0(x) > \varepsilon \mid X \geq x, Z = z) &= S(\varepsilon + \varphi_0(x)|x,z) \\ &= S_0(\varepsilon + \varphi_0(x)|x)^{a(z,\beta(x))}, \text{ for } \varepsilon \geq 0. \end{aligned} \tag{2.15}$$

A very particular case is to assume the homoskedasticity of the inefficiency distribution. In this case $S_0(\varepsilon + \varphi_0(x)|x)$ does not depend on x, and we have $S_0(\varepsilon + \varphi_0(x)|x) = \widetilde{S}_0(\varepsilon)$ for some survivor function $\widetilde{S}_0$ with support on $\mathbb{R}_+$. In this case the conditional survivor function simplifies as

$$S(y|x,z) = \left(\widetilde{S}_0(y - \varphi_0(x))\right)^{a(z,\beta(x))}. \tag{2.16}$$

Even if the model is a cost frontier which does not depend on the environmental variables Z (separability condition), the robust frontier may depend on Z. For the order-m conditional frontiers we have from (1.6) and our PICF model in Definition 2.1:

$$\varphi_m(x,z) = \int_0^\infty S_0(y|x)^{ma(z,\beta(x))}\,dy \tag{2.17}$$

$$= \varphi_0(x) + \int_{\varphi_0(x)}^\infty S_0(y|x)^{ma(z,\beta(x))}\,dy. \tag{2.18}$$

So we see that the conditional m-frontier is the unconditional $M_{x,z}$-frontier of the baseline model $S_0(y|x)$, with $M_{x,z} = ma(z,\beta(x))$. So we may apply previous results of Daouia et al. (2012) which allow to compare $\varphi_m(x,z)$ with $\varphi(x,z) = \varphi_0(x)$.

Lemma 2.1. *Under the regularity condition*

$$1 - S_0(y|x) = \ell_x\big(y - \varphi_0(x)\big)^{\rho_x} + o\Big(\big(y - \varphi_0(x)\big)^{\rho_x}\Big),\ \textit{as } y \downarrow \varphi_0(x), \tag{2.19}$$

where $\ell_x > 0$, $\rho_x > p$ and $\varphi_0(x)$ being differentiable in x with strictly positive first partial derivatives, we have as $m \to \infty$,

$$\varphi_m(x,z) = \varphi_0(x) + \Gamma(1+1/\rho_x)\left(\frac{1}{ma(z,\beta(x))\ell_x}\right)^{1/\rho_x} + o\big((ma(z,\beta(x)))^{-1/\rho_x}\big). \tag{2.20}$$

The proof derives from equation (3.5) in Daouia et al. (2012), where it is shown that condition (2.19) implies that the conditional baseline density f_0 of Y given $X = x$ is given by

$$f_0(y|x) = c_x\big(y - \varphi_0(x)\big)^{\eta_x} + o\Big(\big(y - \varphi_0(x)\big)^{\eta_x}\Big),\ \text{as } y \downarrow \varphi_0(x), \tag{2.21}$$

with $c_x > 0$ and $\eta_x = \rho_x - (p+1) > -1$. As a particular case, if this density has a jump at the frontier, $\eta_x = 0$ and $\rho_x = p+1$, which is a common assumption in the literature on frontiers (parametric and non-parametric).

A similar result holds for the order-α conditional quantile frontier defined in (1.7). Since $a(z,\beta(x)) > 0$, here we have under the PICF model

$$\varphi_\alpha(x,z) = S^{-1}(\alpha|x,z) = S_0^{-1}\big(\alpha^{1/a(z,\beta(x))}|x\big). \tag{2.22}$$

So the conditional order-α frontier $\varphi_\alpha(x,z)$ is the order-$\alpha^{1/a(z,\beta(x))}$ frontier of the baseline model $S_0(y|x)$. Similar approximation results as Lemma 2.1 are established in Daouia et al. (2010), by using appropriate sequences of $\alpha \to 1$ and under the same regularity condition.

In the rest of this paper we will focus on the order-m frontiers: we propose an estimator and we analyze its asymptotic properties. Analog results should be available for the order-α case.

3 Semiparametric Estimation

3.1 Estimation of the components of the PICF model

The interest of our model is that it allows a complete separation of the estimation of its components: the cost frontier and the parameters allowing to analyze the impact of the environmental variables on cost process. We will first consider the estimation of the three components of our model: the cost frontier φ_0, the functional parameters $\beta(x)$ and the baseline survivor S_0. Then we will see how these provide an estimator of the conditional order-m frontier, sharing better properties than the classical one.

3.1.1 Estimation of the frontier

Since the frontier does not depend on Z, we can estimate the frontier $\varphi_0(x)$ by standard nonparametric envelopment estimators from the sample of observations $\{(X_i, Y_i)\}_{i=1}^n$. The most flexible is the FDH estimator defined as

$$\widehat{\varphi}_0(x) = \inf\{Y_i \mid X_i \geq x\}. \tag{3.1}$$

This is a well known estimator with well known properties (see Park et al., 2000). We will not discuss further this estimator since we do not need to use it for the estimation of the pieces of the PICF model. This underlines the major difference of our method with the classical two-stages methods where the estimated inefficiencies $Y_i - \widehat{\varphi}_0(X_i)$ are regressed on the variables Z_i by some appropriate model, with the various drawbacks mentioned above.

3.1.2 Estimation of the parameters

The estimation of the functional parameters $\beta(x)$ has been intensively analyzed in the duration models literature. It is based on the marginal likelihood approach of Cox (1972) (see e.g. Kalbfleisch and Prentice, 1980 for discussion and justifications). We consider a particular value of the output $x \in \mathbb{R}_+^p$ and we define the set $\mathcal{I}_x = \{i \mid X_i \geq x\}$ which determines the set of observations $\{(Y_i, Z_i)\}_{i \in \mathcal{I}_x}$ with output larger or equal to x. It is the latter subsample that will be used to estimate $\beta(x)$. The Cox marginal likelihood is then defined as

$$\ell(\beta(x)) = \prod_{i \in \mathcal{I}_x} \frac{a(Z_i, \beta(x))}{\sum_{j \in \mathcal{I}_x} a(Z_j, \beta(x)) \mathbb{1}(Y_j \geq Y_i)}, \tag{3.2}$$

where $\mathbb{1}(B)$ is the usual indicator function, i.e. $\mathbb{1}(B) = 1$ if B is true and 0 otherwise. We remark that the baseline S_0 does not appear, so the vector $\beta(x)$ can be estimated without

the need of the baseline model.[3] It has been noticed in the literature that only the orderings (the ranks) of the observations Y_i matter and $\ell(\beta(x))$ is known as the distribution of the rank statistics of the observations $Y_i, i \in \mathcal{I}_x$. For any given x, the estimator is defined as

$$\widehat{\beta}(x) = \arg\max_{\beta(x)} \log \ell(\beta(x)), \tag{3.3}$$

where it should be noticed that the number of observations used in $\mathcal{I}_x$ is the number of observations such that $X_i \geq x$, i.e. $n_x = \sum_{i=1}^{n} \mathbb{1}(X_i \geq x)$. The asymptotic properties have been established, we know that if $S_X(x) > 0$, as $n_x \to \infty$, $\widehat{\beta}(x) \xrightarrow{p} \beta(x)$ and that

$$\sqrt{n_x}\big(\widehat{\beta}(x) - \beta(x)\big) \xrightarrow{\mathcal{L}} \mathcal{N}_k\Big(0, \big[-\partial^2 \log \ell(\beta(x))\big]^{-1}\Big), \tag{3.4}$$

where $\partial^2 \log \ell(\beta(x))$ denotes the matrix of second derivatives. Note that only n_x observations are used but we keep the parametric rate of convergence (no curse of dimensionality, as for the traditional two-stage approaches). Most of the statistical packages provide efficient algorithms for solving (3.3) for the particular and popular choice suggested by Cox (1972), $a(z, \beta(x)) = e^{\beta'(x)z}$, where $k = d$. With this latter choice, the score function and the hessian are indeed easy to derive. The asymptotic result in (3.4) can be used for practical inference.

In many applications, it may be reasonable to assume that $\beta(x)$ is constant, i.e. $\beta(x) = \beta$ for all x. If this is the case, we can do the estimation above for only $x = 0$ so that $n_x = n$ and we reach for the estimator $\widehat{\beta}$ the $\sqrt{n}$-rate of convergence in (3.4).

3.1.3 Estimation of the baseline survivor function

The last element of the model is the baseline survivor function $S_0(y|x)$. We may use the Nelson-Aalen estimator, which is a discrete survivor function with jumps at the observed $Y_{(i)}$, where $Y_{(i)}$ is the order statistic of the $Y_i, i \in \mathcal{I}_x$ ($Y_{(1)} < \ldots < Y_{(n_x)}$). It can be written as follows (see e.g. Kalbfleisch and Prentice, 1980 for details):

$$\widehat{S}_0(y|x) = \sum_{i=1}^{n_x} \widehat{p}_i \, \mathbb{1}(y < Y_{(i)}), \tag{3.5}$$

where the jumps are the probabilities $\widehat{p}_i$ given by the equations:

$$\widehat{p}_i = \widehat{\lambda}_i \prod_{j \in \{\mathcal{I}_x \cap Y_{(j)} < Y_{(i)}\}} (1 - \widehat{\lambda}_j), \tag{3.6}$$

[3] We limit our presentation for the case of no ties in the Y_i and no censoring which is mostly the case in our setup of cost efficiency analysis. The marginal likelihood can easily be extended to the case of ties and censored data (only the minimum between Y and some censoring value is observed). See e.g. Kalbfleisch and Prentice (1980).

i.e. $\widehat{p}_1 = \widehat{\lambda}_1$, $\widehat{p}_2 = \widehat{\lambda}_2(1-\widehat{\lambda}_1)$, $\widehat{p}_3 = \widehat{\lambda}_3(1-\widehat{\lambda}_1)(1-\widehat{\lambda}_2)$, etc..., and

$$\widehat{\lambda}_i = 1 - \left(1 - \frac{a(Z_i, \widehat{\beta}(x))}{\sum_{j\in\mathcal{I}_x} a(Z_j, \widehat{\beta}(x))\mathbb{1}(Y_{(j)} \geq Y_{(i)})}\right)^{a^{-1}(Z_i,\widehat{\beta}(x))}. \tag{3.7}$$

The properties of $\widehat{S}_0(y|x)$ have been established and we know that for all y in the support of $S_0(y|x)$, as $n_x \to \infty$, we have

$$\sqrt{n_x}\big(\widehat{S}_0(y|x) - S_0(y|x)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2_{xy}), \tag{3.8}$$

for some some finite σ^2_{xy}.

Finally, if an estimate of the conditional survivor is needed for some given value of the environmental variables z, it is given by

$$\widehat{S}(y|x,z) = \Big(\widehat{S}_0(y|x)\Big)^{a(z,\widehat{\beta}(x))}. \tag{3.9}$$

3.1.4 Functional convergence

We have even a stronger asymptotic results. As above denoting $H_0(y|x) = -\log S_0(y|x)$, Tsiatis (1981) has shown (see proof of Lemma 6.1) that, as $n \to \infty$,

$$\sqrt{n_x}\begin{pmatrix} \widehat{H}_0(y|x) - H_0(y|x) \\ \widehat{\beta}(x) - \beta(x) \end{pmatrix} \longrightarrow \begin{pmatrix} V_{1,x}(y) + \phi_x(y)V_{2,x} \\ V_{2,x} \end{pmatrix}, \tag{3.10}$$

where $V_{1,x}(y)$ is an independent increments Gaussian process with mean zero, $V_{2,x}$ is a normal random variable with mean zero independent of $V_{1,x}(y)$ for all y and $\phi_x(y)$ is a function. The covariance structure of $V_{1,x}(y)$, $\mathrm{Cov}(V_{1,x}(y), V_{1,x}(\tilde{y})$, and the variance of $V_{2,x}$ are given in equation (5.7) of Tsiatis (1981).

Since $S_0 = \exp(-H_0)$, by using a Taylor expansion we obtain as $n \to \infty$,

$$\sqrt{n_x}\begin{pmatrix} \widehat{S}_0(y|x) - S_0(y|x) \\ \widehat{\beta}(x) - \beta(x) \end{pmatrix} \longrightarrow \begin{pmatrix} S_0(y|x)\left[V_{1,x}(y) + \phi_x(y)V_{2,x}\right] \\ V_{2,x} \end{pmatrix}. \tag{3.11}$$

This results can also be applied to the estimator of $S(y|x,z)$. We come back to these results in Section 3.2.

3.1.5 A bootstrap algorithm

The relation (3.4) is useful to produce inference on the components of $\beta(x)$, if the covariance matrix $\partial^2 \log \ell(\beta(x))$ has been derived. An alternative is to use a bootstrap approach to evaluate the covariance matrix. Note that for the estimators of the survivor functions, the

variance are more complex to evaluate, hence the bootstrap will be particularly useful. The asymptotic result in (3.11) is sufficient to prove that the bootstrap is consistent for both $\beta(x)$ and for $S_0(y|x)$ (see e.g. Theorem 1 in Mammen, 1992).

The algorithm for generating bootstrap samples in the setup here is very simple. The idea is to use the representation of the PICF model defined in the semiparametric model (2.13), i.e. $Y = \Lambda_0\left(\frac{V}{a(z,\beta(x))}, x\right)$, where $\Lambda_0(\cdot, x) = H_0^{-1}(\cdot|x)$ and $V \sim \text{Exp}(1)$. The estimates of the model provide $\widehat{\beta}(x)$ and $\widehat{H}_0(y|x) = -\log \widehat{S}_0(y|x)$. So the so-called "Cox-Snell" residuals are given for each $i \in \mathcal{I}_x$, by $\widehat{V}_i = \widehat{H}_0(Y_i|x)a(Z_i, \widehat{\beta}(x))$. After rescaling these residuals to have a mean 1, we obtain the "centered" residuals $\widetilde{V}_i = \widehat{V}_i/\overline{V}_{n_x}$, where $\overline{V}_{n_x} = n_x^{-1}\sum_{i\in\mathcal{I}_x}\widehat{V}_i$.

So one bootstrap sample is generated according to the following steps for a given x.

[1] Draw randomly with replacement n_x values V_i^* among the centered residuals $\widetilde{V}_i$.

[2] Define the bootstrap value for $i \in \mathcal{I}_x$,

$$Y_i^* = \widehat{\Lambda}_0(V_i^*/a(Z_i, \widehat{\beta}(x)), x) = \widehat{H}_0^{-1}(V_i^*/a(Z_i, \widehat{\beta}(x))|x). \tag{3.12}$$

The latter equation (3.12) can be solved numerically by $y = \arg\min_y |\widehat{H}_0(y|x) - \xi|$ where $\xi = v/a(z, \widehat{\beta}(x))$.

This provides a bootstrap sample $\{(Y_i^*, Z_i)\}_{i\in\mathcal{I}_x}$ which leads to the estimates $\widehat{\beta}^*(x), \widehat{S}_0^*(y|x)$ by applying the procedures described above. Repeating all of this B times this produces the bootstrap distribution of the quantities of interest: $\widehat{\beta}^{*,b}(x), \widehat{S}_0^{*,b}(y|x)$, for $b = 1, \ldots, B$.

Note that due to the property of V, the step [1] above can be replaced by drawing the n_x values V_i^* from an Exp(1).

3.2 Estimation of conditional order-m robust frontier

The results of the preceding sections can be used to provide estimator of $\varphi_m(x, z)$, the conditional order-m cost frontier defined in (1.6) and to derive its asymptotic properties. Indeed even under the separability assumption of the full frontier (for all z, $\varphi(x, z) = \varphi_0(x)$), the conditional order-m frontier may still depend on z since $S(y|x, z)$.[4]

Under our PICF model, and using (2.17), we may define the following estimator of $\varphi_m(x, z)$

$$\widehat{\varphi}_m(x, z) = \int_0^\infty \widehat{S}_0(y|x)^{ma(z,\widehat{\beta}(x))}\, dy. \tag{3.13}$$

For analyzing its asymptotic behavior, we can use Taylor expansion of this function of $\widehat{S}_0(y|x)$ and $\widehat{\beta}(x)$ around the true values $S_0(y|x)$ and $\beta(x)$, to approximate, for a given

[4]Similar developments could be done for the conditional order-α frontiers.

(x,z), $\widehat{\varphi}_m(x,z) - \varphi_m(x,z)$ by a linear expression in $\widehat{S}_0(y|x) - S_0(y|x)$ and $\widehat{\beta}(x) - \beta(x)$. We can indeed show that

$$\begin{aligned}\widehat{\varphi}_m(x,z) - \varphi_m(x,z) &= \int_0^\infty \left(\widehat{S}_0(y|x)^{ma(z,\widehat{\beta}(x))} - S_0(y|x)^{ma(z,\beta(x))}\right) dy \\ &\approx \int_0^\infty ma(z,\beta(x)) S_0(y|x)^{ma(z,\beta(x))-1} \left(\widehat{S}_0(y|x) - S_0(y|x)\right) dy \\ &+ \int_0^\infty m \log S_0(y|x) S_0(y|x)^{ma(z,\beta(x))} \left(a(z,\widehat{\beta}(x)) - a(z,\beta(x))\right) dy. \quad (3.14)\end{aligned}$$

Under regularity conditions on the function a, the difference between $a(z,\widehat{\beta}(x))$ and $a(z,\beta(x))$ can also be linearized and written as a function of $\widehat{\beta}(x) - \beta(x)$.

Due to the asymptotic property given in (3.11), we obtain the desired result

$$\sqrt{n_x}\left(\widehat{\varphi}_m(x,z) - \varphi_m(x,z)\right) \to \mathcal{N}(0, \omega_{x,z}), \quad (3.15)$$

where $\omega_{x,z}$ has a complicated expression, but for practical inference we can use the bootstrap method described above.

We see that, under our PICF semiparametric model, we can provide estimates of $\varphi_m(x,z)$ with $\sqrt{n}$-rate of convergence and we avoid the curse of dimensionality shared by the traditional direct estimator which has a rate $\sqrt{nh^d}$, where d is the dimension of z, as derived in Cazals et al. (2002). If least-squares cross-validation methods are used for obtaining optimal bandwidths of order $n^{-1/(d+4)}$ (see Li et al., 2013), this leads to the optimal rate $n^{2/(d+4)}$ for the traditional nonparametric estimator, which deteriorates quickly as d increases.

In practice, for selecting a value for m, we follow the traditional approach described e.g. in Simar (2003) and analyzed in details in Daouia and Gijbels (2011) from a theory of robustness perspective. We can tune the percentage of points that remain outside the order-m frontier as a function of m, knowing that when $m \to \infty$ all the points will be above the cost frontier. A reasonable m is when this curve, as a function of m, shows an elbow effect.

Of course of particular interest for the practitioner are the order-m conditional cost efficiency for a unit operating at the level (x,y) and facing the environmental conditions z, this would be given by $\varphi_m(x,z) - y$, or some appropriate transformation of it. This is easy to obtain since the conditional frontiers function $\varphi_m(x,z)$ that can be computed at any point (x,z). Then we can derive a measure of efficiency for any particular unit and, by using the bootstrap described above, we can also derive confidence intervals of these measures.

4 Numerical Illustrations

4.1 Simulated sample

In Appendix A we explain that care has to be taken for simulating a data set $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ according to some PICF model: the difficulty may arise when we simulate, as usual data from conditional models for $Y|X = x, Z = z$, which is standard in the frontier literature but we want to satisfy the PICF condition $S(y|X \geq x, Z = z) = (S_0(y|X \geq x))^{a(z,\beta(x))}$ for some S_0 and $a(z, \beta(x))$. We illustrate this through a simple example.

We select a frontier function given by $y = \varphi_0(x) = x$ with only one output x. We choose the most popular Cox specification for $a(z, \beta) = \exp(\beta' z)$. We select for the baseline $f_0(x)$ an exponential density with mean 1 and for the baseline $\widetilde{S}_0(y|X = x) = \text{Prob}(Y \geq y|X = x)$ we define an exponential survival of mean 1, shifted above the cost frontier $y \geq \varphi_0(x) = x$. So we have $\widetilde{S}_0(y|X = x) = \exp(y - x)\mathbb{1}(y \geq x)$.[5] Simple algebra leads by (A.6) to

$$S_0(y|X \geq x) = 1 - F_\gamma(y - x, 2, 1), \tag{4.1}$$

where $F_\gamma(t, a, b)$ is a cumulative distribution of a Gamma density with shape parameter a and scale b. To be explicit the gamma density is defined here for $t > 0$ as $f_\gamma(t, a, b) = \frac{t^{(a-1)} b^a e^{-bt}}{\Gamma(a)}$.

In the numerical example we select $d = 2$ with $\beta = (0.5 \quad 0.25)'$ and Z_i as independent uniform variable on $[0, 1]$ then by our model, $X_i|Z_i \sim \text{Exp}(\exp(\beta' Z_i))$, i.e. an exponential with mean $1/\exp(\beta' Z_i)$. Then we generate the Y_i by simulating $U_i \sim \text{Unif}(0, 1)$ and solving $Y_i = Q(U_i, X_i, Z_i)$ where Q is the inverse function of Q^{-1} given in (A.8). In our case here, we have $Q(U_i, X_i, Z_i) = \widetilde{S}^{-1}(U_i|X = X_i, Z = Z_i)$ and $\widetilde{S}(y|X = x, Z = z) = [S_0(y|X \geq x)]^{a(z,\beta)-1} \widetilde{S}_0(y|X = x)$, see (A.8).[6] So to summarize, since $\beta_j > 0$, $j = 1, 2$ large values of Z will produce on the average small values of the output X and higher efficiency since $\partial/\partial z_j(a(z, \beta)) > 0$ for $j = 1, 2$.

Figure 2 displays one such particular sample of size $n = 400$. The figure displays also the true frontier function $\varphi_0(x) = x$ and its FDH estimator. Note that the latter is not used in the estimation of our model (i.e. for estimating β and $S_0(y|X \geq x)$).

[5] As explained in the Appendix, we use the $\widetilde{S}$ notation for survivor functions when we condition to $X = x$, to distinguish form S where we condition on $X \geq x$.

[6] Since $Q^{-1}(y, x, z)$ is specified, the value of y corresponding to a quantile $u \in [0, 1]$ is given by $y = Q(u, x, z)$ and can be found numerically by solving $y = \arg\min_y |Q^{-1}(y, x, z) - u|$, which is easy since Q^{-1} is monotone in y.

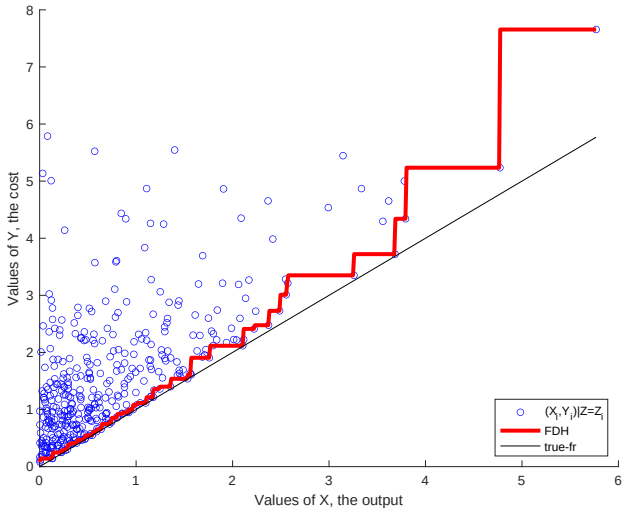

Figure 2: A simulated data set of size $n = 400$.

4.1.1 Estimation of $\beta(x)$

We give in Table 1 the estimation of $\beta(x)$ at 4 values of $x = 0, 0.5, 1$ and 2. If we assume that $\beta(x) = \beta$, as in most of the two-stage approaches, only the case $x = 0$ has an interest since all the n data points are used in the estimation procedure. As expected, when the value of x increases we have less available data and so the quality of the estimation of β is deteriorated (the rate of convergence is $\sqrt{n_x}$). In particular for $x = 2$, only 29 observations are available. We see indeed that the estimated standard deviations increases when x becomes larger. But we will see that the estimation of the basis survivor is fair even in the latter case, the same will be true for the estimation of the order-m frontier $\varphi_m(x, z)$ (see below).

Table 1: Estimation of β for 4 values of x in the simulated example.

	$x = 0, n_x = 400$		$x = 0.5, n_x = 196$		$x = 1.00, n_x = 101$		$x = 2.00, n_x = 29$	
True β	$\widehat{\beta}(x)$	$\widehat{\sigma}_\beta$	$\widehat{\beta}(x)$	$\widehat{\sigma}_\beta$	$\widehat{\beta}(x)$	$\widehat{\sigma}_\beta$	$\widehat{\beta}(x)$	$\widehat{\sigma}_\beta$
0.50	0.4942	0.1752	0.7203	0.2615	0.9614	0.3770	2.6765	0.8929
0.25	0.0465	0.1797	-0.0370	0.2661	-0.0462	0.3817	-0.1650	0.8393

Note that we also computed the standard deviation of the estimates by using the bootstrap algorithm described above and obtained, as expected, very similar values with $B = 2000$ replications. We have $\widehat{\sigma}_{\text{boot}} = (0.1732 \quad 0.1756)'$ for $x = 0$, $\widehat{\sigma}_{\text{boot}} = (0.2440 \quad 0.2623)'$ for $x = 0.5$ and $\widehat{\sigma}_{\text{boot}} = (0.3455 \quad 0.3846)'$ for $x = 1$ and $\widehat{\sigma}_{\text{boot}} = (0.8850 \quad 0.7066)'$ for $x = 2$.

4.1.2 Estimation of the basis survivor $S_0(y|X \geq x)$ and of $S(y|X \geq x, Z = z)$

We display in Figure 3, the true baseline survivor $S_0(y|X \geq x)$ and its estimates. For illustrative purpose we also display the estimate of the survivor $S(y|X \geq x, Z = z)$ for $z = \overline{Z}_{n_x}$, the mean of Z for observations such that $X_i \geq x$. So we have $\widehat{S}(y|X \geq x, Z = \bar{Z}_{n_x}) = \left[\widehat{S}_0(y|X \geq x)\right]^{\exp(\widehat{\beta}'\overline{Z}_{n_x})}$.

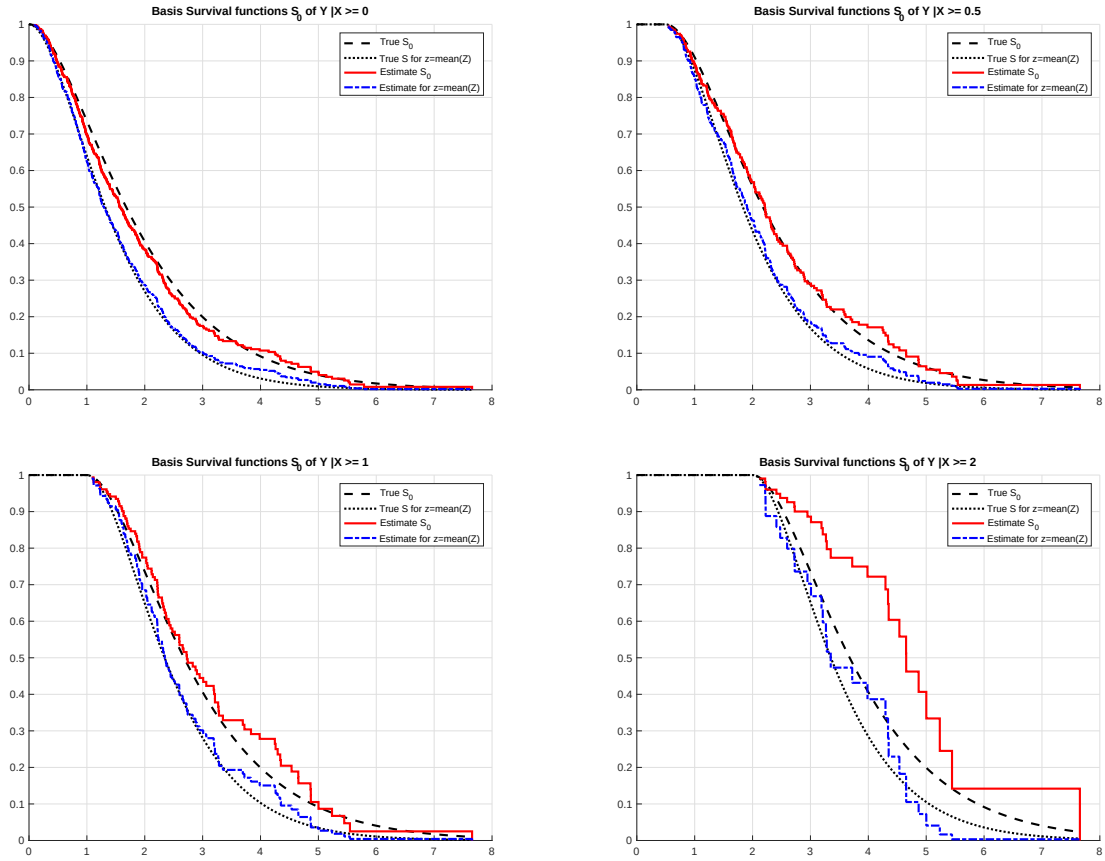

Figure 3: Survivor functions in the simulated examples for several values of $x = 0, 0.5, 1$ and 2. The respective used sample size is $n_x = 400, 196, 101$ and 29. The black dash-dotted is the true function and the solid red is its estimates. The dash-dotted blue is the conditional to Z survivor for an average value of $Z = \overline{Z}_n$, the dotted black is the corresponding true value.

We observe here that even if x increases, the estimation error remains reasonable. We observe that for the survivor conditional on $Z = \overline{Z}_{n_x}$, the estimation error is smaller, in particular near the boundary point. As shown below, this has important (good) consequences when estimating order-m frontiers (when m is large, $S^m(y|X \geq x, Z = z)$ takes values near zero when y leaves the boundary value).

4.1.3 Estimation of conditional order-m frontiers $\varphi_m(x,z)$

We also estimate, for illustration, several order-m frontiers $\varphi_m(x,z)$ by using the estimator $\widehat{\varphi}_m(x,z)$ defined in (3.13). We select here the same values of $x = 0, 0.5, 1$ and 2 and we condition again on $z = \overline{Z}_{n_x}$. The results are displayed in Table 2 where the case $m = \infty$ corresponds to the full frontier true values $\varphi_0(x) = x$ and the estimates are the corresponding FDH values. We observe that as expected the level of the error increases with x, as above and as shown in the length of the 95% confidence intervals (CI) of $\varphi_m(x,z)$, but that still the estimation errors are rather small even when $x = 2$ and only $n_x = 29$ observations are available. We see also, as it should, when m increases, order-m frontiers approach the full frontier values ($m = \infty$). Note also that with our simulated sample, the 95% bootstrap CIs cover the true value in most of the cases of Table 2. Here we used 2000 bootstrap replications and the "robust" basic-bootstrap method for building the confidence intervals.

Table 2: Estimation of order-m frontiers with $z = \overline{Z}_{n_x}$, in the simulated example. The case $m = \infty$ corresponds to the full frontier. For each finite value of m the second row gives the 95% confidence interval of $\varphi_m(x,z)$ computed by bootstrap (basic bootstrap method).

	$x = 0, n_x = 400$		$x = 0.5, n_x = 196$		$x = 1.00, n_x = 101$		$x = 2.00, n_x = 29$	
order-m	$\varphi_m(x,z)$	$\widehat{\varphi}_m(x,z)$	$\varphi_m(x,z)$	$\widehat{\varphi}_m(x,z)$	$\varphi_m(x,z)$	$\widehat{\varphi}_m(x,z)$	$\varphi_m(x,z)$	$\widehat{\varphi}_m(x,z)$
$m = 10$	0.3761	0.3879	0.8802	0.8471	1.3830	1.3991	2.3853	2.3162
95% CI	[0.3335, 0.4359]		[0.7643, 0.9222]		[1.2481, 1.5223]		[2.0618, 2.4561]	
$m = 50$	0.1562	0.1877	0.6579	0.6228	1.1589	1.1368	2.1598	2.1429
95% CI	[0.1402, 0.2311]		[0.5430, 0.6734]		[1.0021, 1.1823]		[2.0191, 2.1688]	
$m = 100$	0.1085	0.1452	0.6096	0.5761	1.1104	1.1005	2.1110	2.1232
95% CI	[0.0994, 0.1834]		[0.4996, 0.6024]		[0.9854, 1.1191]		[2.0179, 2.1295]	
$m = 200$	0.0758	0.1167	0.5765	0.5538	1.0770	1.0865	2.0775	2.1173
95% CI	[0.0689, 0.1453]		[0.4906, 0.5616]		[1.0223, 1.0919]		[2.0104, 2.1177]	
$m = \infty$	0	0.0813	0.5000	0.5458	1	1.0812	2	2.1169

4.2 Some real data examples

We illustrate in this section the flexibility and the power of the PICF model for investigating the effect of environmental variables on the production process. We first use two popular data sets already discussed in the literature and a new data set coming from GRDF, the French company of the network for gaz distribution.

4.2.1 Bank efficiency

We first show how our method works with the data set used in Simar and Wilson (2007) for illustrating two-stage models. We select as in Simar and Wilson (2007), Bădin et al. (2012)

and Florens et al. (2014) a subsample of 303 banks. There are 3 inputs (purchased funds, core deposits and labor) and 4 outputs (consumer loans, business loans, real estate loans and securities helds). In the two latter papers, it is shown how the inputs (being highly correlated) can be aggregated in a one-dimensional inputs without losing much information and the same is true for the 4 outputs (see Bădin et al. (2012) for details). In our example we investigate the effect of the variable Z = "Diversity", a measure of the diversity of the services proposed by the banks. Bădin et al. (2012) and Florens et al. (2014), using traditional conditional efficiency measures recognize that this variable has little effect on the frontier level, but may influence the inefficiency (Bădin et al. (2012) noticed, in a descriptive way, a slight favorable effect, i.e. banks with more diversity seemed to have a distribution of their efficiency slightly more concentrated near the efficient frontier).

We first test the "separability" condition and we apply the test introduced in Daraio et al. (2018) and improved in Simar and Wilson (2020). According to the terminology of the latter, we used 100 splittings of the sample with 20 shuffles for the jackknife bias corrections, and we used 200 bootstrap replications to evaluate the p-value which is around 0.46. So there is indeed no evidence against the separability assumption.

Then we estimate our PICF model and we will focus in this illustration, on the minimal input (or cost) frontier and we choose the traditional $\exp(\beta z)$ for the proportionality factor. The results of the estimation are shown in Table 3 for four values of $x \in \{\min(X_i), Q_1(X), Q_2(X), Q_3(X)\}$, i.e. the minimal observed output X and its 3 quartiles.

Table 3: Estimation of β for 4 values of x in the Bank example.

x	0.1401	0.5570	1.1330	2.1567
n_x	303	227	152	76
$\widehat{\beta}(x)$	1.2157	1.2771	1.2871	1.5950
$\widehat{\sigma}_{\beta_x}$	0.3026	0.3480	0.4241	0.6023

We first note that in each case, $\widehat{\beta}(x) > 0$ which confirms in our model the favorable effect of "Diversity" on the distribution of the efficiencies, and the standard deviations show it is significant. Of course when $x = Q_3(X)$ we only have $n_x = 76$ and the standard deviation of the estimator is larger, but still it seems the value of $\beta(x)$ seems to be very stable over the various values of x. So in practice the analysis could be concentrated on the first case, with $n_x = 303$.

The Cox proportional hazard model offers a battery of graphical diagnostic to check the validity of the model. We have first a ProbPlot of the "Cox-Snell" residuals $\widehat{V}_i$, defined above, which are supposed to be distributed as an Exp(1) random variable. In our case here

we display the plot in Figure 4 for the value minimal value of $x = 0.1401$. We see indeed a reasonable fit by our model.

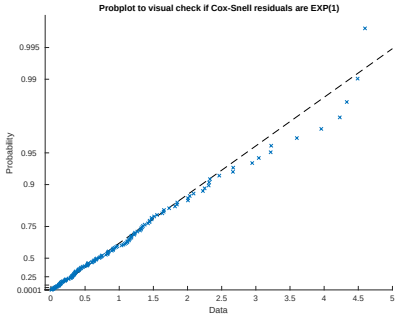

Figure 4: ProbPlot of the Cox-Snell residuals when $x = 0$ versus an Exp(1), Bank example.

The Cox model also provides the scaled Schoenfeld residuals for the Z variables when $x = 0$ (we could do the same for other values of x). As explained in Grambsch and Therneau (1994), these pictures provide graphical diagnostic of the validity of the Cox proportional model as they reveal if the coefficients $\beta(x)$ follow a particular pattern as a function of y. Here, as seen in Figure 5, there is no clear pattern confirming that the PICF model seems to be reasonable to fit our conditional survival functions.

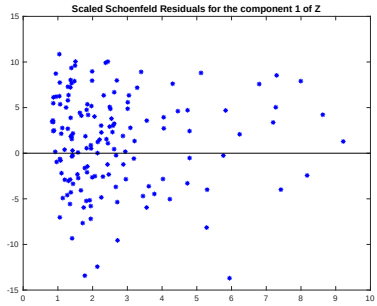

Figure 5: Scaled Schoenfeld residuals for the variable Z variable when $x = 0$ as a function of Y_i (horizontal axis), for the Bank example.

Finally we can look at the estimates of the basis survivor function $S_0(y|x)$ and the conditional ones $S(y|x, z)$, when (case 1) x is the minimum of X, for 4 selected values of z. We have also the pictures for the larger values of x, the functions are quite similar except a shift to the right since the cost frontier increases with the value of the output. To save space we only provide the case (case 3) where x is the median of X. The results are shown in Figures 6 and 7.

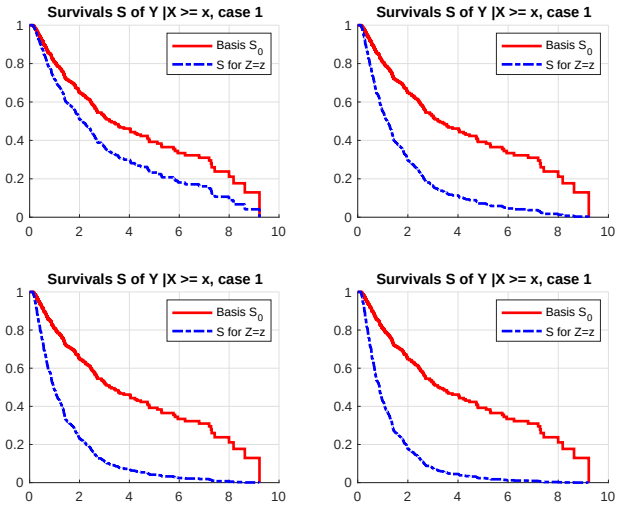

Figure 6: Estimates of the survival functions $S_0(y|x)$ and $S(y|x,z)$, when (case 1), x is the minimum of X, for 4 selected values of z. From left to right and top to bottom panels, the minimum and the 3 quartiles of Z.

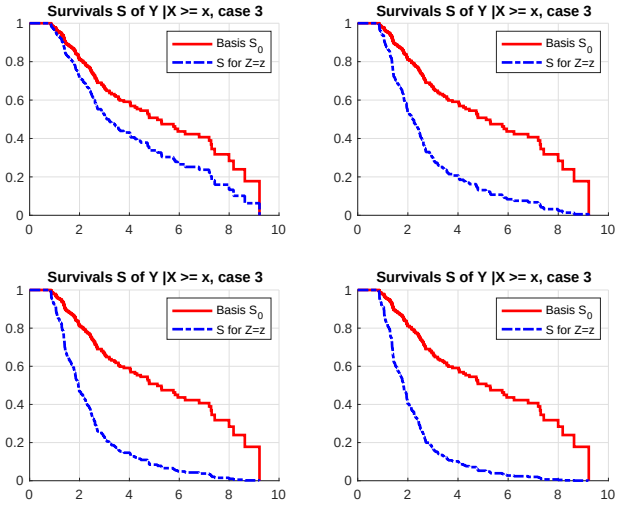

Figure 7: Estimates of the survival functions $S_0(y|x)$ and $S(y|x,z)$, when (case 3), x is the median of X, for 4 selected values of z. From left to right and top to bottom panels, the minimum and the 3 quartiles of Z.

Finally, we also may provide estimates of the conditional order-m frontiers $\varphi_m(x,z)$ and confidence intervals at any selected value of (x,z). For the points estimates, Table 4 displays the estimates for 2 values of x: the minimum of X and the median of X and the 4 selected

values of z described in Figures 6 and 7 for both values of x. For choosing a value for m we follow the traditional approach described at the end of Section 3. We select $m = 300$ which left around 16% of the banks below the order-m input frontier (much larger values of m would finally get all the data points above the frontier, e.g. when $m \to \infty$). We see the effect of Z on the order-m frontiers which is switched to the left when z increases, since, as noticed above, z is favorable to efficiency.

Table 4: Estimation of conditional order-m frontiers $\varphi_m(x, z)$ for 2 values of x and 4 values of z, in the Bank data set. Here $m = 300$.

x	$\min(X)$, $n_x = 303$	median(X), $n_x = 152$	
$\varphi(x)$	0.1253	0.8345	
$\varphi_m(x, z_1)$	0.1439	0.8452	$z_1 = 0.3652$
$\varphi_m(x, z_2)$	0.1359	0.8389	$z_2 = 0.8518$
$\varphi_m(x, z_3)$	0.1338	0.8375	$z_3 = 1.0055$
$\varphi_m(x, z_4)$	0.1322	0.8365	$z_4 = 1.1417$

Note that, just for illustrative purpose, at $x = \min(X)$ and $z = z_3$, the bootstrap provides a 95% confidence interval for $\varphi_m(x, z_3)$ given by $[0.1059, 0.1420]$ (2000 bootstrap loops and basic bootstrap method).

4.2.2 The CCR Schools data

We illustrate now the flexibility of the PICF model with the popular data set from Charnes et al. (1981), referred as the CCR data, where the performance of 70 schools is analyzed, 49 of them benefited from Program Follow Through (PFT) and 21 called the No-Follow Through (NFT). This CCR-paper gives the data for 3 outputs (achievements of students on some tests) and 5 inputs describing 4 characteristics of the family and the number of teachers. In our illustration, and due to the limited number of data points, we limit our model, as in Simar and Zelenyuk (2011) or Simar et al. (2016), to the simplest production model with one input (X :the number of teachers) and one output (Y: a weighted average of the 3 achievement tests which are highly correlated, using factorial methods as described in Section 6.2.3 of Daraio and Simar, 2007). We will consider rather three characteristics of the family (Z_1: Education level of the mother, Z_2: Highest occupation of a family member and Z_3: Parental counseling index) as environmental variables and we will investigate their eventual role on the production process. We also eliminate the observation for units #44 and #59, considered as outliers in many studies (see e.g. Wilson, 1993 and Simar, 2003). We will focus in this illustration, as in the preceding sections, on the minimal input (or cost) frontier and we choose again $\exp(\beta' z)$ as the proportionality factor.

First we investigate if the three Z variables influence the shape of the efficient frontier. We apply the test of separability (we used 100 splittings of the sample with 20 shuffles for the jackknife bias corrections, and we used 1000 bootstrap replications to evaluate the p-values) and we obtained a p-values near 1, indicating that there is no reason to reject the null hypothesis of separability. Due to the small sample size ($n = 68$) we estimate the $\beta(x)$ only for $x = 0$ and for $x = 3.34$ which is the median of the X_i. The results for the estimation of $\beta(x)$ are shown in Table 5, where we see a very small positive effect (favorable to efficiency) for Z_1 and Z_3 and a more important unfavorable effect for Z_2, which dominates the two other effects. In addition, the effects seem rather "stable" relative to the two values of x. However, in this illustration, the number of data points is too small to detect significant effects (too large standard deviations).

Table 5: Estimation of β for 2 values of x in the School data set.

	$x = 0, n_x = 68$		$x = 3.34, n_x = 34$	
	$\widehat{\beta}(x)$	$\widehat{\sigma}_\beta$	$\widehat{\beta}(x)$	$\widehat{\sigma}_\beta$
Z_1	0.0003	0.0181	0.0007	0.0240
Z_2	-0.1518	0.2380	-0.2329	0.3324
Z_3	0.0045	0.0514	0.0400	0.0668

Still, to illustrate the possibilities of the method, we also provide the scaled Schoenfeld residuals for the three Z variables when $x = 0$ (we obtained qualitatively similar picture for the case $x = 3.34$). Here, as seen in Figure 8, there is no pattern confirming that the Cox model seems to be reasonable to fit our conditional survival functions for the 3 components.

Also, even if in this example the β do not seem to be really significant, still it is useful to look at the basis survivor function $S_0(y|x)$ and at the conditional survivor functions $S(y|x, z)$ for selected values of x and of z. Figures 9 and 11 display the results for $x = 0$ and x, the median of the output X, respectively. In each figures the values of z are chosen from small to large: from the minimum of each components and the 3 quartiles of each components (see Table 6 for these 4 values). The pictures indicate a global (jointly) unfavorable, negative effect of the Z variables on the efficiency distribution, because as seen in Table 5, the unfavorable effect of Z_2 dominates. We see also that the two figures are qualitatively similar for the two selected values of x.

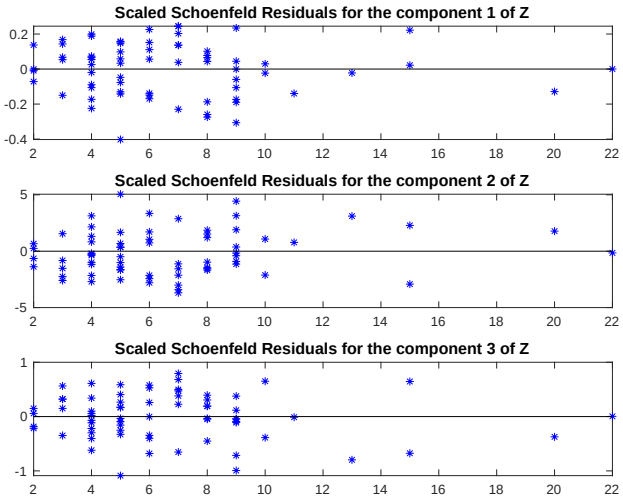

Figure 8: Scaled Schoenfeld residuals for the three components of the Z variables when $x = 0$ as a function of Y_i.

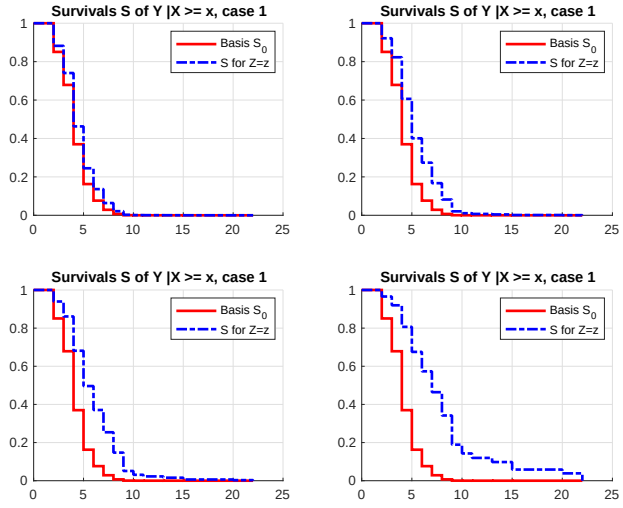

Figure 9: Estimates of the survival functions $S_0(y|x)$ and $S(y|x, z)$, when $x = 0$ (case 1), for 4 selected values of z. From left to right and top to bottom panels, the minimum and the 3 quartiles, simultaneously for the 3 components of Z (see Table 6).

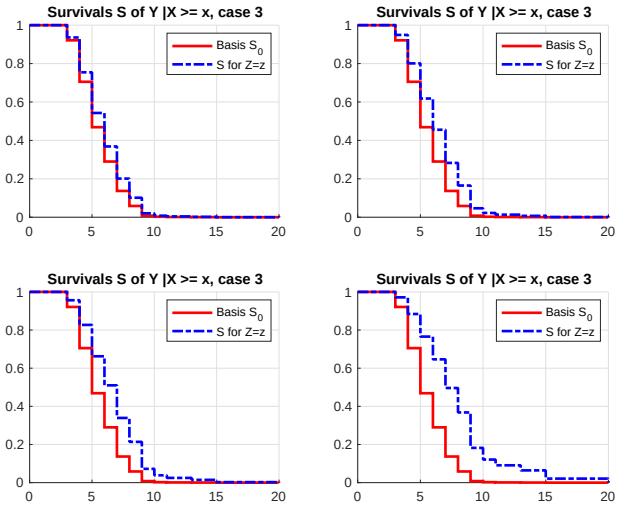

Figure 10: Estimates of the survival functions $S_0(y|x)$ and $S(y|x,z)$, when $x = 3.34$ (case 3, the median of X), for 4 selected values of z. From left to right and top to bottom panels, the minimum and the 3 quartiles, simultaneously for the 3 components of Z.

Here too, we provide estimates of the conditional order-m frontiers $\varphi_m(x,z)$ and confidence intervals at any selected value of (x,z). For the points estimates, Table 6 displays the estimates for $x = 0$ and $x = 3.34$ at the 4 selected values of z described in Figures 9 and 11. We fix $m = 10$ which left around 5% of the schools below the order-m input frontier. We see the global effect of Z on the order-m frontiers which is switched to the right when z increases.

Table 6: Estimation of conditional order-m frontiers $\varphi_m(x,z)$ for 2 values of x and 4 values of z, in the School data set. Here $m = 10$.

	$x = 0, n_x = 68$	$x = 3.34, n_x = 34$				
$\varphi(x)$	2.0000	3.0000				
$\varphi_m(x,z_1)$	2.3360	3.5774	$z_1 =$	3.2400	1.8500	5.4600
$\varphi_m(x,z_2)$	2.5923	3.7106	$z_2 =$	12.2700	5.0950	18.5700
$\varphi_m(x,z_3)$	2.7829	3.8068	$z_3 =$	25.1900	7.1100	26.6700
$\varphi_m(x,z_4)$	3.2789	4.1235	$z_4 =$	37.4750	11.3700	40.8850

Note that, just for illustrative purpose, at $x = 0$ and $z = z_3$, the bootstrap provides a 95% confidence interval for $\varphi_m(x,z_3)$ given by $[1.9068, 2.9610]$ (2000 bootstrap loops and basic bootstrap method).

Fianlly, since the original study in CRR-paper was focused on the comparisons between PFT and NFT schools, we complete our illustration by selecting this qualitative factor as

an environmental factor. First the test of separability (200 random splits, 20 shuffles for jackknife bias corrections and 200 bootstrap replications) provides high p-values of the order 0.60, so no reason to reject the separability assumptions. Applying then the PICF model with the qualitative factor ($Z = 1$ for PFT schools and $Z = 0$ for NFT schools) we obtain the following results for $x = 0$; $\widehat{\beta}(0) = 0.5503$ with $\widehat{\sigma}_\beta = 0.2849$ and a p-value slightly above 0.05: so not very significant but still a positive value of β indicating a possible favorable effect of being a PFT school. Again, more data would be required to get more discriminant results. We computed the values of $\widehat{\beta}(x)$ far higher values of x with similar magnitude but even less significant, since the number of used observations decreases when x increases. We display below, for $x = 0$ the plots of the basis survivor function and the conditional to $Z = 1$ (PFT schools) one. In the case here the basis survivor is also the conditional to $Z = 0$ survivor function. We see clearly the favorable effect of being in the PFT group since the conditional survivor is shifted to the left.

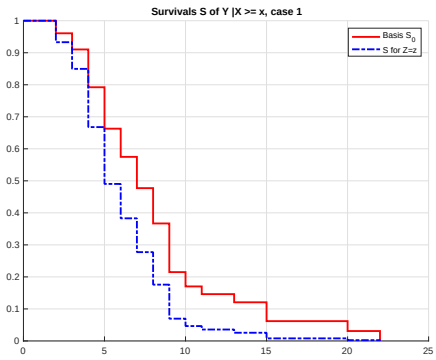

Figure 11: Estimates of the survival functions $S_0(y|x)$ and $S(y|x, z)$, when $x = 0$ (case 1) for the PFT schools ($z = 1$). The basis survivor is also the conditional to $Z = 0$ (NFT schools) survivor function.

As a conclusion, it is clear that in this illustration, no definite conclusions can be drawn due to the limited sample size, except the almost significant role of the effect of being a PFT or a NPT school. But the exercise shows the flexibility of the method and the possibilities of various analysis. Note that the effect of the family descriptors in the CCR data seems to play a very weak role on the efficiency distributions but not on the shape of the frontier (the separability condition is not rejected). So it is certainly better to use these descriptors as environmental variables rather than as "inputs", as in the original study of Charnes et al. (1981). In particular the variable Z_2, the highest occupation of a family member, does not seem, at first glance, to be favorable to efficiency.

4.2.3 The data from GRDF

The next data set comes from the company "Gaz Réseau de Distribution de France" (GRDF): we have information on 1241 distribution units for the year 2019. In this illustrative example we will consider a cost variable Y (the annual cost of the piping network) and two outputs X_1, X_2 being the gaz consumption by the customers (in kWH) and the length of the piping network (in kms). As continuous environmental variables, we have Z_1 the minimal yearly temperature in the area covered by the unit and Z_2 the surface of the area covered by the unit (in km^2). The variables Y, X and Z_2 behaves like log-normal variables so we transform these variables in logs. We have also for illustration a qualitative variable that is the region of France in which the units belong (they are 6 regions). We will show below how to handle this kind of variable.

There are also some outliers in the data for Y and X (even in logs) so we use the robust technique to eliminate these points that might create numerical instability (we used the boxplots approach as described, e.g. in Section 1.1 of Härdle and Simar, 2019). This reduces the sample size to 1186 units. Preliminary tests of separability indicate that both Z_1 and Z_2 does not influence in a significant way the frontier level (p-values between 0.66 and 1.00, according the chosen test statistics). So we implement our PICF model to investigate the effect of these variables on the cost efficiency. We use again the traditional $\exp(\beta'(x)Z)$ as proportionality factor.

We computed the estimation of $\beta(x)$ at 4 values for x: the minimum of the observed X_i for each component (Case 1), and at the quartiles of the observed X_i for each component (Case 2 to 4). To save space we present only here the value for Case 1 (the min) and Case 3 (the median), corresponding to a median size unit (the two outputs are highly correlated). The results are summarized in Table 7.

Table 7: Estimation of β for 2 values of x (min and median) in the GRDF data.

$(x_1, x_{2)}$	$(2.2809, 0.8007)$	$(5.9134, 3.7209)$
n_x	1186	526
$\widehat{\beta}_1$	0.0379	0.0176
$\widehat{\sigma}_{\beta_1}$	0.0193	0.0289
p-value	0.0490	0.5424
$\widehat{\beta}_2$	-1.3501	-1.4471
$\widehat{\sigma}_{\beta_2}$	0.0411	0.0755
p-value	0.0000	0.0000

We see clearly that the variable Z_2, the surface of the area has a negative effect for the

efficiency distribution whch is highly significant and seems rather stable across the 2 values of x considered. The minimal temperature (Z_1) has a slightly significant positive effect on efficiency, for small units only but even in this case, the effect is quite negligible compared to the effect of Z_1 which dominates.

As explained above, the literature on Cox model offers a battery of graphical diagnostic to check the validity of the model. We first may look to the ProbPlot of the Cox-Snell residuals versus an Exp(1). To save space we present only in Figure 12 the case for x being at its minimal level (that would be sufficient if we assume that $\beta(x) = \beta$, as in most of the traditional two-stage approaches). We see a reasonable plot with only some discrepancy with the Exp(1) for high values. In a real analysis, the practitioner would identify these few points and try to understand the reason of these isolated discrepancies.

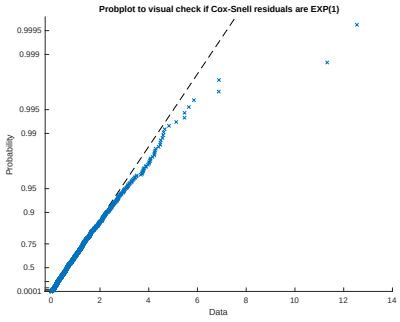

Figure 12: ProbPlot of the Cox-Snell residuals when x is the $\min(X_i)$ versus an Exp(1), GRDF data.

We can also have a look on the Schoenfeld residuals for both components Z_1 and Z_2 to check the proportional hazard assumption. Figure 13 does not indicate any clear pattern, so the PICF model seems to ba reasonable to fit our data.

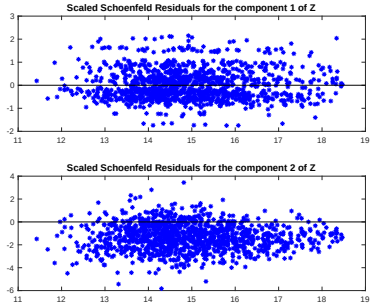

Figure 13: Scaled Schoenfeld residuals for the two variables Z variable when x is the $\min(X_i)$, as a function of the cost Y_i (horizontal axis), for the GRDF data.

We did the same plots for larger x equal to the median of each components and the pictures are qualitatively quite similar.

To appreciate the effect of Z on the conditional efficiency, Figures 14 and 15 display the basis survivor function and the conditional to $Z = z$ survivor functions, for two values of x, as above (Case 1) the minimum of the X and the median value of X (Case 3) respectively. In each Figure we display 4 different values for z. Since the effect of Z_1 is weakly significant and negligible compared to the effect of Z_2, we fix $z_1 = \text{median}(Z_1)$ and for z_2, we select the minimum and the 3 quartiles of Z_2.

We see clearly the shift to the right of the conditional survivor function when Z_2 increases, indicating larger probabilities to be far from the frontier. We observe also a change of the basis survivor function which becomes steeper descendent for x larger, indicating more concentrated distribution of the cost near the boundary for these values of $X \geq x$.

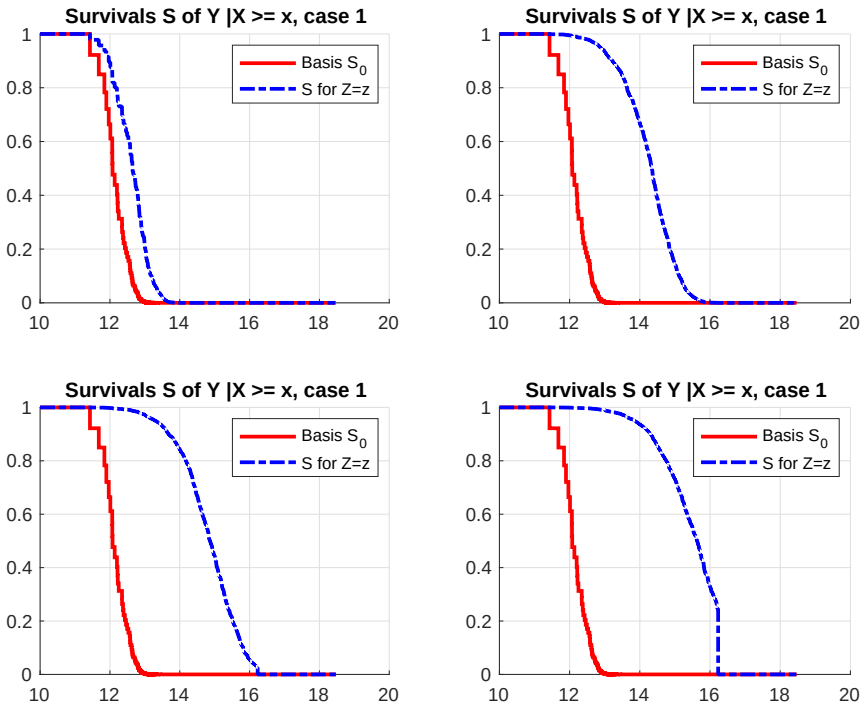

Figure 14: Estimates of the survival functions $S_0(y|x)$ and $S(y|x, z)$, when (case 1), x is the minimum of the 2 components of X. For 4 selected values of Z, we have z_1 which is fixed at its median and from left to right and top to bottom panels, z_2 is the minimum and the 3 quartiles of Z_2.

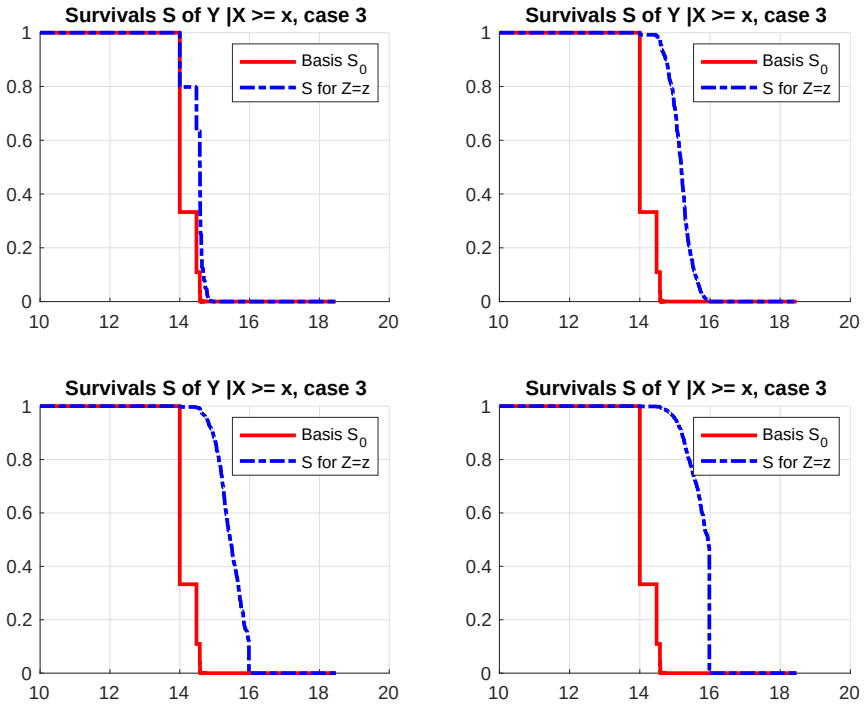

Figure 15: Estimates of the survival functions $S_0(y|x)$ and $S(y|x,z)$, when (case 3), x is the median of the 2 components of X. For 4 selected values of Z, we have z_1 which is fixed at its median and from left to right and top to bottom panels, z_2 is the minimum and the 3 quartiles of Z_2.

Of course of particular interest for the practitioner are the conditional frontiers function $\varphi_m(x,z)$ that can be computed at any point (x,z), allowing to derive e.g. the order-m conditional cost efficiency for a unit operating at the level (x,y) and facing the environmental conditions z. We fix the value of $m = 350$ which left around 5% of the data points outside the order-m cost frontier. The values have been computed here for the two selected values of x and the 4 values of z described in the Figures 14 and 15. The results are shown in Table 8, showing, as expected that the robust frontiers move to the right when Z_2, the surface of the area covered by the units increase.

Table 8: Estimation of conditional order-m frontiers $\varphi_m(x, z)$ for 2 values of x and 4 values of z, in the School data set. Here $m = 350$.

	$x = \min(X), n_x = 1186$	$x = \text{median}(X), n_x = 526$			
$\varphi(x)$	11.4275	14.0003			
$\varphi_m(x, z_1)$	11.4277	14.0003	$z_1 =$	7.8857	1.1990
$\varphi_m(x, z_2)$	11.7525	14.0316	$z_2 =$	7.8857	3.5127
$\varphi_m(x, z_3)$	11.9604	14.1708	$z_3 =$	7.8857	4.1454
$\varphi_m(x, z_4)$	12.2639	14.3759	$z_4 =$	7.8857	4.8575

We can also provide confidence intervals for the values of $\varphi_m(x, z)$ and of course for the values of the derived efficiency measures. Just for illustration we did it for the case where $x = \min(X)$ and $z = z_3$ in Table 8. We obtain by bootstrap (1000 bootstrap loops and basic bootstrap method) the 95% confidence interval $[11.6214, 12.0597]$ for $\varphi_m(x_{\min}, z_3)$, which is rather narrow due to the large number of observation used for this value of x.

Finally, to illustrate the flexibility of the method, we analyze the effect of the regions on cost efficiency, if any, by considering in our PICF model the conditional survival function, conditional to a qualitative variable. Since there are 6 regions we need to build a design matrix Z with n rows and 5 columns to identify the effect of the 6 regions. We adopt here the approach described, e.g., in Section 8.1 of Härdle and Simar (2019). For instance we would have as row of Z for the region j, $j = 1, \ldots, 5$, the jth row of the identity matrix I_5, whereas for the region $j = 6$, the row of Z is $-i_5$, i.e. minus a vector of ones of length 5. Then we may use the proportionality factor as above, $\exp(\beta' Z_i)$ with $\beta \in \mathbb{R}^5$. Then the effect of the first 5 regions are directly given by $(\beta_1, \ldots, \beta_5)$ and the effect of the 6th region is given by $\beta_6 = -\sum_{j=1}^{5} \beta_j$ (see details in Härdle and Simar, 2019). In our case here the exercise gives the values for $x = x_{\min}$, $\widehat{\beta}(x) = (0.0425, -0.0234, 0.0213, 0.0885, -0.0687, -0.0177)$, but none of these values are significantly different from zero, all the standard errors lie between 0.11 and 0.48 so that all the p-values are above 0.50. So we do not develop this analysis further, but it indicates how easy it is to handle with qualitative variables.

5 Conclusions

We have introduced the PICF model for analyzing cost efficiency and in particular the effect of environmental variables on the production process. This can be viewed as a semiparametric alternative to the semiparametric popular two-stage methods in this literature. As for these two-stage production models, the "separability" assumption is implicit in the PICF model. The methodology avoids to estimate in a first stage the nonparametric frontier and

due to that, it avoids the drawbacks of the semiparametric two-stage approaches: poor rates of convergence like n^{κ}, where κ is the rate of the nonparametric estimator of the frontier ($\kappa = 1/(p+1)$ for the FDH case) and bias correction problem in the frontier estimates.

We keep the $\sqrt{n}$-rates of convergence for the parameters of interest β (that may be x-dependent, i.e. $\beta = \beta(x)$) and asymptotic normality which makes inference quite easy. We provide also a semiparametric estimator of the order-m conditional frontier, with $\sqrt{n}$ rate of convergence and asymptotic normality (compared to the rate $n^{2/(d+4)}$ obtained for traditional nonparametric estimators, but of course at a cost of our semiparametric flexible assumption). Here also the bootstrap is easy to implement and provides the estimation of the variance of the estimators, when needed.

These various examples have shown the flexibility and the power of the tool with the battery of existing simple graphical diagnostic checks for the validity of the Cox proportional hazard model.

Straightforward extension is the use of Lasso methods for variables selection as described in Tibshirani (1997). Future work would include testing methods for testing if $\beta(x) = \beta$ and more generally, adapting the existing literature on Cox model, develop a formal test for the validity of the PICF model.

A Appendix: Simulation of data

Simulating a data set $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ should be done with care. Usually researchers specify a model for the frontier function $\varphi_0(x)$ then select a model to simulate values of X_i and Z_i and finally generate Y_i for $X = X_i$ and $Z = Z_i$. Here we have to generate the sample according to our PICF model which specifies that the survival function $S(y|X \geq x, Z = z) = [S_0(y|X \geq x)]^{a(z,\beta(x))}$ for some basic survival function $S_0(y|X \geq x)$ and some given functions $a(\cdot,\cdot)$ and $\beta(\cdot)$. So we need to recover from our model the conditional distribution of Y given $X = x$ and $Z = z$ derived from the PICF. We will denote $\widetilde{S}(y|X = x, Z = z)$ this conditional survival function where we use the $\widetilde{S}$ notation when we condition to $X = x$, to distinguish form $S(y|X \geq x, Z = z)$ defined above, where we condition on $X \geq x$. Consider for instance the corresponding quantile function

$$y = Q(u, x, z) = \widetilde{S}^{-1}(y|X = x, Z = z), \tag{A.1}$$

where Q is monotone decreasing with u. We know that $U = Q^{-1}(Y, x, z)$ is uniform on $[0, 1]$ and independent of X and Z, so an easy way to simulate Y given $X = x$ and $Z = z$, is to generate U_i as uniform on $[0, 1]$ and then define $Y_i = Q(U_i, X_i, Z_i)$.

The general form of $Q(u, x, z)$ can be obtained as follows in order to satisfy the PICF

model. Some simple algebra leads to the equation

$$\text{Prob}(Y \geq y \mid X \geq x, Z = z) = \frac{\int_x^\infty Q^{-1}(y,t,z) f_X(t|z) dt}{S_X(x|z)}. \tag{A.2}$$

So, for the PICF model, the function Q must satisfy

$$\int_x^\infty Q^{-1}(y,t,z) f_X(t|z) dt = S_X(x|z) \left[S_0(y|X \geq x)\right]^{a(z,\beta(x))}. \tag{A.3}$$

Taking the derivative with respect to x (with some abuse of notations below, for $x \in \mathbb{R}^p$ the derivative ∂_x^p has to be understood as $\partial^p/(\partial x_1 \dots \partial x_p)$) and equating both sides we obtain

$$-Q^{-1}(y,x,z) f_X(x|z) = -f_X(x|z) \left[S_0(y|X \geq x)\right]^{a(z,\beta(x))} + S_X(x|z) \partial_x^p \left\{ \left[S_0(y|X \geq x)\right]^{a(z,\beta(x)} \right\}. \tag{A.4}$$

After some tedious but simple mathematical developments, this leads to the equation

$$\begin{aligned} Q^{-1}(y,x,z) &= \widetilde{S}(y|X = x, Z = z) \\ &= \left[S_0(y|X \geq x)\right]^{a(z,\beta(x))} \times \\ &\left\{ 1 - \frac{S_X(x|z)}{f_X(x|z)} \left[\log S_0(y|X \geq x) \partial_x^p a(z,\beta(x)) + a(z,\beta(x)) \partial_x^p \log S_0(y|X \geq x) \right] \right\}. \end{aligned} \tag{A.5}$$

which allows to define (at least numerically) its reciprocal $Q(u,x,z)$ for any (u,x,z). The expression in greatly simplified if we introduce additional assumption in the model we want to simulate.

Indeed, if we assume that the joint conditional survival function satisfies the Cox model, i.e. $S_{XY}(x,y|z) = \left[S_0(x,y)\right]^{a(z,\beta(x))}$ where $S_0(x,y) = S_0(y|X \geq x) S_0(x)$, we have

$$S_0(y|X \geq x) = \frac{\int_x^\infty \widetilde{S}_0(y|t) f_0(t) dt}{S_0(x)}, \tag{A.6}$$

where again $\widetilde{S}_0(y|x)$ is $\widetilde{S}_0(y|X = x)$ is the correspondent of the baseline survivor $S_0(y|X \geq x)$ when conditioning on $X = x$. We also have $S_X(x|z) = (S_0(x))^{a(z,\beta)}$. Therefore, the equation (A.3) simplifies into

$$\int_x^\infty Q^{-1}(y,t,z) f_X(t|z) dt = \left[\int_x^\infty \widetilde{S}_0(y|t) f_0(t) dt \right]^{a(z,\beta(x))}, \tag{A.7}$$

In addition if we assume that $\beta(x) = \beta$, the derivative of both sides of (A.7) with respect to x simplifies. Note also that $f_X(x|z) = a(z,\beta) f_0(x) (S_0(x))^{a(z,\beta)-1}$. After some simplifications this leads to the equation

$$\begin{aligned} Q^{-1}(y,x,z) =& \widetilde{S}(y|X = x, Z = z) \\ &= \left[S_0(y|X \geq x)\right]^{a(z,\beta)-1} \widetilde{S}_0(y|X = x). \end{aligned} \tag{A.8}$$

So given the function $a(z,\beta)$, the survival $\widetilde{S}_0(y|X=x)$ and the baseline density of X, $f_0(x)$, we can compute $S_0(y|X \geq x)$ by (A.6), and then the conditional survival $\widetilde{S}(y|X=x,Z=z)$. By inverting (A.8), we have the quantile function $y = Q(u,x,z)$ for any u (at least numerically) and then we can simulate a value Y_i, for a given (X_i,Z_i) according to the PICF model.

References

[1] Aragon, Y., A. Daouia and C. Thomas-Agnan (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] Cazals, C., Fève F., Florens, J.P. and L. Simar (2016), Nonparametric Instrumental Variables estimation for efficiency frontier, *Journal of Econometrics*, 190, 349–359.

[4] Cazals, C. Florens, J.P. and L. Simar (2002), Nonparametric Frontier Estimation: a Robust Approach , *Journal of Econometrics*, 106, 125.

[5] Charnes, A., Cooper, W.W. and E. Rhodes (1981), Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through, *Management Science* 27, 668–697.

[6] Cox DR (1972), Regression Models and Life Tables, *JRSS*, B34, 187-220

[7] Daouia, A., Florens, J.P. and L. Simar (2010), Frontier estimation and extreme values theory. *Bernoulli*, 16(4), 1039–1063.

[8] Daouia, A., Florens, J.P. and L. Simar (2012), Regularization of Non-parametric Frontier Estimators, *Journal of Econometrics*, 168, 285–299.

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

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

[11] 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.

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

[13] 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.

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

[15] Grambsch, P.M. and T.M. Therneau (1994), Proportional Hazards Tests and Diagnostics Based on Weighted Residuals, *Biometrika*, vol 81, 3, 515–526.

[16] Härdle, W.K. and L. Simar (2019), *Applied Multivariate Statistical Analysis*, Fifth Edition, Springer Nature, Switzerland.

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

[18] Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John. Wiley & Sons.

[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] Mammen, E. (1992), *When does bootstrap work? Asymptotic results and simulations*. Springer-Verlag, Berlin.

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

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

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

[25] Simar, L and P. Wilson (2007), Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes, *Journal of Econometrics*, vol 136, 1, 3164.

[26] Simar, L. and P.W. Wilson (2011), Two-Stage DEA: *Caveat Emptor*. *Journal of Productivity Analysis*, 36, 205–218.

[27] Simar, L. and P.W. Wilson (2020), Hypothesis Testing in Nonparametric Models of Production using Multiple Sample Splits. *Journal of Productivity Analysis*, 53, 287–303.

[28] Tibshirani, R. (1997), The Lasso Method for Variable Selection in the Cox Model, *Statistics in Medicine*, vol 16, 385–395.

[29] Tsiatis, A.A. (1981), A Large Sample Study of Cox's Regression Model, *The Annals of Statistics*, 9(1), 93–108.

[30] Wilson, P. W. (1993), Detecting outliers in deterministic nonparametric frontier models with multiple outputs, *Journal of Business and Economic Statistics* 11, 319–323.