TECHNICAL
REPORT

11032

Frontier estimation in nonparametric location-scale models

FLORENS, J.P., SIMAR, L. and I. VAN KEILEGOM

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Frontier estimation in nonparametric location-scale models

Jean-Pierre FLORENS [*]
Toulouse School of Economics

Léopold SIMAR [§]
Université catholique de Louvain

Ingrid VAN KEILEGOM [§,**]
Université catholique de Louvain

October 25, 2011

Abstract

Conditional efficiency measures are very natural tools to capture the efficiency of firms facing heterogeneous environmental conditions. They are defined by the distance of a unit to the support of a conditional distribution, conditional to the level of these external factors. The traditional approach is to estimate nonparametrically this distribution and this requires the use of appropriate smoothing techniques. In this paper, we consider an alternative approach to estimate its support. We first assume flexible nonparametric location-scale models linking the inputs and the outputs to the environmental factors to eliminate in the inputs/outputs the dependence on Z. Then we use these "pre-whitened" inputs and outputs to define the optimal frontier function. This provides a "pure" measure of efficiency more reliable to produce rankings or benchmarks of units among themselves, since the influence of external factors has been eliminated. We estimate both the full frontier and its more robust version, the order-m frontier. The asymptotic properties are established. We can also recover the frontiers in the original inputs/outputs space and we give their asymptotic properties. The approach is illustrated with some selected simulated data but also with a real dataset from the bank industry.

Key Words: Nonparametric Frontiers, Efficiency Analysis, Robust Estimation of Frontiers, Conditional Efficiency.

[*] This research was supported by the French "Agence Nationale pour la Recherche" under grant ANR-08-BLAN-0106-01/EPI project.

[§] Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. Research supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy), and by the INRA-GREMAQ, Toulouse.

[**] Research also supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

1 Introduction

In production theory and efficiency analysis, we analyze how firms transform their inputs (factors of production) to produce a set of outputs. The efficient production frontier is then defined in the input-output space as the locus of the maximal attainable level of outputs, given the level of the inputs. In other setups, we are rather willing to estimate an input (or cost) frontier that is defined as the minimal attainable level of the inputs for producing a given level of outputs. In both cases the problem can be viewed as estimating a surface under shape constraints induced by the underlying economic model (monotonicity, ...). The efficiency score of a given production unit is then determined by an appropriate distance (in the output direction, or in the input direction) of this unit to the optimal frontier. Farrell–Debreu radial distances (Debreu, 1951, Farrell, 1957) are often used in this perspective. For the empirical researcher, the attainable set of inputs and outputs is not known, neither its production frontier and the derived efficiency scores. The econometric literature has proposed many ways for producing estimators based on a sample of observed units. Nonparametric estimators are particularly attractive because they do not rely on restrictive parametric hypotheses on the process that generates the data. This includes assumptions on the shape of the attainable set and on the distribution of inputs and outputs in the attainable set (see Simar and Wilson, 2008, for a recent survey).

During the last decades, the efficiency literature has become more concerned with connecting the efficiency measures to environmental factors that cannot be controlled by the producers, but might influence the production process. In this paper, we address this latter issue and we propose an original way to complement previous approaches of the literature. Our presentation is for the input orientation case, where we want to estimate the minimal input (cost) frontier.[1] Formally, let $Y \in \mathbb{R}_+$ denote the input (or the cost of production), $X \in \mathbb{R}_+^{d_x}$ be the vector of goods or services produced, and we will denote by $Z \in \mathbb{R}^{d_z}$ the set of environmental factors.

A traditional approach in the efficiency literature to investigate the effect of these environmental factors on the efficiencies, is a two-stage procedure. In this approach, the efficiency scores are nonparametrically estimated in a first stage, in the input-output space and then, in a second stage, the estimated efficiency scores are regressed, by some appropriate model (mainly parametric models) on the environmental variables (see Simar

[1] The presentation for the output oriented case, where we want to estimate the maximal production frontier, is a straightforward adaptation of what is done here.

and Wilson, 2007, 2011, and the dozens of references quoted there). However, as pointed out by Simar and Wilson, these two-stage approaches are restricted to situations where these factors do not influence the shape of the production set, but can only affect the probability of being more or less efficient. As demonstrated in Simar and Wilson (2011) in some very simple examples, if this is not the case (i.e., if the environmental conditions may affect the attainable set), the first stage efficiency estimates in the input-output space have no economic meaning. If the two-stage approach is validated, by some appropriate test (see e.g. Daraio et al., 2010), one can indeed regress the first stage efficiencies on the environmental factors. However, usual inference on the regression coefficients is not available in this framework (because the first stage efficiency estimators are biased and correlated) and if used, this may lead to wrong inference. Bootstrap based procedures may help to solve the problem (see Simar and Wilson, 2007, for details).

A more general and appealing approach is to consider the probabilistic formulation of the production process proposed by Cazals et al. (2002). Here the production set is the support of some probability measure in the input-output space and the traditional Debreu–Farrell efficiency scores can be defined in terms of some nonstandard conditional distribution function. This approach also allows to define the concepts of partial order frontiers (order-m or order-α quantile), which are less extreme than the boundary of the support and allow to determine frontier and efficiency estimators that are less sensitive to extreme or outlying data points. See Cazals et al. (2002), Aragon et al. (2005), Daouia and Simar (2007), and Daraio and Simar (2007) for an overview of these approaches. The probabilistic formulation of the production process allows also to accommodate quite naturally the model to the presence of environmental factors. This leads to efficient conditional frontiers and to conditional Debreu–Farrell efficiency scores.

Nonparametric estimators of the conditional frontiers and efficiency scores can easily be derived (see Cazals et al., 2002, Daraio and Simar, 2005, and Daouia and Simar, 2007) and their asymptotic properties have been established (see Cazals et al., 2002, Daouia and Simar, 2007, and Jeong et al., 2010). These estimators are based on nonparametric estimators of conditional distribution functions (or conditional survival functions), where the conditioning is on the environmental factors Z. This requires smoothing techniques for the environmental variables including selection of smoothing parameters (bandwidths). Data-driven procedures have been proposed in this setup by Bădin et al. (2010), providing optimal bandwidths. However, the resulting estimators have rates of convergence deteriorated by the dimension of Z (known as the "curse of dimensionality").

The approach developed in this paper contributes to the literature on conditional frontiers and efficiency scores, but avoids (or at least reduces the impact of) this curse of dimensionality. This will be obtained by assuming flexible nonparametric location-scale models linking the input and the outputs to the environmental factors. In a sense, our approach could be seen as a two-stage method, but the other way around. First we eliminate in the input and the outputs the dependence on Z by means of nonparametric location-scale models and then, in a second step, we estimate the frontier and the efficiencies of the units using "pure" or "pre-whitened" inputs and outputs, whitened from the influence of Z. This allows to define a "pure" measure of managerial efficiency, pure in the sense that all the external influence of the Z-factors has been eliminated. This measure of pure efficiency is certainly more reliable to produce rankings or benchmarks of units among themselves, since the influence of external factors has been eliminated. We will see that the resulting estimators will be free of the curse of dimensionality due to the dimension of Z, which is another advantage. We can estimate both the full frontier and their more robust versions. In this paper we will only focus on the order-m frontiers, but the extension to order-α quantile frontiers is immediate. The asymptotic properties of the estimators will be established. We will also be able to recover estimators of the full and of the order-m conditional frontier in the original input-output space and we give their asymptotic properties.

The paper is organized as follows. In the next section we summarize and recall some basic concepts and notations. Then, in Section 3 we introduce our nonparametric location-scale regression model. An estimator of the frontier under this model is proposed in Section 4, and its asymptotic properties are presented. Section 5 comments on some practical aspects of our procedure and suggests some variants of the model. In Section 6 we illustrate how the procedure works in practice through some simulated data and through a real data example on bank efficiencies previously discussed in the literature. Finally, Section 7 states some general conclusions.

2 Basic Notations

It is useful to consider the production process as a process generating the random variables (X, Y, Z) on an appropriate probability space. Cazals et al. (2002) and Daraio and Simar (2005) consider a probability model, where the conditional distribution of (X, Y) given a particular value of Z will be of particular interest. This conditional process can be described by the conditional survival function

$$S_{X,Y|Z}(x,y|z) = P(X \geq x, Y \geq y|Z = z) = S_{Y|X,Z}(y|x,z)S_{X|Z}(x|z), \tag{1}$$

where $S_{Y|X,Z}(y|x,z) = P(Y \geq y \,|\, X \geq x, Z = z)$ and $S_{X|Z}(x|z) = P(X \geq x|Z = z)$. It should be noticed that the only difference between the mathematical treatment of X and Z in $S_{Y|X,Z}$, is that we condition on $Z = z$ for the environmental factors but on $X \geq x$ for the outputs.

The conditional minimum input frontier is then defined as the minimal achievable input level for units producing at least the level x of outputs, but facing the environmental conditions z. This defines the conditional frontier

$$\tau(x,z) = \inf\{y : S_{Y|X,Z}(y|x,z) < 1\}. \tag{2}$$

Cazals et al. (2002) introduce also the order-m conditional frontier as less extreme frontier for benchmarking the different units. For a given integer $m \geq 1$, it is defined as

$$\tau_m(x,z) = E(\min(Y_1,\ldots,Y_m) \,|\, X \geq x, Z = z) = \int_0^\infty S^m_{Y|X,Z}(y|x,z)\,dy. \tag{3}$$

Of course, as $m \to \infty$, $\tau_m(x,z) \to \tau(x,z)$. Nonparametric estimators of the frontier functions are obtained by plugging in a nonparametric estimator of the conditional survival function. This requires some smoothing relative to the Z variables and provides

$$\widehat{S}_{Y|X,Z}(y \,|\, x,z) = \frac{\sum_{i=1}^n I(X_i \geq x, Y_i \geq y) \prod_{j=1}^{d_z} k((z_j - Z_{ij})/h_j)}{\sum_{i=1}^n I(X_i \geq x) \prod_{j=1}^{d_z} k((z_j - Z_{ij})/h_j)}, \tag{4}$$

where for a vector a, a_j denotes its j^{th} component, and we choose here a product kernel for Z, with each $k(\cdot)$ being a univariate kernel with compact support and $h_j > 0$, $j = 1,\ldots,d_z$ being the bandwidths.

Of course, when no environmental factors Z are considered, the main object of interest is the survival function $S_{Y|X}(y|x) = P(Y \geq y \,|\, X \geq x)$ providing the unconditional frontiers $\psi(x)$ and $\psi_m(x)$ when replacing $S_{Y|X,Z}(y|x,z)$ by $S_{Y|X}(y|x)$ in the expressions (2) and (3) above. In this case, nonparametric estimators of the frontiers are obtained by plugging in the empirical conditional survival function (no smoothing is required).

The asymptotic properties of these estimators are established, see Cazals et al. (2002) for the order-m frontiers and Park et al. (2000), Daouia et al. (2010) and Jeong et al. (2010) for the full frontier case. For the conditional frontier estimator, the rates of convergence are deteriorated by the smoothing in Z to get the estimator $\widehat{S}_{Y|X,Z}$, in the sense that n is to be replaced by $n\prod_{j=1}^{d_z} h_j$ when product kernels are used for smoothing the d_z components of Z, which leads to rates of order $n^{4/((d_z+4)(d_x+1))}$ for full frontiers and $n^{2/(d_z+4)}$ for the order-m frontiers (see Jeong et al., 2010 for details). In the next section we propose to model the links between (X, Y) and Z by flexible nonparametric location-scale models, and we will derive estimators of the conditional frontiers, that suffer less from the "curse of dimensionality" problem than the traditional nonparametric estimators based on $\widehat{S}_{Y|X,Z}$. We will also propose so-called "whitened" or "pure" frontier estimators, which will have parametric rate of convergence, and which will be free of the influence of Z.

3 The Location-Scale Model

Suppose the vector (X, Y, Z) follows the following location-scale regression model:

$$\begin{cases} X_j = \mu_{1j}(Z) + \sigma_{1j}(Z)\varepsilon_{1j} \qquad (j = 1, \ldots, d_x) \\ Y = \mu_2(Z) + \sigma_2(Z)\varepsilon_2, \end{cases} \tag{5}$$

where we assume that $(\varepsilon_1, \varepsilon_2)$ is independent of Z, and where $\varepsilon_1 = (\varepsilon_{11}, \ldots, \varepsilon_{1d_x})^t$, $\mu_{1j}(Z) = E(X_j|Z)$, $\mu_2(Z) = E(Y|Z)$, $\sigma_{1j}^2(Z) = \text{Var}(X_j|Z)$ and $\sigma_2^2(Z) = \text{Var}(Y|Z)$. Also, denote $\mu_1(Z) = (\mu_{11}(Z), \ldots, \mu_{1d_x}(Z))^t$ and $\sigma_1^2(Z) = (\sigma_{11}^2(Z), \ldots, \sigma_{1d_x}^2(Z))^t$. Our goal is to estimate the conditional full frontier $\tau(x, z)$ and the conditional order-m frontier $\tau_m(x, z)$, under the above model (5). We will see that this can be done without estimating directly the conditional survival function $S_{Y|X,Z}(y \,|\, x, z)$. Indeed, note that

$$\begin{aligned} \tau(x,z) &= \inf\Big\{y : P\Big(\frac{Y-\mu_2(Z)}{\sigma_2(Z)} \geq \frac{y-\mu_2(z)}{\sigma_2(z)}\Big| X \geq x, Z = z\Big) < 1\Big\} \\ &= \mu_2(z) + \inf\Big\{t_2 : P\Big(\varepsilon_2 \geq t_2 \Big| \varepsilon_1 \geq \frac{x-\mu_1(z)}{\sigma_1(z)}\Big) < 1\Big\}\sigma_2(z) \\ &= \mu_2(z) + \varphi\Big(\frac{x-\mu_1(z)}{\sigma_1(z)}\Big)\sigma_2(z), \end{aligned} \tag{6}$$

where

$$\varphi(t_1) = \inf\{t_2 : S_{\varepsilon_2|\varepsilon_1}(t_2|t_1) < 1\}$$

and $S_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = P(\varepsilon_2 \geq t_2|\varepsilon_1 \geq t_1)$. Hence, under model (5) the frontier $\tau(x, z)$ can be derived from appropriate estimators of the functions $\mu_j(z)$ and $\sigma_j(z)$ $(j = 1, 2)$ (which require only smoothing in z in the *center* of the data cloud), and an estimator of the survival function $S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)$, which can be estimated at parametric rate. The fact that we avoid smoothing at the frontier is important, as typically the data can be rather sparse there and estimators are sensitive to outliers.

In a similar way we can show the following identity for the conditional order-m frontier:

$$\tau_m(x,z) = \mu_2(z) + \varphi_m\Big(\frac{x-\mu_1(z)}{\sigma_1(z)}\Big)\sigma_2(z),$$

where

$$\varphi_m(t_1) = E\Big(\min(\varepsilon_{21},\ldots,\varepsilon_{2m})\Big|\varepsilon_1 \geq t_1\Big) \tag{7}$$

is the order-m frontier of ε_2 given that $\varepsilon_1 \geq t_1$, and where $\varepsilon_{21},\ldots,\varepsilon_{2m}$ are i.i.d. copies of ε_2.

Interpretation

To appreciate the flexibility of our model (5), consider the particular case where Z would be independent of the input Y and of all the outputs X. In such a case, all the functions $\mu_{1j}(Z)$, $\sigma_{1j}(Z)$, $j = 1,\ldots d_x$, $\mu_2(Z)$ and $\sigma_2(Z)$ would be constant and the vector $(\varepsilon_1,\varepsilon_2)$ would simply be a standardized version (mean zero and unit variance) of the original input and outputs. Model (5) is much more flexible since it allows more general setups where some dependence is possible between (X,Y) and Z. So, under the assumptions of model (5), ε_2 and ε_1 can be interpreted as "pure" input and outputs, because due to the independence between the vector $(\varepsilon_1,\varepsilon_2)$ and Z, they can be viewed as being whitened versions of Y and X, respectively. Note that no particular assumption is made on the distribution of $(\varepsilon_1,\varepsilon_2)$. So our model remains basically nonparametric.

The minimum input frontier defined as $\varphi(\varepsilon_1)$, gives the minimal achievable level of the input ε_2 for units producing at least the level of output ε_1. In the same spirit, we can also define in this space of pure input and outputs the order-m frontier, to obtain a less extreme benchmark frontier.

For a particular unit with current value $(\varepsilon_1,\varepsilon_2)$, it will be easy to define a pure (input) inefficiency measure by the distance between this value and the point $(\varepsilon_1,\varphi(\varepsilon_1))$. Since values of the variables can be negative (they are centered and scaled), directional distances can be used. We choose to work with the difference as measure of pure inefficiency:

$$\rho(\varepsilon_1,\varepsilon_2) = \varepsilon_2 - \varphi(\varepsilon_1). \tag{8}$$

Note that $\rho(\cdot,\cdot)$ is always non-negative for points in the attainable set, and $\rho(\varepsilon_1,\varepsilon_2) = 0$ indicates an efficient unit.

It should be noticed, that under model (5), comparison and ranking of firms is legitimate, since the effect of environmental factors has been eliminated. In the standard approach, a measure of efficiency based on $\psi(x)$ ignores completely the possible effect of the factors Z, and a measure based solely on $\tau(x, z)$ is useless when comparing firms with different values of z.

Updating the ideas of Daraio and Simar (2005, 2007), the analysis of the global effect of the environmental factors on the production process can be captured by analyzing the ratios $\tau(x, z)/\psi(x)$ and $\tau_m(x, z)/\psi_m(x)$ as a function of z, at various fixed levels of the outputs x. For the input orientation, and for a fixed level of the outputs, when these ratios are globally increasing with z, this indicates an unfavorable effect of z on the production process (Z behaves like an undesirable output). On the contrary, when these ratios are globally decreasing with z, we have a favorable effect of z on the process (Z behaves like a free available input). Note that the analysis of these ratios has to be done for fixed levels of outputs x, averaging over possible values of x would introduce additional features that make the interpretation much more delicate (depending on the dependence between Z and X). As pointed out in Bădin et al. (2011), the full frontier ratios indicate only the effect of z on the shape of the frontier, whereas with partial frontiers (unless m is large), this effect may combine effects on the shape of the frontier and effects on the conditional distribution of the inefficiencies (for instance, in the limiting case where $m = 1$ the order-m frontier captures the average behavior of the input and not its boundary). Of course, in practice, we are mainly interested in frontier estimation and so m will be large for this purpose; however, for the practitioner, the analysis of both ratios, eventually with several values of m, may also be useful.

What might also be of interest, is the analysis of the changes in the frontier levels in the input-output space, when varying the value of z. A graphical view of this would be possible in the bivariate case only (one input y and one output x) by plotting in the (x, y) coordinates, the function $y = \tau(x, z)$ for various values of z. We could also compare these conditional frontiers with the marginal frontier $y = \psi(x)$, the latter having no particular economic interpretation if Z is not "separable" from (X, Y), but can always be viewed as $\min_z \tau(x, z)$.

Note finally, that our approach also offers to the practitioner the additional tools $\mu_1(z)$ and $\mu_2(z)$ allowing to appreciate, marginally, the mean behavior of the input and the outputs as a function of z, in a flexible model.

Of course all these quantities are unknown and have to be estimated. In the next section we address the problem of estimating the model and its various components.

4 Estimation

4.1 The proposed estimator

Let $(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)$ be i.i.d. data generated from model (5), where we denote the components of X_i by $(X_{i1}, \dots, X_{id_x})$ and the components of Z_i by $(Z_{i1}, \dots, Z_{id_z})$ $(i = 1, \dots, n)$. We start by estimating the regression function $\mu_2(z)$ and the variance function $\sigma_2^2(z)$ for an arbitrary point $z = (z_1, \dots, z_{d_z})$ in the support R_Z of Z in $\mathbb{R}^{d_z}$, which we suppose to be compact. We estimate $\mu_2(z)$ by a local polynomial estimator of degree p [see Fan and Gijbels (1996) or Ruppert and Wand (1994), among others], i.e. $\widehat{\mu}_2(z) = \widehat{\beta}_0$, where $\widehat{\beta}_0$ is the first component of the vector $\widehat{\beta}$, which is the solution of the local minimization problem

$$\min_{\beta} \sum_{i=1}^{n} \Big\{ Y_i - P_i(\beta, z, p) \Big\}^2 K_h(Z_i - z), \tag{9}$$

where $P_i(\beta, z, p)$ is a polynomial of order p built up with all $0 \leq k \leq p$ products of factors of the form $Z_{ij} - z_j$ $(j = 1, \dots, d_z)$. The vector β is the vector consisting of all coefficients of this polynomial. Here, for $u = (u_1, \dots, u_{d_z}) \in \mathbb{R}^{d_z}$, $K(u) = \prod_{j=1}^{d_z} k(u_j)$ is a d_z-dimensional product kernel, k is a univariate kernel function, $h = (h_1, \dots, h_{d_z})$ is a d_z-dimensional bandwidth vector converging to zero when n tends to infinity, and $K_h(u) = \prod_{j=1}^{d_z} k(u_j/h_j)/h_j$. In a similar way, we define $\widehat{\sigma}_2^2(z) = \widehat{\gamma}_0$, where $\widehat{\gamma}_0$ is the first component of the vector $\widehat{\gamma}$, which is the solution of the local minimization problem[2]

$$\min_{\gamma} \sum_{i=1}^{n} \Big\{ (Y_i - \widehat{\mu}_2(Z_i))^2 - P_i(\gamma, z, p) \Big\}^2 K_h(Z_i - z). \tag{10}$$

Now, let for $i = 1, \dots, n$,

$$\widehat{\varepsilon}_{2i} = \frac{Y_i - \widehat{\mu}_2(Z_i)}{\widehat{\sigma}_2(Z_i)}.$$

In order to estimate the components of $\mu_1(\cdot)$ and $\sigma_1^2(\cdot)$, we follow the same local polynomial estimation procedure as above but with Y_i replaced by the components of X_i, which leads to

$$\widehat{\varepsilon}_{1i} = \frac{X_i - \widehat{\mu}_1(Z_i)}{\widehat{\sigma}_1(Z_i)}$$

(where the ratio has to be understood componentwise).

[2] For simplifying the presentation the order of the polynomial and the bandwidth vector are taken to be the same as for the estimation of $\mu_2(z)$, but we could also work with different orders and bandwidths.

We are now ready to estimate the (full) frontier $\varphi(t_1)$ of ε_2 given $\varepsilon_1 \geq t_1$:

$$\widehat{\varphi}(t_1) = \min\{\widehat{\varepsilon}_{2i} : \widehat{\varepsilon}_{1i} \geq t_1\}.$$

For the order-m frontier $\varphi_m(t_1)$ defined in (7), note that for any $M \leq \varphi(t_1)$, this can also be written as

$$\varphi_m(t_1) = -\int_{-\infty}^{\infty} t_2 dS^m_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = -\int_M^{\infty} t_2 dS^m_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = M + \int_M^{\infty} S^m_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\, dt_2.$$

Hence, a natural estimator of $\varphi_m(t_1)$ is

$$\widehat{\varphi}_m(t_1) = M + \int_M^{\infty} \widehat{S}^m_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\, dt_2,$$

where

$$\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = \frac{n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_{1i} \geq t_1, \widehat{\varepsilon}_{2i} \geq t_2)}{n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_{1i} \geq t_1)}.$$

We suppose in the sequel that M is known. In practice, M can be chosen as any value smaller than $\min(\widehat{\varepsilon}_{21}, \ldots, \widehat{\varepsilon}_{2n})$. Note that the value of M has no influence on the estimator $\widehat{\varphi}_m(t_1)$.

Finally, define

$$\widehat{\tau}(x,z) = \widehat{\mu}_2(z) + \widehat{\varphi}\Big(\frac{x - \widehat{\mu}_1(z)}{\widehat{\sigma}_1(z)}\Big)\widehat{\sigma}_2(z),$$

and

$$\widehat{\tau}_m(x,z) = \widehat{\mu}_2(z) + \widehat{\varphi}_m\Big(\frac{x - \widehat{\mu}_1(z)}{\widehat{\sigma}_1(z)}\Big)\widehat{\sigma}_2(z).$$

4.2 Asymptotic results

From now on, for simplicity we take all h_j equal: $h_j = h$ for $j = 1, \ldots, d_z$. We restrict attention to the case where p is odd, as it is well known that this case outperforms the case where p is even in terms of the mean squared error of the estimators $\widehat{\mu}_j(z)$ and $\widehat{\sigma}_j(z)$ $(j = 1, 2)$.[3] For any $z \in R_Z$ and for $j = 1, 2$, we know that there exists functions b_{μ_j} and b_{σ_j} such that (where $f_Z(z) = F_Z'(z)$ and $F_Z(z) = P(Z \leq z)$)

$$\widehat{\mu}_j(z) - \mu_j(z) = n^{-1}\sum_{i=1}^n K_h(Z_i - z)\varepsilon_{ji}\sigma_j(z)f_Z^{-1}(z) + h^{p+1}b_{\mu_j}(z) + o_P((nh^{d_z})^{-1/2})$$

$$\widehat{\sigma}_j(z) - \sigma_j(z) = \frac{1}{2}n^{-1}\sum_{i=1}^n K_h(Z_i - z)(\varepsilon_{ji}^2 - 1)\sigma_j(z)f_Z^{-1}(z) + h^{p+1}b_{\sigma_j}(z) + o_P((nh^{d_z})^{-1/2}),$$

[3] Our asymptotic theory also works in the case where p is even, but the formulae of the bias of $\widehat{\mu}_j(z)$ and $\widehat{\sigma}_j(z)$ are different in that case.

provided $(nh^{d_z})h^{2(p+1)} = O(1)$. See e.g. Fan and Gijbels (1996) or Masry (1997) for the precise formula of b_{μ_j}. The formula of b_{σ_j} can be obtained using a similar development. For what follows, we also need the notation $f_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = -\frac{\partial}{\partial t_2} S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)$, which is the conditional probability density function of ε_2 given that $\varepsilon_1 \geq t_1$.

The assumptions under which the results below are valid, can be found in the Appendix. The results below rely on Akritas and Van Keilegom (2001) and Neumeyer and Van Keilegom (2010), who studied the asymptotic properties of the estimator $\widehat{S}_{\varepsilon_1}(t_1) = n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_{1i} \geq t_1)$ when $d_z = 1$ and when $d_z \geq 1$ respectively.

Theorem 1 *Assume (C1)-(C5) and assume h satisfies $nh^{2p+2} = O(1)$. Then,*

$$\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = n^{-1}\sum_{i=1}^n h(X_i, Y_i, Z_i, t_2|t_1) + h^{p+1} b(t_2|t_1) + R_n(t_2|t_1),$$

where (with $\varepsilon_1 = (X - \mu_1(Z))/\sigma_1(Z)$ and $\varepsilon_2 = (Y - \mu_2(Z))/\sigma_2(Z)$)

$$\begin{aligned} h(X,Y,Z,t_2|t_1) &= \frac{I(\varepsilon_1 \geq t_1)}{S_{\varepsilon_1}(t_1)}\Big[I(\varepsilon_2 \geq t_2) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\Big] \\ &\quad + \frac{\partial}{\partial t_1} S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\Big[\varepsilon_1 + \frac{t_1}{2}\{\varepsilon_1^2 - 1\}\Big] - f_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\Big[\varepsilon_2 + \frac{t_2}{2}\{\varepsilon_2^2 - 1\}\Big], \\ b(t_2|t_1) &= \frac{\partial}{\partial t_1} S_{\varepsilon_2|\varepsilon_1}(t_2|t_1) \int \frac{b_{\mu_1}(z) + t_1 b_{\sigma_1}(z)}{\sigma_1(z)} f_Z(z)\, dz \\ &\quad - f_{\varepsilon_2|\varepsilon_1}(t_2|t_1) \int \frac{b_{\mu_2}(z) + t_2 b_{\sigma_2}(z)}{\sigma_2(z)} f_Z(z)\, dz, \end{aligned}$$

and where

$$\sup_{t_1,t_2} |R_n(t_2|t_1)| = o_P(n^{-1/2}).$$

Corollary 2 *Assume (C1)-(C5). Then,*

(i) If h satisfies $h = C_1 n^{-1/(2p+2)}(1 + o(1))$ for some $0 \leq C_1 < \infty$, the process $n^{1/2}(\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1))$, $t_1 \in \mathbb{R}^{d_x}, t_2 \in \mathbb{R}$, converges weakly to a Gaussian process $W(t_2|t_1)$ with covariance function given by

$$Cov(W(s_2|s_1), W(t_2|t_1)) = E[h(X,Y,Z,s_2|s_1) h(X,Y,Z,t_2|t_1)],$$

and mean function given by $E(W(t_2|t_1)) = C_1^{p+1} b(t_2|t_1)$.

(ii) If h satisfies $h = C_1 n^{-1/(2p+2)}(1 + o(1))$ for some $0 \leq C_1 < \infty$, the process $n^{1/2}(\widehat{\varphi}_m(t_1) - \varphi_m(t_1))$, $t_1 \in \mathbb{R}^{d_x}$ (m fixed), converges weakly to a Gaussian process

$Z(t_1)$ with covariance function given by

$$\begin{aligned} &Cov(Z(s_1), Z(t_1)) \\ &= m^2 \int\int S^{m-1}_{\varepsilon_2|\varepsilon_1}(s_2|s_1) S^{m-1}_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\, Cov(W(s_2|s_1), W(t_2|t_1))\, ds_2 dt_2, \end{aligned}$$

and mean function given by

$$E(Z(t_1)) = C_1^{p+1} m \int S^{m-1}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) b(t_2|t_1)\, dt_2.$$

(iii) If h satisfies $h = C_2 n^{-1/(2p+2+d_z)}(1 + o(1))$ for some $0 \leq C_2 < \infty$, the process $(nh^{d_z})^{1/2}(\widehat{\tau}_m(x, z) - \tau_m(x, z))$, $x \in \mathbb{R}^{d_x}$ (with both $z \in R_Z$ and m fixed), converges weakly to a Gaussian process $V(x)$ with covariance function given by

$$Cov(V(x_1), V(x_2)) = E\Big[g(X, Y, Z, x_1) g(X, Y, Z, x_2)|Z = z\Big] \int K^2(u)\, du\, f_Z^{-1}(z),$$

where

$$\begin{aligned} g(X, Y, Z, x) = \Big[&\varepsilon_2 + \frac{1}{2}\varphi_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)(\varepsilon_2^2 - 1) \\ &-\varphi'_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)\Big\{\varepsilon_1 + \frac{x - \mu_1(z)}{2\sigma_1(z)}(\varepsilon_1^2 - 1)\Big\}\Big]\sigma_2(z), \end{aligned}$$

and mean function given by

$$\begin{aligned} E(V(x)) = C_2^{p+1+d_z/2}\Big[&b_{\mu_2}(z) + \varphi_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big) b_{\sigma_2}(z) \\ &+\sigma_2(z) m \int S^{m-1}_{\varepsilon_2|\varepsilon_1}\Big(t_2\Big|\frac{x - \mu_1(z)}{\sigma_1(z)}\Big) b\Big(t_2\Big|\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)\, dt_2 \\ &-\varphi'_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)\Big\{\frac{b_{\mu_1}(z)}{\sigma_1(z)} + \frac{x - \mu_1(z)}{\sigma_1^2(z)} b_{\sigma_1}(z)\Big\}\sigma_2(z)\Big]. \end{aligned}$$

Note that as z is kept fixed, we do not have a process in z and hence there are no tightness problems in z. Moreover, the estimator of $\varphi_m(\cdot)$ has no effect on the limit, since it converges at faster rate than the estimators $\widehat{\mu}_j(z)$ and $\widehat{\sigma}_j(z)$ $(j = 1, 2)$. Also, note that if μ_j and σ_j $(j = 1, 2)$ would be estimated parametrically, $\widehat{\tau}_m(x, z)$ would have a parametric rate of convergence, and both $\widehat{\varphi}_m$, $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ would contribute to the limit.

As a last asymptotic result of this section we show the weak consistency of the estimators $\widehat{\varphi}(t_1)$ and $\widehat{\tau}(x, z)$ of the full frontiers $\varphi(t_1)$ and $\tau(x, z)$. The weak convergence of these estimators is a much harder problem, and is beyond the scope of this paper.

Theorem 3 *Assume (C1)-(C5). Then,*

$$\widehat{\varphi}(t_1) - \varphi(t_1) = o_P(1),$$

and

$$\widehat{\tau}(x,z) - \tau(x,z) = o_P(1),$$

for all $t_1, x \in \mathbb{R}^{d_x}$ *and* $z \in R_Z$*.*

5 Practical Aspects

In this section we focus on a number of important issues related to our method, and we explain how they can be dealt with in practice: the use of a bootstrap method to approximate the distribution of the pure and the conditional frontier functions, the validity of our model when working with other location and scale functionals, and the verification of our location-scale model. Other issues (like the selection of the bandwidth parameters) will be explained in detail in the simulation section.

Bootstrap approximation

Although the asymptotic limits obtained in the previous section are normal distributions and the formulas of the asymptotic bias and variance are explicit, their estimation in practice can be cumbersome. It might therefore be useful to have a bootstrap procedure at hand, which can be used to calculate confidence intervals or test hypotheses concerning the frontier function. We propose to work with the following bootstrap method. For $i = 1, \ldots, n$, generate

$$\begin{cases} Z_i^* = Z_i, \\ (\varepsilon_{1i}^*, \varepsilon_{2i}^*) \sim \widetilde{F}_{\varepsilon_1,\varepsilon_2} \quad \text{i.i.d.}, \\ X_i^* = \widehat{\mu}_1(Z_i^*) + \widehat{\sigma}_1(Z_i^*)\varepsilon_{1i}^* \quad \text{and} \quad Y_i^* = \widehat{\mu}_2(Z_i^*) + \widehat{\sigma}_2(Z_i^*)\varepsilon_{2i}^*, \end{cases}$$

where $\widetilde{F}_{\varepsilon_1,\varepsilon_2}(t_1,t_2)$ is the distribution corresponding to the density

$$\widetilde{f}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) = \frac{1}{na_n^{d_x+1}} \sum_{i=1}^n K\Big(\frac{\widetilde{\varepsilon}_{1i} - t_1}{a_n}\Big) k\Big(\frac{\widetilde{\varepsilon}_{2i} - t_2}{a_n}\Big),$$

where $\widetilde{\varepsilon}_{ji} = [\widehat{\varepsilon}_{ji} - \overline{\widehat{\varepsilon}_j}]/\text{std}(\widehat{\varepsilon}_j)$ is the standardized version of the residual $\widehat{\varepsilon}_{ji}$ (with $\overline{\widehat{\varepsilon}_j} = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_{ji}$ and $\text{std}^2(\widehat{\varepsilon}_j) = (n-1)^{-1}\sum_{i=1}^n [\widehat{\varepsilon}_{ji} - \overline{\widehat{\varepsilon}_j}]^2$), and where k is a univariate kernel, K is a d_x-dimensional kernel, and a_n is an appropriate bandwidth sequence. The consistency of this smooth residual bootstrap procedure has been shown by Neumeyer (2006, 2009a) in the case of a single error variable, and has been applied in various papers in the literature, see e.g. Dette et al. (2007). Note that the smoothness of the distribution of the bootstrap residuals $(\varepsilon_{1i}^*, \varepsilon_{2i}^*)$ is crucial to prove the consistency of the bootstrap. This can be explained by the fact that the asymptotic representation of $\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1, t_2)$ (see Theorem 1) depends on the derivatives of $S_{\varepsilon_1,\varepsilon_2}(t_1, t_2)$ with respect to t_1 and t_2. In practice the bandwidth a_n can be chosen very small, usually much smaller than bandwidths used for estimation purposes, although the bootstrap procedure is quite stable with respect to the choice of a_n.

Based on this new bootstrap sample (X_i^*, Y_i^*, Z_i^*) $(i = 1, \ldots, n)$, we can now recalculate the pure and the conditional frontier. By repeating this procedure a large number of times, we obtain an approximation of the distribution of these frontiers.

Choice of μ_j and σ_j

Suppose that instead of working with the conditional mean ($\mu_1(Z) = E(X|Z)$ and $\mu_2(Z) = E(Y|Z)$) and the conditional variance ($\sigma_1^2(Z) = (\mathrm{Var}(X_1|Z), \ldots, \mathrm{Var}(X_{1d_x}|Z))^t$ and $\sigma_2^2(Z) = \mathrm{Var}(Y|Z)$), one would like to work with another location function $\widetilde{\mu}_j$ (e.g. the conditional median or trimmed mean) and another scale function $\widetilde{\sigma}_j$ (e.g. the conditional interquartile range). Then, we can write

$$\begin{cases} X = \widetilde{\mu}_1(Z) + \widetilde{\sigma}_1(Z)\widetilde{\varepsilon}_1 \\ Y = \widetilde{\mu}_2(Z) + \widetilde{\sigma}_2(Z)\widetilde{\varepsilon}_2, \end{cases} \tag{11}$$

for certain new error terms $\widetilde{\varepsilon}_1$ and $\widetilde{\varepsilon}_2$. An important question is then the following one: "Is model (11) again a location-scale model, i.e. is $(\widetilde{\varepsilon}_1, \widetilde{\varepsilon}_2)$ independent of Z ?". In Lemma 4 in the Appendix we show that this is indeed the case for any location function $\widetilde{\mu}_j$ and any scale function $\widetilde{\sigma}_j$. Hence, all the results and interpretations of this paper are valid not only for the classical mean-variance model, but for any location-scale model.

Testing the independence between $(\varepsilon_1, \varepsilon_2)$ and Z

A crucial assumption of our location-scale model is the independence between the vector of errors $(\varepsilon_1, \varepsilon_2)$ and the vector of environmental variables Z. In the literature procedures have been developed for testing the independence between a single error (say ε_1) and a single variable (say Z_1), see e.g. Einmahl and Van Keilegom (2008a, 2008b) and Neumeyer (2009b). These tests can however be easily generalized to multivariate errors and environmental variables. Consider e.g. the test developed in Einmahl and Van Keilegom (2008a), which relies on the difference $\widehat{F}_{\varepsilon_1,Z_1} - \widehat{F}_{\varepsilon_1}\widehat{F}_{Z_1}$. The obvious extension to our case consists in considering the process

$$\widehat{F}_{\varepsilon_1,\varepsilon_2,Z}(\cdot,\cdot) - \widehat{F}_{\varepsilon_1,\varepsilon_2}(\cdot)\widehat{F}_Z(\cdot),$$

from which a Kolmogorov-Smirnov or Cramer-von Mises type test statistic can be easily developed. Einmahl and Van Keilegom (2008a) show that the process is asymptotically distribution free (and the limit equals the limit one would obtain when ε_1 and ε_2 would be observed!). However, the convergence to this limit being rather slow, they advocate the use of bootstrap methods to obtain the critical values of the proposed test statistics.

Note that, as shown above, either the assumption of independence holds for all location and scale functions μ_j and σ_j, either for none.

6 Numerical Illustrations

6.1 Simulated samples

We will illustrate our approach in two simulated scenarios. We restrict the analysis to the univariate case, because it allows to illustrate the different components of our model in 2 or 3-dimensional pictures.

Example 1

We first simulate data according to the following scheme. Let the exogenous factor Z_i be uniformly distributed on the interval $[1, 3]$ and the "pure" output ε_{1i} be uniformly distributed on the interval $[-\sqrt{3}, \sqrt{3}]$ (note that the ε_{1i}'s have mean zero and variance one). We define the pure efficient frontier by the function $\widetilde{\varphi}(\varepsilon_1) = \exp(\varepsilon_1)$ and we generate the inefficient pure input as $\widetilde{\varepsilon}_{2i} = \widetilde{\varphi}(\varepsilon_{1i}) + U_i$ where $U_i \sim \mathcal{N}^+(0, \sigma_U^2)$ with $\sigma_U = 1$. We rescale then the values of the $\widetilde{\varepsilon}_{2i}$'s so that they have mean zero and variance one. The new (standardized) frontier is then denoted by $\varphi(\varepsilon_{1i})$.

For the location-scale model, we choose to impact mostly the inputs by Z (location and scale) and to impact only the scale function for the outputs:

$$\begin{aligned} \mu_1(Z) &= 6 \quad \text{and} \quad \sigma_1(Z) = 1 + Z, \\ \mu_2(Z) &= 10 - 2Z \quad \text{and} \quad \sigma_2(Z) = 1 + 4(Z-2)^2. \end{aligned}$$

We present the results for a sample of size $n = 100$. Bandwidths were selected for each nonparametric regression by least-squares cross validation. Figure 1 displays the true and estimated location and scale functions for both the output X (left panel) and the input Y

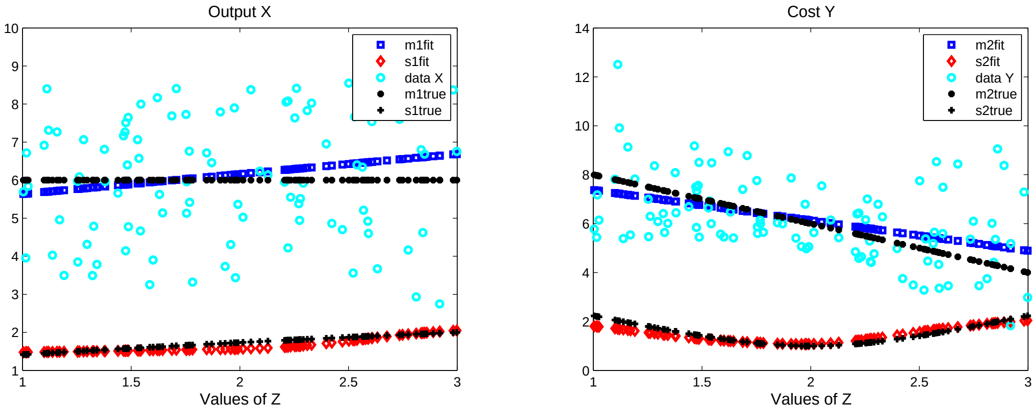

Figure 1: *Example 1: True and estimated location and scale functions: output case (left panel) and input case (right panel).*

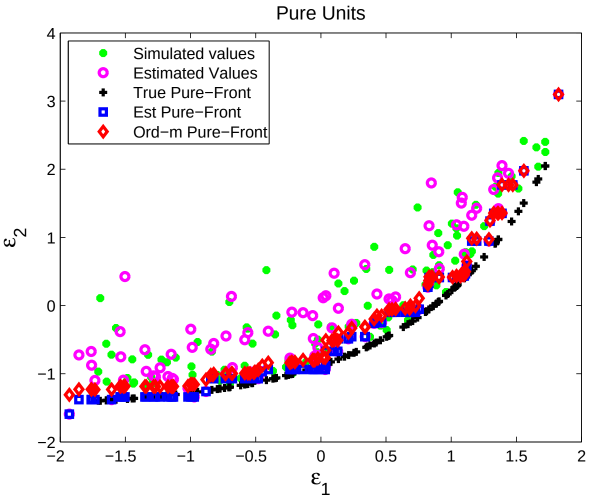

Figure 2: *Example 1: Pure input and output: simulated (true values) and estimated pairs, with the true pure efficient frontier and its estimates (full and order-m, with $m = 25$).*

(right panel). In Figure 2, we see how well the pure input and output have been estimated: both the simulated original "true" values and the estimates are displayed. The Figure shows also the true full frontier and the two estimators: full-frontier and order-m, with $m = 25$. If we go back to the original units, in the input and output space we obtain the results shown in Figure 3. For each data point (X_i, Y_i), we have the estimate of the conditional efficient frontier (corresponding to its value of Z_i) and the corresponding true value. We can appreciate how well the estimator performs even with a relatively small

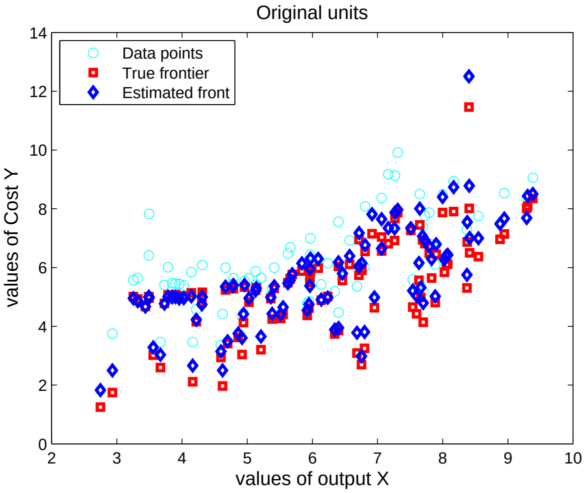

Figure 3: *Example 1: Input and output in the original units: the data points, the true conditional efficient frontier for each data point and its estimate (we display here only the full frontier estimates).*

sample size.

Figure 4 is another way to appreciate the quality of the fit: a perfect fit would give a straight line as the solid line in the figure. The figure also plots the fitted values one would obtain by following the traditional nonparametric approach, based on the nonparametric estimators of $S_{Y|X,Z}(y|x,z)$, as described in Cazals et al. (2002) and Daraio and Simar (2005). Of course, this is just an example with one simulated sample, but it indicates a very good behavior of our estimator, and as expected, better than the nonparametric estimator since we simulated the data according to a location-scale model. The integrated squared error (ISE) of the two estimators, estimated over the 100 data points, is respectively 0.8065 for the nonparametric estimator and 0.0969 for our location-scale estimator.

Example 2

We keep the same dataset as above but we add an outlier at the point $(\varepsilon_{1i}, \varepsilon_{2i}) = (0.5, -1.25)$ (it appears clearly in Figure 6). It is remarkable how the estimators of the location and of the scale functions are not sensitive to this data point (because they smooth the data in their center). We also see how, as expected, even with $m = 25$, the order-m frontier is well resistant to this outlying data point, which is not the case for the full frontier estimate. The results are displayed below. In Figure 8 we observe again a

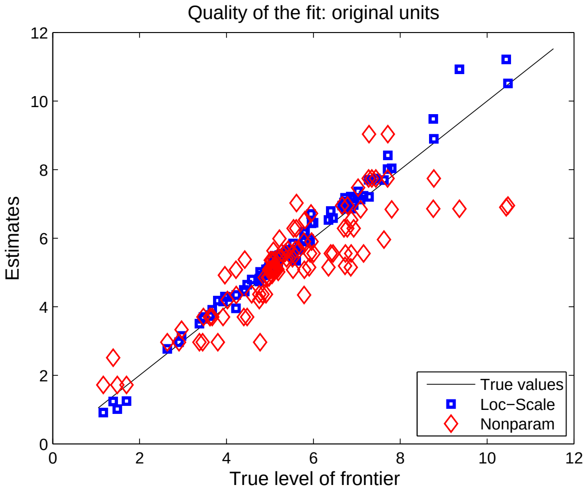

Figure 4: *Example 1: The horizontal axis is for the true levels of the conditional efficient full frontier, evaluated at all the data points, the squares are their corresponding location scale estimates and the diamonds are the corresponding nonparametric estimates.*

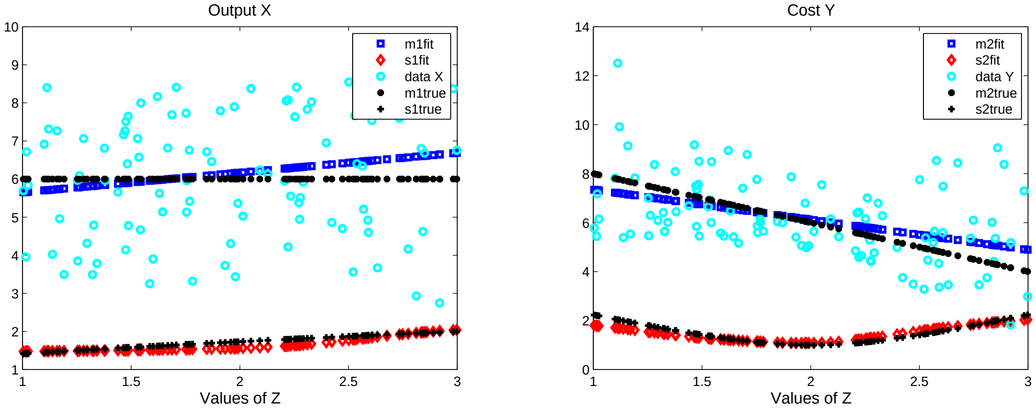

Figure 5: *Example 2: True and estimated location and scale functions: output case (left panel) and input case (right panel).*

very good quality of the fit of our estimator, again better than the nonparametric one (the respective ISE's are here 0.2417 and 0.9036).

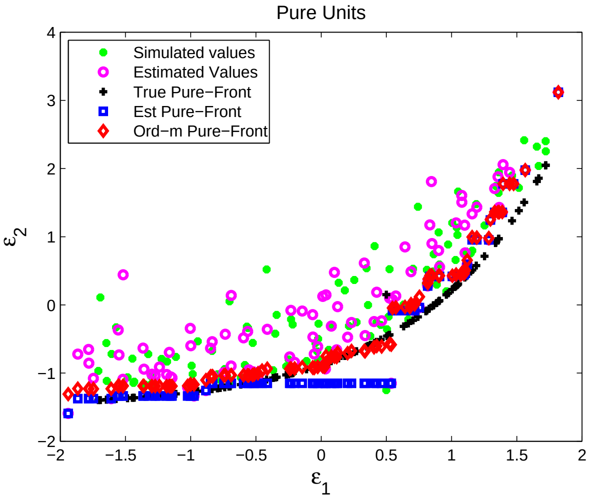

Figure 6: *Example 2: Pure input and output: simulated (true values) and estimated pairs, with the true pure efficient frontier and its estimates (full and order-m, with $m = 25$).*

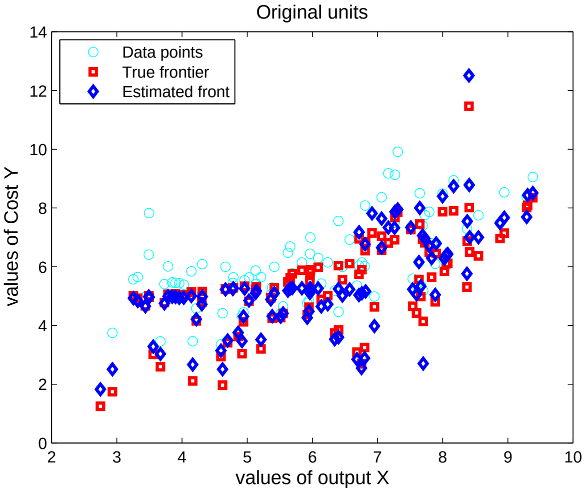

Figure 7: *Example 2: Input and output in the original units: the data points, the true conditional efficient frontier for each data point and its estimate (we display here only the full frontier estimates).*

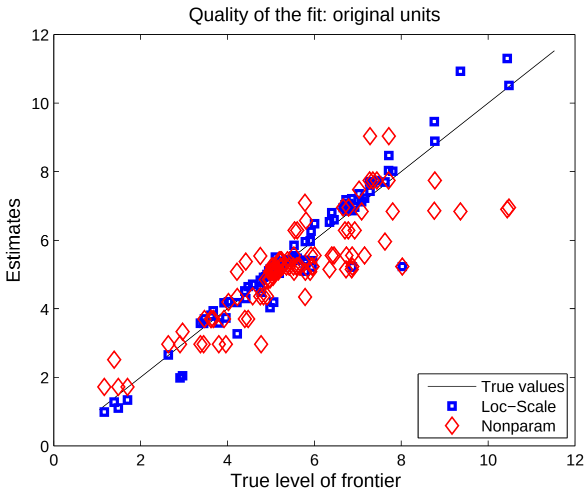

Figure 8: *Example 2: The horizontal axis is for the true levels of the conditional efficient full frontier, evaluated at all the data points, the squares are their corresponding location scale estimates and the diamonds are the corresponding nonparametric estimates.*

6.2 Efficiency in the banking sector

We illustrate our procedure with a real dataset coming from the banking sector.[4] Simar and Wilson (2007) analyzed these data based on Aly et al. (1990) using data on 6.955 US commercial banks observed at the end of the 4th quarter of 2002.

The original dataset contains 3 inputs (purchased funds, core deposits and labor) and 4 outputs (consumer loans, business loans, real estate loans, and securities held) for banks. Aly et al. (1990) considered 2 continuous environmental factors, the size of the banks Z_1, and a measure of the diversity of the services proposed by the banks Z_2 (see Aly. et al., 1990, for details). We will use, as in Simar and Wilson (2007), a measure of the size of the banks by the log of the total assets, rather than the total deposit.

Daraio et al. (2010) used the same dataset to test the "separability" condition in the same setup. The hypothesis was rejected at any reasonable level, indicating that any traditional two-stage procedure is meaningless for this dataset. So, the approach using conditional efficiency scores seems to be much more appropriate. This was suggested in Bădin et al. (2011) where they used the traditional nonparametric estimator of $S_{Y|X,Z}(y\,|\,x,z)$. We will rather use our location-scale model. We select, for the illustration, a subset of 303 banks as in Simar and Wilson (2007) and in Bădin et al. (2011). In

[4] We would like to thank Paul W. Wilson who provided us this dataset.

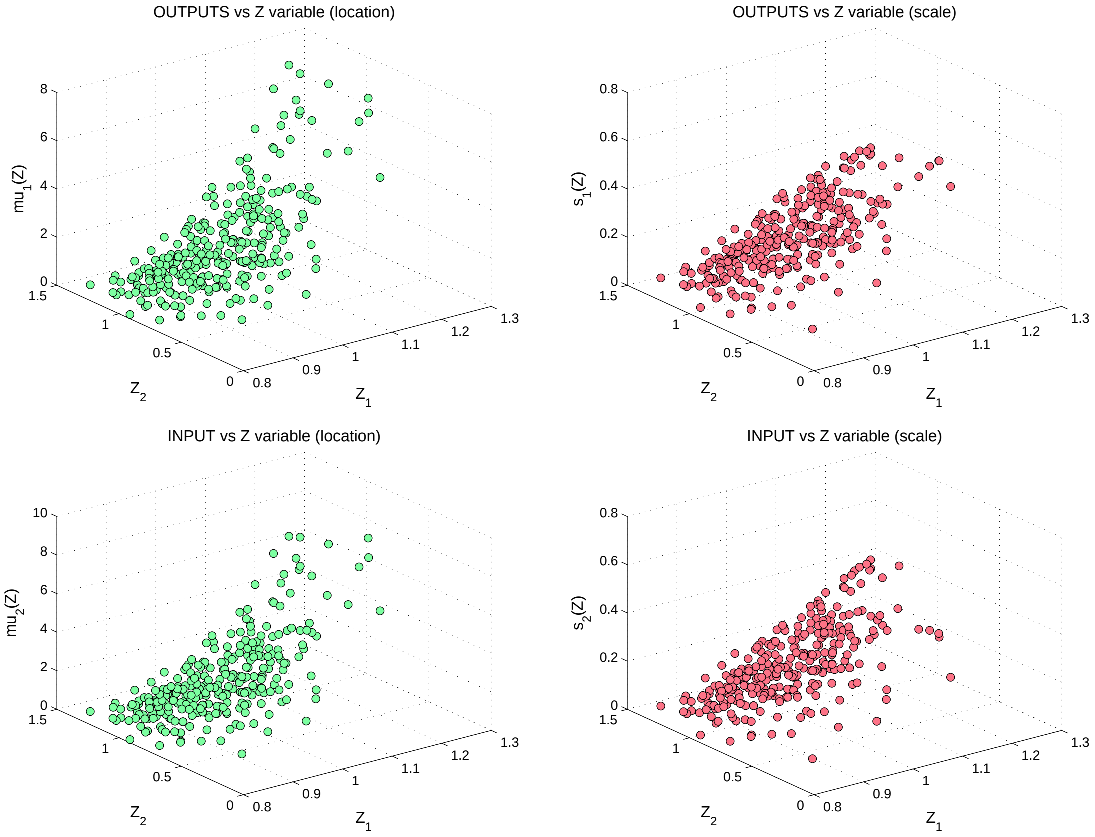

Figure 9: *Bank example: location (left) and scale (right) regression at observed data points for the output X (top panels) and for the input Y (bottom panels). Z_1 is SIZE and Z_2 is DIVERSITY.*

the latter paper it is explained that the inputs can be aggregated in a one dimensional input measure, without loosing much information and the same is true for the outputs. The final output X is highly correlated (more than 0.93) with all the original outputs and the same is true for the input Y (correlation with the original inputs of more than 0.97). This facilitates the presentation of our empirical illustration.

Figure 9 gives the location and the scale functions for the output and for the input. We see from this first result that the variable Z_1 (SIZE) has much more influence on the two variables X and Y than Z_2 (DIVERSITY).

We also observe that the variable Z_1 has a positive effect on both the input and the output (exponential effect on the location and linear effect on the scale), as expected, since the size of the banks is certainly determining the levels of the inputs and the outputs. The variable Z_2 does not show substantial effects on neither input nor output.

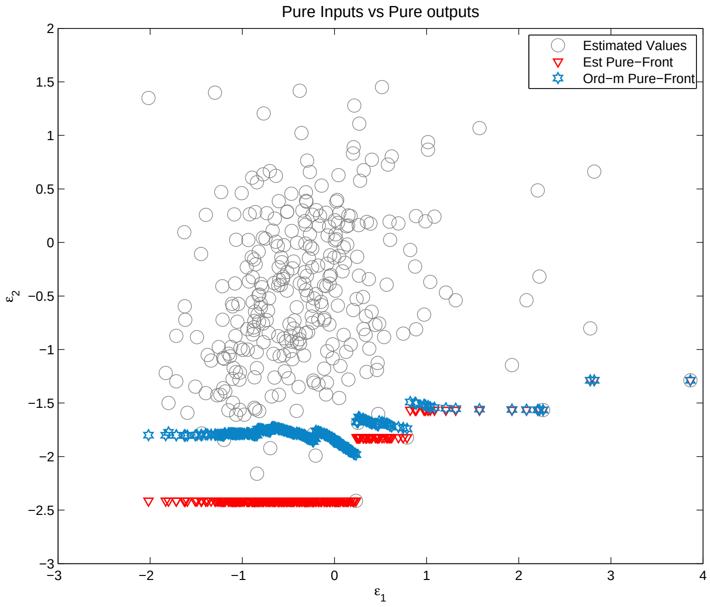

Figure 10: *Bank example: Estimated "pure" inputs and outputs, and efficient frontiers.*

Figure 10 displays the values of the estimated errors $\widehat{\varepsilon}_1$ and $\widehat{\varepsilon}_2$ together with the full frontier $\widehat{\varphi}$ and the order-m frontier $\widehat{\varphi}_m$ with $m = 30$ (we select $m = 30$ for illustration: the units are benchmarked against 10 percent of the full dataset). We observe that there is one extreme data point around $(\widehat{\varepsilon}_1, \widehat{\varepsilon}_2) = (0.2, -2.5)$, that is rather influential for both the full and, to a lesser extent, the order-m frontier (note that for $m = 10$, the order-m frontier is not attracted by this extreme data point). The distribution of the resulting measures of inefficiency $\widehat{\rho}(\widehat{\varepsilon}_1, \widehat{\varepsilon}_2)$ and $\widehat{\rho}_m(\widehat{\varepsilon}_1, \widehat{\varepsilon}_2)$ (defined in (8) above) is displayed in Figure 11. The distributions seem to be regular and reasonably bell-shaped, with some rare very inefficient banks. In a practical application, the detailed analysis of these efficiencies would be very informative. Tables 1 and 2 below give some detailed results for 15 randomly chosen banks.

It is also possible to draw a picture of the conditional frontier (and its order-m version) in the original units to stress the difference with the preceding one. Figure 12 displays the data points (X_i, Y_i) and the estimates $\widehat{\tau}(x, z)$ evaluated for each pair (X_i, Z_i). Of course, the frontier points (diamonds in the figure) do not follow the shape of the lower boundary of the cloud of data points (as would do the marginal frontier estimate $\widehat{\psi}(X_i)$), because here, each bank is facing different exogenous conditions determined by Z_i. Neither the marginal nor the order-m frontier estimates are displayed to make the picture more clear.

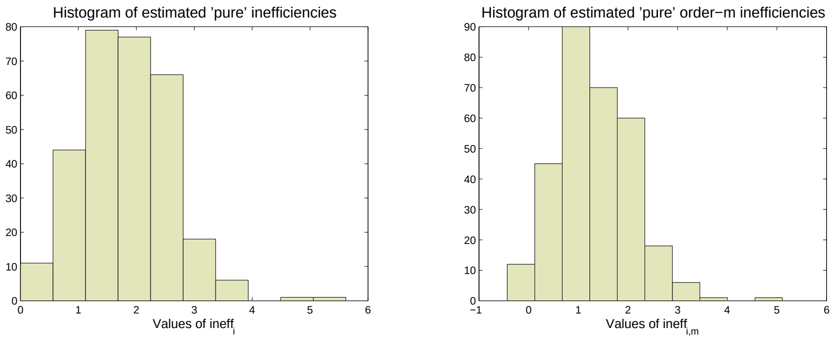

Figure 11: *Bank example: distribution of the estimated inefficiencies, relative to the full frontier and to the order-m frontier.*

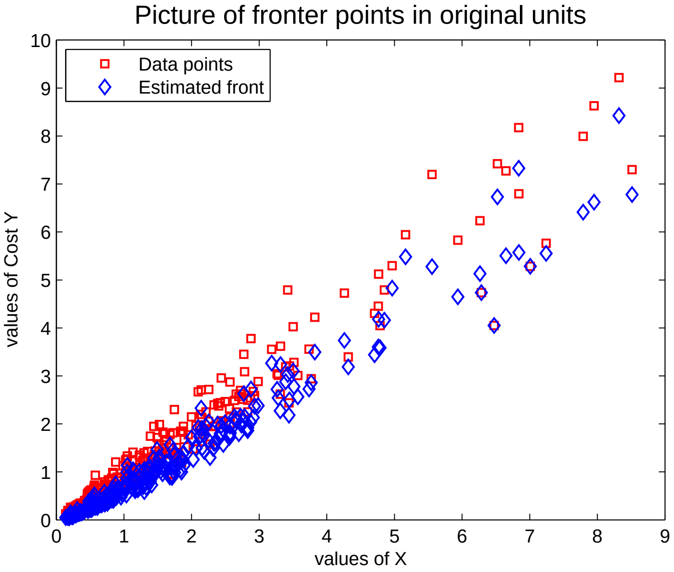

Figure 12: *Bank example: Data points* (X_i, Y_i) *and estimated (full) frontier points* $\widehat{\tau}(X_i, Z_i)$ *in the original units of the inputs and outputs.*

Let us now look at the shape of the frontier in the (X, Y) space for fixed values of the environmental conditions Z. Here Z is bivariate, so we do the exercise for selected fixed values of Z_1 and Z_2. We selected the 9 pairs $(QZ_{1k}, QZ_{2\ell})$, for $k, \ell = 1, 2, 3$, where QZ_{ik} is the kth quartile of Z_i $(i = 1, 2)$. Of course for each selected pair $(QZ_{1k}, QZ_{2\ell})$ we do not have many data points (for a fixed value of e.g. SIZE, we do not have so many data points with largely varying output values X) and the nonparametric evaluation of

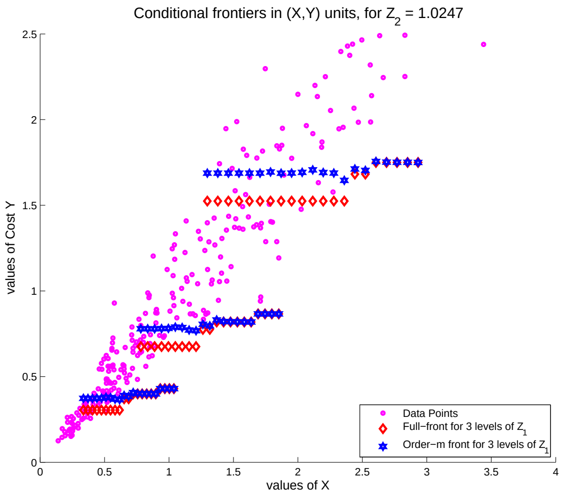

Figure 13: *Frontier estimates when fixing the level of Z. Here Z_2 is fixed at its median value, and Z_1 is fixed at its 3 quartiles (from the left to the right).*

the location-scale model remains rather imprecise. We would require more data points to give more sensible results. Still, for the illustration, the results are presented in Figure 13 for the median value of Z_2 and for the 3 different quartile values of Z_1. We see indeed that the frontier is moving when changing the level of Z_1, but the effect of X is difficult to capture when Z_1 is fixed. The pictures (not displayed here) were almost the same for the other 2 quartiles of Z_2 confirming the fact that Z_2 does not seem to have an effect on the production process.

Another question of interest, as explained in Section 3, is to see if the support of (X, Y) is changing with Z. We look at the ratios $\widehat{\tau}(x, z)/\widehat{\psi}(x)$ as a function of Z, for fixed values of X. Figure 14 shows the results for the output X fixed at its median value. We clearly see an effect of the variable Z_1 (SIZE), indicating a shift of the support of Y (non-favorable effect of Z_1 on the support of Y). The variable Z_2 seems to be without effect on the support of Y. The pictures for other values of X have the same shape (only the level of the surface is changing) and the pictures for the order-m ratios are also similar and are not reproduced for saving space. The analysis done on these ratios, with the same data, but using traditional nonparametric estimators of $S_{Y|X,Z}(y|x, z)$ (see Bădin et al., 2011) provided a less clear picture, but they arrived at the same qualitative conclusions (even if there, an output orientation was selected). Note also that this confirms that the separability condition is not reasonable, and so, two stage approaches are meaningless.

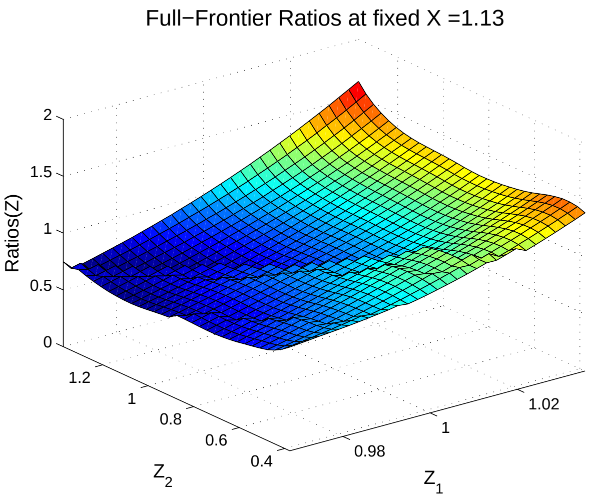

Figure 14: *Bank example: analysis of the ratios* $\widehat{\tau}(x,z)/\widehat{\psi}(x)$ *for a fixed value of* $x = Median(X)$.

Note that Aly et al. (1990) arrived by using this meaningless method to an opposite conclusion on the effect of SIZE.

Finally we use the bootstrap algorithm to test the independence between $(\varepsilon_1, \varepsilon_2)$ and Z (see Section 5 for more details about the bootstrap approximation and about the tests for independence). We know that the bandwidths for determining $(\widehat{\varepsilon}_1, \widehat{\varepsilon}_2)$ should be of smaller order than the optimal bandwidths determined by least-squares cross validation (LSCV). We computed the p-value of the null hypothesis scaling the optimal bandwidths by a factor c ranging from 0.25 to 1. Figure 15 shows the results, based on 2000 bootstrap replications. We see that the resulting p-values do not provide any evidence that we have to reject the null hypothesis of independence, since the p-values range from 0.83 $(c = 0.25)$ to 0.09 $(c = 1)$. Tables 1 and 2 give detailed individual results for 15 randomly selected banks, including 95% confidence intervals for the order-m individual efficiency scores. Note in the second column of Table 2 the ranking of these units based on the measure of pure inefficiencies.

7 Conclusions

We have considered the estimation of an input oriented frontier, conditional on environmental factors. The output oriented case can be considered following the same ideas. Traditional methods are based on a nonparametric estimator of a conditional survival

Unit i	X_i	Y_i	Z_{1i}	Z_{2i}	$\widehat{\varphi}(\widehat{\varepsilon}_{1i})$	$\widehat{\varphi}_m(\widehat{\varepsilon}_{1i})$	$\widehat{\tau}(X_i, Z_i)$	$\widehat{\tau}_m(X_i, Z_i)$
259	8.5127	7.2986	1.1778	1.1388	-1.5648	-1.5486	6.7794	6.7871
237	0.5186	0.3505	0.9409	0.8371	-2.4136	-1.8333	0.3026	0.3682
258	0.2958	0.1998	0.8700	1.2650	-2.4136	-1.8113	0.1207	0.1511
1	1.3999	1.1985	1.0199	0.8539	-2.4136	-1.8967	0.8848	0.9895
241	0.8173	0.8693	0.9863	0.8737	-2.4136	-1.7430	0.5892	0.6913
66	0.3546	0.3421	0.9028	0.8407	-2.4136	-1.8017	0.1463	0.2024
164	2.1868	1.8694	1.0551	1.1069	-2.4136	-1.8419	1.4248	1.5724
274	0.4185	0.4026	0.9152	1.0590	-2.4136	-1.7635	0.2086	0.2676
303	0.5053	0.2969	0.9024	1.1611	-1.8215	-1.7326	0.2263	0.2328
199	2.9132	2.6751	1.0786	0.9124	-1.5648	-1.5217	2.1336	2.1468
216	6.6494	7.2741	1.1670	1.0387	-2.4136	-1.7935	5.5037	5.8232
125	1.1407	1.0559	1.0111	0.6974	-2.4136	-1.8344	0.8100	0.9170
239	1.7164	1.3945	1.0330	1.2376	-2.4136	-1.7849	1.1322	1.2698
170	2.4389	2.9572	1.0807	0.8205	-2.4136	-1.8250	1.9605	2.1451
242	2.1842	1.8388	1.0735	0.8725	-2.4136	-1.8001	1.7629	1.9470

Table 1: *Results for 15 randomly selected banks. Data values and estimates of the frontier levels at the data points. Full and order-m frontiers, with* $m = 30$.

Unit i	Rank	$\widehat{\rho}_{m,i}$	low	up	$\widehat{\theta}_m(X_i, Z_i)$	low	up
259	129	1.0800	0.3436	1.7342	0.9299	0.8512	1.0141
237	3	-0.1572	-0.4273	0.1667	1.0507	0.8655	1.2895
258	105	0.9659	0.6929	1.2543	0.7560	0.8350	1.2592
1	119	1.0322	0.7212	1.4994	0.8256	0.6829	0.8974
241	139	1.1700	0.8825	1.2963	0.7952	0.7341	0.9146
66	189	1.5240	1.2444	1.7963	0.5917	0.4068	0.7686
164	138	1.1503	0.8907	1.4845	0.8411	0.7197	0.8687
274	184	1.4887	1.2002	1.6789	0.6647	0.6267	0.8582
303	94	0.8806	0.3203	1.5229	0.7841	0.6713	1.1445
199	209	1.7194	0.9145	2.2559	0.8025	0.6845	0.8691
216	294	2.8157	2.5357	3.0725	0.8005	0.7342	0.8432
125	72	0.7521	0.4226	1.1091	0.8685	0.7242	0.9907
239	50	0.5693	0.3276	0.6972	0.9106	0.7999	0.9635
170	290	2.5893	2.3239	2.8951	0.7254	0.6284	0.7402
242	2	-0.3607	-0.5861	-0.1834	1.0589	0.9307	1.0793

Table 2: *Results for the same units. Ranks obtained by ranking the* $\widehat{\rho}_i$ *and the estimates of the order-m inefficiencies with 95% confidence intervals. Here* $\widehat{\theta}_m = \widehat{\tau}_m(X_i, Z_i)/Y_i$.

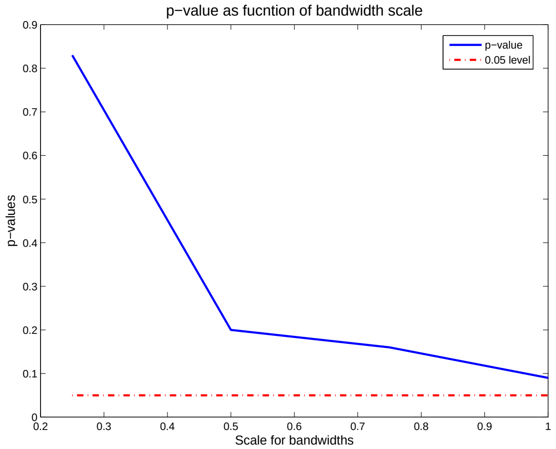

Figure 15: *Bank example: p-values for selected values of the scaling factor c of the LSCV optimal bandwidths.*

function, where one conditions on the level of these external factors. This requires the use of appropriate smoothing techniques, and hence these methods suffer from the curse of dimensionality.

In order to overcome these problems, we rather assume in this paper a location-scale model for both the input and the outputs, which avoids the curse of dimensionality at the boundary in the space of the whitened input and outputs.

This allows to estimate "pure" efficiencies by eliminating the effect of the environmental factors. Both full and order-m frontiers are estimated and we derive the asymptotic properties of these estimators. These measures of efficiency allow more reliable ranking of firms facing heterogeneous conditions of production.

We can also recover the frontiers in the original inputs/outputs space and we give their asymptotic properties. The approach is illustrated with some selected simulated data and also with a real dataset from the bank industry.

Appendix: Proofs

(C1) k is a symmetric probability density function supported on $[-1, 1]$, k is d_z times continuously differentiable, and $k^{(j)}(\pm 1) = 0$ for $j = 0, \ldots, d_z - 1$.

(C2) h satisfies $nh^{3d_z+\delta} \to \infty$ for some small $\delta > 0$.

(C3) All partial derivatives of F_Z up to order $2d_z+1$ exist on the interior of R_Z, they are uniformly continuous and $\inf_{z\in R_Z} f_Z(z) > 0$.

(C4) For $j = 1, 2$, all partial derivatives of μ_j and σ_j up to order $p+2$ exist on the interior of R_Z, they are uniformly continuous and $\inf_{z\in R_Z} \sigma_j(z) > 0$.

(C5) $S_{\varepsilon_1,\varepsilon_2}$ is twice continuously differentiable with respect to its components, $\max_{j,k=1,2} \sup_{t_1,t_2} |t_j t_k \frac{\partial^2}{\partial t_j \partial t_k} S_{\varepsilon_1,\varepsilon_2}(t_1, t_2)| < \infty$, and $E(\varepsilon_j^6) < \infty$ $(j = 1, 2)$.

Proof of Theorem 1. Write

$$\begin{aligned}
&\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\\
&= \widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big[\frac{1}{\widehat{S}_{\varepsilon_1}(t_1)} - \frac{1}{S_{\varepsilon_1}(t_1)}\Big] + \frac{1}{S_{\varepsilon_1}(t_1)}\Big[\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big],
\end{aligned} \tag{1}$$

where $\widehat{S}_{\varepsilon_1}(t_1) = n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_{1i} \geq t_1)$ and $\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) = n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_{1i} \geq t_1, \widehat{\varepsilon}_{2i} \geq t_2)$. We start with considering the second term of (1). Along the same lines as in the proof of Lemma A.3 in Neumeyer and Van Keilegom (2010), we can easily show that

$$\begin{aligned}
\sup_{t_1,t_2}\Big|n^{-1}\sum_{i=1}^n\Big\{I(\widehat{\varepsilon}_{1i} \geq t_1, \widehat{\varepsilon}_{2i} \geq t_2) - I(\varepsilon_{1i} \geq t_1, \varepsilon_{2i} \geq t_2) - S_{\widehat{\varepsilon}_1,\widehat{\varepsilon}_2}(t_1,t_2) + S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big\}\Big|\\
= o_P(n^{-1/2}),
\end{aligned}$$

where $S_{\widehat{\varepsilon}_1,\widehat{\varepsilon}_2}$ is the bivariate survival function of $(\widehat{\varepsilon}_1, \widehat{\varepsilon}_2)$ conditionally on the data (X_i, Y_i, Z_i), $i = 1, \ldots, n$ (i.e. considering $\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ as fixed functions). It now follows that

$$\begin{aligned}
&\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\\
&= n^{-1}\sum_{i=1}^n I(\varepsilon_{1i} \geq t_1, \varepsilon_{2i} \geq t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\\
&\quad + \int\Big\{S_{\varepsilon_1,\varepsilon_2}\Big(\frac{t_1\widehat{\sigma}_1(z) + \widehat{\mu}_1(z) - \mu_1(z)}{\sigma_1(z)}, \frac{t_2\widehat{\sigma}_2(z) + \widehat{\mu}_2(z) - \mu_2(z)}{\sigma_2(z)}\Big) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big\}dF_Z(z)\\
&\quad + o_P(n^{-1/2})\\
&= n^{-1}\sum_{i=1}^n I(\varepsilon_{1i} \geq t_1, \varepsilon_{2i} \geq t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\\
&\quad + \sum_{j=1}^2\Big[\frac{\partial}{\partial t_j}S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\int\sigma_j^{-1}(z)\big\{t_j(\widehat{\sigma}_j(z) - \sigma_j(z)) + \widehat{\mu}_j(z) - \mu_j(z)\big\}\,dF_Z(z)\Big]\\
&\quad + O_P\Big(\max_{j=1,2}\sup_z|\widehat{\mu}_j(z) - \mu_j(z)|^2\Big) + O_P\Big(\max_{j=1,2}\sup_z|\widehat{\sigma}_j(z) - \sigma_j(z)|^2\Big) + o_P(n^{-1/2}).
\end{aligned}$$

The O_P-terms above are $O_P((nh^{d_z})^{-1}\log n) = o_P(n^{-1/2})$, which follows from the proof of Lemma A.1 in Neumeyer and Van Keilegom (2010). Therefore, using Lemma A.2 in Neumeyer and Van Keilegom (2010), we have that

$$\begin{aligned}
&\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) \\
&= n^{-1}\sum_{i=1}^{n} I(\varepsilon_{1i} \geq t_1, \varepsilon_{2i} \geq t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) \\
&\quad + \sum_{j=1}^{2}\Big[\frac{\partial}{\partial t_j} S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big\{n^{-1}\sum_{i=1}^{n}\Big(\varepsilon_{ji} + \frac{t_j}{2}[\varepsilon_{ji}^2 - 1]\Big) \\
&\qquad + h^{p+1}\int \frac{b_{\mu_j}(z) + t_j b_{\sigma_j}(z)}{\sigma_j(z)} f_Z(z)\,dz\Big\}\Big] + o_P(n^{-1/2}),
\end{aligned} \tag{2}$$

uniformly in t_1 and t_2. Note that contrary to Neumeyer and Van Keilegom (2010), the asymptotic bias of $\widehat{\mu}_j(z)$ and $\widehat{\sigma}_j(z)$ is not negligible here, since we assume that $nh^{2p+2} = O(1)$ (instead of $nh^{2p+2} \to 0$ in the latter paper).

Consider now the first term of (1):

$$\begin{aligned}
&\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big[\frac{1}{\widehat{S}_{\varepsilon_1}(t_1)} - \frac{1}{S_{\varepsilon_1}(t_1)}\Big] \\
&= -S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) S_{\varepsilon_1}^{-2}(t_1)\Big[n^{-1}\sum_{i=1}^{n} I(\widehat{\varepsilon}_{1i} \geq t_1) - S_{\varepsilon_1}(t_1)\Big] + o_P(n^{-1/2}) \\
&= -S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) S_{\varepsilon_1}^{-2}(t_1)\Big[n^{-1}\sum_{i=1}^{n} I(\varepsilon_{1i} \geq t_1) - S_{\varepsilon_1}(t_1) \qquad\qquad (3) \\
&\quad - f_{\varepsilon_1}(t_1)\int \sigma_1^{-1}(z)\big\{t_1(\widehat{\sigma}_1(z) - \sigma_1(z)) + \widehat{\mu}_1(z) - \mu_1(z)\big\}\,dF_Z(z)\Big] + o_P(n^{-1/2}),
\end{aligned}$$

uniformly in t_1 and t_2, which follows using similar arguments as in the bivariate case, and provided $\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) = O_P(n^{-1/2})$ uniformly in t_1 and t_2. Indeed, it is easily seen that the class

$$\begin{aligned}
\mathcal{F} = \Big\{&(e_1,e_2) \to I(e_1 \geq t_1, e_2 \geq t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) + \sum_{j=1}^{2}\frac{\partial}{\partial t_j} S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)\Big[e_j + \frac{t_j}{2}(e_j^2 - 1)\Big] : \\
&t_1 \in \mathbb{R}^{d_x}, t_2 \in \mathbb{R}\Big\}
\end{aligned} \tag{4}$$

is Donsker (see again Neumeyer and Van Keilegom (2010) for more details). It now follows from (2) that $n^{1/2}(\widehat{S}_{\varepsilon_1,\varepsilon_2} - E\widehat{S}_{\varepsilon_1,\varepsilon_2})$ converges weakly to a Gaussian process, and hence

$$\sup_{t_1,t_2}|\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)| = O_P(n^{-1/2}),$$

since we also know that $\sup_{t_1,t_2} |E\widehat{S}_{\varepsilon_1,\varepsilon_2}(t_1,t_2) - S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)| = O(h^{p+1}) = O(n^{-1/2})$.

The result now follows from (1), (2) and (3), since it is readily seen that

$$\frac{S_{\varepsilon_1,\varepsilon_2}(t_1,t_2)}{S^2_{\varepsilon_1}(t_1)} f_{\varepsilon_1}(t_1) + \frac{1}{S_{\varepsilon_1}(t_1)} \frac{\partial}{\partial t_1} S_{\varepsilon_1,\varepsilon_2}(t_1,t_2) = \frac{\partial}{\partial t_1} S_{\varepsilon_2|\varepsilon_1}(t_2|t_1).$$

This completes the proof. □

Proof of Corollary 2. (i) Define the class

$$\begin{aligned}\mathcal{F} = \Big\{(e_1,e_2) \to \frac{I(e_1 \geq t_1)}{S_{\varepsilon_1}(t_1)} &\Big[I(e_2 \geq t_2) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\Big] \\ &+ \frac{\partial}{\partial t_1} S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\Big[e_1 + \frac{t_1}{2}\{e_1^2 - 1\}\Big] - f_{\varepsilon_2|\varepsilon_1}(t_2|t_1)\Big[e_2 + \frac{t_2}{2}\{e_2^2 - 1\}\Big] : \\ &t_1 \in \mathbb{R}^{d_x}, t_2 \in \mathbb{R}\Big\}.\end{aligned}$$

In a similar way as for the class defined in (4), we can show that the class $\mathcal{F}$ is Donsker, and hence by Theorem 1 the process $n^{1/2}(\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1))$ $(t_1 \in \mathbb{R}^{d_x}, t_2 \in \mathbb{R})$ converges weakly to a Gaussian process, with mean function given by $C_1^{p+1} b(t_2|t_1)$.
(ii) It can be easily seen that

$$\widehat{S}^m_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S^m_{\varepsilon_2|\varepsilon_1}(t_2|t_1) = m\, S^{m-1}_{\varepsilon_2|\varepsilon_1}(t_2|t_1)(\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1)) + o_P(n^{-1/2}).$$

Hence,

$$\begin{aligned}&n^{1/2}(\widehat{\varphi}_m(t_1) - \varphi_m(t_1)) \\ &= m\, n^{1/2} \int S^{m-1}_{\varepsilon_2|\varepsilon_1}(t_2|t_1)(\widehat{S}_{\varepsilon_2|\varepsilon_1}(t_2|t_1) - S_{\varepsilon_2|\varepsilon_1}(t_2|t_1))\, dt_2 + o_P(1),\end{aligned}$$

and hence the result follows from part (i).
(iii) Note that, uniformly in x,

$$\begin{aligned}&\widehat{\tau}_m(x,z) - \tau_m(x,z) \\ &= \widehat{\mu}_2(z) - \mu_2(z) + \varphi_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)[\widehat{\sigma}_2(z) - \sigma_2(z)] \\ &\quad + \Big[\varphi_m\Big(\frac{x - \widehat{\mu}_1(z)}{\widehat{\sigma}_1(z)}\Big) - \varphi_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)\Big]\sigma_2(z) + o_P((nh^{d_z})^{-1/2}) \\ &= \widehat{\mu}_2(z) - \mu_2(z) + \varphi_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)[\widehat{\sigma}_2(z) - \sigma_2(z)] \\ &\quad -\varphi'_m\Big(\frac{x - \mu_1(z)}{\sigma_1(z)}\Big)\Big[\frac{\widehat{\mu}_1(z) - \mu_1(z)}{\sigma_1(z)} + \frac{x - \mu_1(z)}{\sigma_1^2(z)}(\widehat{\sigma}_1(z) - \sigma_1(z))\Big]\sigma_2(z) \\ &\quad + o_P((nh^{d_z})^{-1/2}),\end{aligned}$$

by part (*ii*). Hence, the result follows from the i.i.d. expansions of $\widehat{\mu}_j(z)$ and $\widehat{\sigma}_j(z)$ ($j = 1, 2$) given in Section 4.2. □

Proof of Theorem 3. For simplifying the presentation suppose that Z is one-dimensional (the general case where $d_z \geq 1$ can be treated similarly, but is technically a bit more involved). Let

$$\begin{aligned} i_1 &= \text{argmin}_j Z_j, \\ i_2 &= \text{argmin}_{j:Z_j - Z_{i_1} > 2h}(Z_j - Z_{i_1}), \\ i_3 &= \text{argmin}_{j:Z_j - Z_{i_2} > 2h}(Z_j - Z_{i_2}), \\ &\vdots \end{aligned}$$

define $V = \{i_1, i_2, i_3, \ldots\}$, and denote the cardinality of V by $r_n = O(h^{-1})$. By construction, the variables $\widehat{\varepsilon}_{ji_1}, \widehat{\varepsilon}_{ji_2}, \ldots, \widehat{\varepsilon}_{ji_{r_n}}$ are mutually independent (for $j = 1, 2$). Also, note that $0 \leq \widehat{\varphi}(t_1) - \varphi(t_1) \leq \min\{\widehat{\varepsilon}_{2i} : i \in V, \widehat{\varepsilon}_{1i} \geq t_1\} - \varphi(t_1)$.

Now, fix $s > 0$ and consider

$$\begin{aligned} & P\big(\widehat{\varphi}(t_1) - \varphi(t_1) \leq s\big) \\ & \geq P\big(\min\{\widehat{\varepsilon}_{2i} : i \in V, \widehat{\varepsilon}_{1i} \geq t_1\} \leq \varphi(t_1) + s\big) \\ & = 1 - P\big(\widehat{\varepsilon}_{1i_1} < t_1 \text{ or } \widehat{\varepsilon}_{2i_1} > \varphi(t_1) + s\big) \ldots P\big(\widehat{\varepsilon}_{1i_{r_n}} < t_1 \text{ or } \widehat{\varepsilon}_{2i_{r_n}} > \varphi(t_1) + s\big). \qquad (5) \end{aligned}$$

We will show that

$$p := \max_{k=1,\ldots,r_n} P\big(\widehat{\varepsilon}_{1i_k} < t_1 \text{ or } \widehat{\varepsilon}_{2i_k} > \varphi(t_1) + s\big) < 1. \qquad (6)$$

From (6) it will follow that (5), which is greater than or equal to $1 - p^{r_n}$, converges to one, since $r_n \to \infty$ as $n \to \infty$. For $\delta_n \to 0$, write

$$\begin{aligned} & P\big(\widehat{\varepsilon}_{1i_k} < t_1 \text{ or } \widehat{\varepsilon}_{2i_k} > \varphi(t_1) + s\big) \\ & \leq P\Big(\big\{\widehat{\varepsilon}_{1i_k} < t_1 \text{ or } \widehat{\varepsilon}_{2i_k} > \varphi(t_1) + s\big\} \text{ and } \sup_{z,j=1,2} |\widehat{\mu}_j(z) - \mu_j(z)| \leq \delta_n\Big) \\ & \quad + P\big(\sup_{z,j=1,2} |\widehat{\mu}_j(z) - \mu_j(z)| > \delta_n\big) \\ & \leq P\big(\varepsilon_{1i_k} < t_1 + \delta_n \text{ or } \varepsilon_{2i_k} > \varphi(t_1) + s - \delta_n\big) + \nu_n \\ & \leq P\big(\varepsilon_{1i_k} < t_1 + \delta_n \text{ or } \varepsilon_{2i_k} > \varphi(t_1 + \delta_n) + \frac{s}{2}\big) + \nu_n \\ & := 1 - q + \nu_n \\ & \leq 1 - \frac{q}{2} < 1, \end{aligned}$$

for ν_n sufficiently small, since $q > 0$. This shows (6), and hence $\widehat{\varphi}(t_1) - \varphi(t_1) \overset{P}{\to} 0$.

Next, consider

$$\begin{aligned}
&\widehat{\tau}(x,z) - \tau(x,z) \\
&= \widehat{\mu}_2(z) - \mu_2(z) + \varphi\Big(\frac{x-\mu_1(z)}{\sigma_1(z)}\Big)(\widehat{\sigma}_2(z) - \sigma_2(z)) \\
&\quad \Big[\widehat{\varphi}\Big(\frac{x-\widehat{\mu}_1(z)}{\widehat{\sigma}_1(z)}\Big) - \varphi\Big(\frac{x-\mu_1(z)}{\sigma_1(z)}\Big)\Big]\sigma_2(z) + o_P(1),
\end{aligned}$$

and this converges in probability to zero thanks to the weak consistency of $\widehat{\mu}_2(z)$ and $\widehat{\sigma}_2(z)$, and provided we can show that

$$\widehat{\varphi}\Big(\frac{x-\widehat{\mu}_1(z)}{\widehat{\sigma}_1(z)}\Big) - \varphi\Big(\frac{x-\mu_1(z)}{\sigma_1(z)}\Big) \overset{P}{\to} 0. \tag{7}$$

By following the same proof as for $\widehat{\varphi}(t_1) - \varphi(t_1)$, it is easily seen that (7) holds true, using the weak consistency of $\widehat{\mu}_1(z)$ and $\widehat{\sigma}_1(z)$. □

Lemma 4 *Assume that model (5) is valid, i.e. assume that $(\varepsilon_1, \varepsilon_2)$ is independent of Z. Let $\widetilde{\mu}_1$ and $\widetilde{\mu}_2$ be arbitrary location functions, let $\widetilde{\sigma}_1$ and $\widetilde{\sigma}_2$ be arbitrary scale functions, and define*

$$\widetilde{\varepsilon}_1 = \frac{X - \widetilde{\mu}_1(Z)}{\widetilde{\sigma}_1(Z)} \quad \textit{and} \quad \widetilde{\varepsilon}_2 = \frac{Y - \widetilde{\mu}_2(Z)}{\widetilde{\sigma}_2(Z)}.$$

Then, $(\widetilde{\varepsilon}_1, \widetilde{\varepsilon}_2)$ is also independent of Z.

Proof. First, by definition of a location and scale function, we can write $\widetilde{\mu}_2(z) = T(F_{Y|Z}(\cdot|z))$ and $\widetilde{\sigma}_2(z) = S(F_{Y|Z}(\cdot|z))$ for some functionals T and S, that satisfy

$$T(F_{aY+b|Z}(\cdot|z)) = aT(F_{Y|Z}(\cdot|z)) + b \;\text{ and }\; S(F_{aY+b|Z}(\cdot|z)) = aS(F_{Y|Z}(\cdot|z)),$$

for all $a \geq 0$ and $b \in \mathbb{R}$, where for the purpose of this proof, we use the notations $F_{Y|Z}(\cdot|z)$, respectively $F_Y(\cdot)$, for the conditional (on the vector Z), respectively marginal, distribution of any random variable Y (see also Huber (1981), p. 59, 202).

Write

$$\widetilde{\varepsilon}_2 = \frac{\mu_2(Z) - \widetilde{\mu}_2(Z)}{\widetilde{\sigma}_2(Z)} + \frac{\sigma_2(Z)}{\widetilde{\sigma}_2(Z)}\varepsilon_2.$$

In a first step, we will show that $\sigma_2(Z)/\widetilde{\sigma}_2(Z)$ is independent of Z. We have that

$$\widetilde{\sigma}_2(z) = S\Big(F_{\mu_2(z)+\sigma_2(z)\varepsilon_2|Z}(\cdot|z)\Big) = \sigma_2(z)S\Big(F_{\varepsilon_2|Z}(\cdot|z)\Big) = \sigma_2(z)S\Big(F_{\varepsilon_2}(\cdot)\Big),$$

for arbitrary z, i.e. $\sigma_2(z)/\widetilde{\sigma}_2(z)$ is the same for all values of z. In a similar way we can show that $(\mu_2(z)-\widetilde{\mu}_2(z))/\widetilde{\sigma}_2(z)$, $\sigma_1(z)/\widetilde{\sigma}_1(z)$ and $(\mu_1(z)-\widetilde{\mu}_1(z))/\widetilde{\sigma}_1(z)$ are also independent of z. It now follows that

$$\begin{aligned}
&P(\widetilde{\varepsilon}_1 \le t_1, \widetilde{\varepsilon}_2 \le t_2 | Z) \\
&= P\Big(\frac{\mu_1(Z)-\widetilde{\mu}_1(Z)}{\widetilde{\sigma}_1(Z)} + \frac{\sigma_1(Z)}{\widetilde{\sigma}_1(Z)}\varepsilon_1 \le t_1, \frac{\mu_2(Z)-\widetilde{\mu}_2(Z)}{\widetilde{\sigma}_2(Z)} + \frac{\sigma_2(Z)}{\widetilde{\sigma}_2(Z)}\varepsilon_2 \le t_2 \Big| Z\Big) \\
&= P(A_1 + A_2\varepsilon_1 \le t_1, C_1 + C_2\varepsilon_2 \le t_2 | Z) \\
&= P(\widetilde{\varepsilon}_1 \le t_1, \widetilde{\varepsilon}_2 \le t_2),
\end{aligned}$$

where $A_1 := (\mu_1(z)-\widetilde{\mu}_1(z))/\widetilde{\sigma}_1(z)$, $A_2 := \sigma_1(z)/\widetilde{\sigma}_1(z)$ $C_1 := (\mu_2(z)-\widetilde{\mu}_2(z))/\widetilde{\sigma}_2(z)$ and $C_2 := \sigma_2(z)/\widetilde{\sigma}_2(z)$ for all z, and where the third equality follows from the fact that $(\varepsilon_1, \varepsilon_2)$ is independent of Z. □

References

Akritas, M.G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, **28**, 549-568.

Aly, H.Y., Grabowski, R., Pasurka, C. and Rangan, N. (1990). Technical, scale, and allocative efficiencies in U.S. banking: An empirical investigation. *Review of Economics and Statistics*, **72**, 211-218.

Aragon, Y., Daouia, A. and Thomas-Agnan, C. (2005). Nonparametric frontier estimation: a conditional quantile-based approach. *Econometric Theory*, **21**, 358-389.

Bădin, L., Daraio, C. and Simar, L. (2010). Optimal bandwidth selection for conditional efficiency measures: a data-driven approach. *European Journal of Operational Research*, **201**, 633-640.

Bădin, L., Daraio, C. and L. Simar (2011), How to measure the impact of environmental factors in a nonparametric production model? Discussion paper 2011/19, Institut de Statistique, UCL.

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

Daouia, A., Florens, J.-P. and Simar, L. (2010). Frontier estimation and extreme value theory. *Bernoulli*, **16**, 1039-1063.

Daouia, A. and Gijbels, I. (2011). Robustness and inference in nonparametric partial-frontier modeling. *Journal of Econometrics* (in press).

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

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

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

Daraio, C., Simar, L. and Wilson, P.W. (2010). Testing whether two-stage estimation is meaningful in non-parametric models of production. Discussion paper 1031, Institut de Statistique, UCL.

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

Dette, H., Neumeyer, N. and Van Keilegom, I. (2007). A new test for the parametric form of the variance function in non-parametric regression. *Journal of the Royal Statistical Society - Series B*, **69**, 903-917.

Einmahl, J. and Van Keilegom, I. (2008a). Specification tests in nonparametric regression. *Journal of Econometrics*, **143**, 88-102.

Einmahl, J. and Van Keilegom, I. (2008b). Tests for independence in nonparametric regression. *Statistica Sinica*, **18**, 601-616.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall, London.

Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society - Series A*, **120**, 253-290.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

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

Masry, E. (1997). Multivariate regression estimation: local polynomial fitting for time series. *Nonlinear Analysis, Theory, Methods & Applications*, **30**, 3575-3581.

Neumeyer, N. (2006). *Bootstrap Procedures for Empirical Processes of Nonparametric Residuals.* Habilitationsschrift, Fakultät für Mathematik, Ruhr-Universität Bochum, Bochum.

Neumeyer, N. (2009a). Smooth residual bootstrap for empirical processes of nonparametric regression residuals. *Scandinavian Journal of Statistics*, **36**, 204-228.

Neumeyer, N. (2009b). Testing independence in nonparametric regression. *Journal of Multivariate Analysis*, **100**, 1551-1566.

Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, **101**, 1067-1078.

Park, B., Simar, L. and Weiner, Ch. (2000). The FDH estimator for productivity efficiency scores: asymptotic properties. *Econometric Theory*, **16**, 855-877.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346-1370.

Shephard, R.W. (1970). *Theory of Cost and Production Function.* Princeton University Press, Princeton, New-Jersey.

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

Simar, L. and Wilson, P.W. (2008). Statistical inference in nonparametric frontier models: recent developments and perspectives. In: *The Measurement of Productive Efficiency*, 2nd Edition, H. Fried, C.A.Knox Lovell and S. Schmidt (Eds.), Oxford University Press.

Simar, L. and Wilson, P.W. (2011). Two-stage DEA: *Caveat emptor. Journal of Productivity Analysis* (in press).