Fast and efficient computation of directional distance estimators

Cinzia Daraio, Léopold Simar & Paul W. Wilson

Annals of Operations Research

ISSN 0254-5330
Volume 288
Number 2

Ann Oper Res (2020) 288:805-835
DOI 10.1007/s10479-019-03163-9

Springer

Annals of Operations Research (2020) 288:805–835
https://doi.org/10.1007/s10479-019-03163-9

S.I.: DATA ENVELOPMENT ANALYSIS: FOUR DECADES ON

Fast and efficient computation of directional distance estimators

Cinzia Daraio[1] · Léopold Simar[1,2] · Paul W. Wilson[3]

Published online: 11 February 2019

Abstract
Directional distances provide useful, flexible measures of technical efficiency of production units relative to the efficient frontier of the attainable set in input-output space. In addition, the additive nature of directional distances permits negative input or outputs quantities. The choice of the direction allows analysis of different strategies for the units attempting to reach the efficient frontier. Simar et al. (Eur J Oper Res 220:853–864, 2012) and Simar and Vanhems (J Econom 166:342–354, 2012) develop asymptotic properties of full-envelopment, FDH and DEA estimators of directional distances as well as robust order-m and order-α directional distance estimators. Extensions of these estimators to measures conditioned on environmental variables Z are also available (e.g., see Daraio and Simar in Eur J Oper Res 237:358–369, 2014). The resulting estimators have been shown to share the properties of their corresponding radial measures. However, to date the algorithms proposed for computing the directional distance estimates suffer from various numerical drawbacks (Daraio and Simar in Eur J Oper Res 237:358–369, 2014). In particular, for the order-m versions (conditional and unconditional) only approximations, based on Monte-Carlo methods, have been suggested, involving additional computational burden. In this paper we propose a new fast and efficient method to compute *exact* values of the directional distance estimates for all the cases (full and partial frontier cases, unconditional or conditional to external factors), that overcome all previous difficulties. This new method is illustrated on simulated and real data sets. Matlab code for computation is provided in an "Appendix".

Keywords Directional distances · Conditional efficiency · Robust frontiers · Environmental factors · Nonparametric methods

1 Introduction

Production theory and efficiency analysis examine how production units (i.e., Decision Making Units or DMUs) transform quantities of inputs (e.g., labor, energy and capital) into quantities of outputs (e.g., goods and services). The technical efficiency of a particular unit

Cinzia Daraio: Financial support from the Italian Ministry of Education and Research (PRIN Project No. 2015RJARX7), from Sapienza University of Rome (Sapienza Awards No. 6H15XNFS), and of the Lazio Region (Project FILAS-RU-2014-1186) is gratefully acknowledged.

Extended author information available on the last page of the article

is then measured by distance in some direction from the unit's location in input-output space to the technology, i.e., the frontier of the production set.

Traditional nonparametric efficiency estimators based on radial contractions of inputs or radial expansions of outputs to reach the frontier have been proposed by Farrell (1957), Charnes et al. (1978) and Deprins et al. (1984).[1] More recently, estimators of directional distance efficiency have been proposed by Chambers et al. (1996, 1998). The directional measures of efficiency and their corresponding estimators nest the input and output-oriented versions of the original DEA and FDH estimators, but also permit estimation of efficiency along other paths to the frontier. In addition, the directional estimators permit negative values of input or output quantities, unlike the earlier radial estimators. This enhanced flexibility has made directional measures and their estimators popular in recent years.

The conditional efficiency estimators based on FDH and DEA have been extended to robust order-m and order-α type estimators; see Daraio and Simar (2014) for an introduction and Simar and Wilson (2013, 2015) for comprehensive summaries. These robust estimators are based on the idea of estimating distance from a given DMU's position in input-output space to a *partial frontier* lying "close" to the full frontier (i.e., the boundary of the production set). Partial frontiers provide an alternative benchmark, and provide advantages over the full-envelopment FDH and DEA estimators in terms of the resulting statistical properties. Inclusion of environmental variables may reflect heterogeneity of the DMUs and their operating environments. Environmental variables are neither inputs nor outputs, but instead are external (to the DMU) factors that may affect the performance of the units. Efficiency estimates are *conditioned* on these variables in the sense that efficiency is estimated *given the environment described by the environmental variables.* Bădin et al. (2014) provide an overview.

The statistical properties of both conditional and unconditional directional distance estimators have been derived by Simar and Vanhems (2012) and Simar et al. (2012) for both the full-envelopment and robust, partial frontier cases. However, as observed by Daraio and Simar (2005), computation of directional distance estimates is problematic due to numerical issues as well as a substantial computational burden due to reliance on Monte-Carlo approximations required to computed the estimates.

This paper provides a new, fast and efficient method to compute exact values of the directional distance estimates for all the cases (i.e., both the full frontier case as well as the robust, partial-frontier cases, and both conditional or unconditional cases). The new method eliminates the need for Monte-Carlo approximations and provides exact solutions. This avoids the substantial computational burden that has been incurred until now. In addition, the new method avoids numerical problems that can arise in applications when the previous computational methods are used in applications. This new method is illustrated on both simulated and real data, and Matlab code is provided for use by practitioners.

The results provided in this paper are relevant to practitioners, in particular because the robust directional distance estimators (both conditional and unconditional) are widely used. Conditional efficiency analyses have been applied to carry out innovation studies at regional level (Broekel 2012) and environmental analyses at both national (Halkos and Tzeremes 2014; Halkos et al. 2016; Halkos and Managi 2016; Manello 2017) as well as regional levels accounting for governance issues (Halkos et al. 2015) and growth (Halkos et al. 2016; Halkos and Tzeremes 2013b). Applications in agriculture (Serra and Lansink 2014) include the efficiency of family firms (Baležentis and De Witte 2015) and the analysis of the effect of public subsidies on farm efficiency (Minviel and De Witte 2017). Examples of applications

[1] The Data Envelopment Analysis (DEA) estimators proposed by Farrell (1957) and Charnes et al. (1978) impose convexity on the production set, while the Free Disposal Hull (FDH) estimator proposed by Deprins et al. (1984) does not.

in the financial sector include Mallick et al. (2016), Matousek and Tzeremes (2016) and Tzeremes (2015). Other interesting applications examine libraries (De Witte and Geys 2011), primary schools (Cordero et al. 2017a, b), secondary schools (Haelermans and De Witte 2012), municipalities (Cordero et al. 2017a, b), the health care sector (Varabyova et al. 2016; Varabyova and Schreyögg 2017; Ferreira et al. 2018), water utilities (Zschille 2015), waste management (Fuentes et al. 2015; Guerrini et al. 2016), culture and eco-efficiency (Halkos and Tzeremes 2013a) and local police departments (Verschelde and Rogge 2012).

The paper is organized as follows. The next section introduces the basic concepts and notation and provides an outline of the issues addressed by the paper. Section 3 presents the full frontier cases distinguishing between unconditional and conditional analyses, Sects. 4 and 5 analyze the partial frontier approaches, presenting again for each of them the unconditional and conditional cases. Section 6 reports the outcome of the application of the new proposed method for computing directional distances to simulated as well as real data. Section 7 provides conclusions and a brief summary of the main results. Matlab code implementing the new computational method is provided in "Appendix".

2 Statistical framework and notation

This section introduces the basic concepts and notation needed to present the new computational methods for directional distances in the various cases of interest. We first summarize the concepts of directional distance functions and their conditional versions which allow analysis of possible heterogeneity due to some environmental factors. We then give the intuition behind the robust partial frontiers (order-α and order-m) in the context of directional distances. Finally, we discuss the drawbacks of the existing algorithms for computing these various directional distances and the need for the new computational methods provided later in this paper.

2.1 Directional distances and their probabilistic formulation

Consider a production process in which p inputs are used to produce q outputs. The production set

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q} \mid x \text{ can produce } y\} \tag{2.1}$$

is the set of technically feasible combinations of inputs and outputs. The efficient frontier of Ψ is defined by

$$\Psi^{\partial} = \{(x, y) \in \Psi \mid (\gamma^{-1}x, \gamma y) \notin \Psi \ \forall\, \gamma > 1\}. \tag{2.2}$$

Traditional approaches to efficiency measurement based on the ideas of Farrell (1957), Debreu (1951) and Shephard (1970) involve measuring the distance from a production plan (x, y) to the efficient frontier Ψ^{∂} in either the input or output direction by considering either the maximum feasible, proportionate reduction in input quantities (without lowering any output) or the maximum feasible, proportionate increase in output quantities (without raising any input). With these *radial* measures of efficiency, only non-negative values of input and output quantities can be accommodated.

Nonparametric estimators of the attainable set Ψ are often based on envelopment of the cloud of observed points $\mathcal{X}_n = \{(X_i, Y_i)\}_{i=1}^n$. The Free Disposal Hull (FDH), suggested by Deprins et al. (1984), only assumes free disposability of both inputs and outputs, whereas the

Data Envelopment Analysis (DEA) estimators proposed by Farrell (1957) and popularized by Charnes et al. (1978) assume convexity of Ψ as well as free disposability of inputs and outputs. The properties of the resulting estimators of efficiency measures, in the radial cases, have been derived in Park et al. (2000) for the FDH case and in Kneip et al. (2008) for the DEA with varying returns to scale and Park et al. (2010) for the DEA with constant returns to scale. It is now well-known that these estimators suffer from the "curse of dimensionality." When the dimension $p+q$ increases, the rates of convergence become slower. For individual efficiency measures, bootstrap techniques are required to make inference, estimate bias and estimate confidence intervals. For more details, see the recent surveys by Simar and Wilson (2013, 2015) and the references therein.

Directional distances introduced by Chambers et al. (1996) and discussed by Färe and Grosskopf (2004) provide useful and flexible ways to measure technical efficiency of units relative to the efficient frontier. The directional distance function

$$\beta(x, y \mid d_x, d_y) = \sup\{\beta > 0 \mid (x - \beta d_x, y + \beta d_y) \in \Psi\} \tag{2.3}$$

projects the input-output vector (x, y) onto the technology in a direction specified by a vector $d = (d_x, d_y) \geq 0$. The choice of the directions d_x and d_y for measuring the distance from the unit operating at $(x, y) \in \Psi$ to the frontier allows analysis of different strategies for the units to reach the efficient frontier. Note that some directions (but not all) can be set equal to zero, indicating the components of X and Y that are "inactive" in the optimization described in (2.3). For instance is the vector $d_x = 0$, and if all the outputs take positive values, then the Farrell–Debreu radial output efficiency measure is given by $1 + \beta(x, y \mid 0, y)$. Alternatively, in the input orientation, if all the inputs take positive values, the Farrell–Debreu radial efficiency is given by $1 - \beta(x, y \mid x, 0)$. Note that the additive nature of directional distances allows to treat negative inputs and outputs, which is not the case for radial distances.

In practice all these quantities are unknown and must be estimated from a sample of observations $\mathcal{X}_n = \{(X_i, Y_i)\}_{i=1}^n$. Therefore, in order to evaluate the properties of the resulting estimates, and to make inference, a statistical model is required. We adopt the probabilistic formulation of Cazals et al. (2002) and extended by Daraio and Simar (2007). The production process is characterized by the process that generates a vector of inputs and outputs defined over an appropriate probability space. Let $X \in \mathbb{R}^p$ denote a p-vector of inputs and $Y \in \mathbb{R}^q$ denote a q-vector of outputs. The joint distribution of (X, Y) has support over Ψ. Now consider the joint probability $H_{XY}(x, y) = \Pr(X \leq x, Y \geq y)$, which is the probability of finding a unit (X, Y) dominating the point (x, y). As shown by Cazals et al. (2002), under the free disposability assumption[2]

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q} \mid H_{XY}(x, y) > 0\}. \tag{2.4}$$

Simar and Vanhems (2012) show that under free disposability,

$$\beta(x, y \mid d_x, d_y) = \sup\{\beta > 0 \mid H_{XY}(x - \beta d_x, y + \beta d_y) > 0\}. \tag{2.5}$$

Nonparametric estimators of the attainable set are typically obtained by envelopment techniques. Simar and Vanhems (2012)) and Simar et al. (2012) show that the resulting estimators of the directional distances share properties similar to those of the radial measures.

In this paper we will focus on the FDH family of estimators, without imposing convexity of the attainable set. In this case, it can be shown that the FDH estimator of $\beta(x, y \mid d_x, d_y)$ can also be obtained by plugging (2.5) into the empirical version of H_{XY} given by

[2] Free disposability of inputs and outputs means that if $(x, y) \in \Psi$, then $(\widetilde{x}, \widetilde{y}) \in \Psi$ for all $(\widetilde{x}, \widetilde{y})$ such that $\widetilde{x} \geq x$ and $\widetilde{y} \leq y$. In a sense, it assumes the possibility of wasting resources.

$$\widehat{H}_{n,XY}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y), \tag{2.6}$$

where $\mathbb{1}(\cdot)$ is the indicator function [$\mathbb{1}(a) = 1$ if a is true and 0 otherwise]. We will see below how to implement this in practice, in particular when some (but not all) elements of (d_x, d_y) are set at zero.

2.2 Introducing environmental variables

The probabilistic characterization of the production process defined above allows quite naturally introduction of environmental factors into the process. Consider the case where external environmental variables $Z \in \mathcal{Z} \subset \mathbb{R}^r$ represent heterogeneity factors that may influence the production process. To accommodate these variables, the probability space considered so far has to be augmented. We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the random variables X, Y, Z are defined and we denote by $\mathcal{P}$ the support of the joint distribution of (X, Y, Z). Let Ψ^z denote the support of (X, Y) given $Z = z$. Thus the attainable set for firms facing external conditions $Z = z$ is given by

$$\begin{aligned}\Psi^z &= \{(x, y) \in \mathbb{R}^{p+q} \mid x \text{ can produce } y \text{ if } Z = z\} \\ &= \{(x, y) \in \mathbb{R}^{p+q} \mid H_{XY|Z}(x, y \mid z) > 0\},\end{aligned} \tag{2.7}$$

where $H_{XY|Z}(x, y \mid z) = \Pr(X \leq x, Y \geq y \mid Z = z)$. The variables in Z can affect the production process either (i) through Ψ^z the support of (X, Y), (ii) through the conditional distribution of (X, Y) given Z, affecting e.g. only the probability of a firm to reach its optimal boundary, or (iii) through both (i) and (ii).

It is easy to see that $\Psi = \bigcup_{z\in\mathcal{Z}} \Psi^z$, so that $\Psi^z \subseteq \Psi$, for all $z \in \mathcal{Z}$. In the very particular case where the joint support of (X, Y, Z) can be written as the Cartesian product $\mathcal{P} = \Psi \times \mathcal{Z}$, Z has no impact on the boundaries of Ψ and $\Psi^z = \Psi$ for all $z \in \mathcal{Z}$ (this is called the "separability condition" in the literature; e.g., see Simar and Wilson 2007, 2011). In the latter case, Z may eventually influence the production process only through the probability of reaching its optimal boundary. Daraio et al. (2018) provide a procedure for testing this separability condition.

Now we can define the conditional directional distance function

$$\beta(x, y \mid d_x, d_y, z) = \sup\{\beta > 0 \mid H_{XY|Z}(x - \beta d_x, y + \beta d_y \mid z) > 0\}. \tag{2.8}$$

Here again, we can recover the conditional version of the radial Farrell–Debreu measures by the appropriate choice of the distance vector. A nonparametric estimator of $\beta(x, y \mid d_x, d_y, z)$ is obtained from a sample $\mathcal{X}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ by plugging a nonparametric estimator of $H_{XY|Z}(x, y \mid z)$ into (2.8). This conditional version requires smoothing over the values of Z_i in a neighborhood of z since observations with exact values $Z_i = z$ are typically not available. We use the empirical, localized analog of $H_{XY|Z}(x, y \mid z)$ given by

$$\widehat{H}_{n,XY|Z}(x, y \mid Z = z) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x, Y_i \geq y) K_h(Z_i, z)}{\sum_{i=1}^n K_h(Z_i, z)}, \tag{2.9}$$

where $K_h(Z_i, z)$ are appropriate kernel functions (with compact support) and h is a vector of r bandwidths, one for each component of z. In fact, it can be shown that the conditional FDH estimator is a localized version of the unconditional case, where the localization is tuned by the bandwidths h. The estimator is the FDH estimator computed over the subsample of observations $i = 1, \ldots, n$ such that $||Z_i - z|| \leq h$ (see Cazals et al. 2002; Daraio and

Simar 2005 for details).[3] The asymptotic properties of the resulting estimators of the radial conditional measures have been established in Jeong et al. (2010) and adapted to directional distances in Simar and Vanhems (2012). To summarize, we keep similar properties as in the unconditional case but with a reduced number of observations: n is replaced by $n\prod_{j=1}^{r} h^{(j)}$.

Conditioning on Z requires the determination of an r-vector of bandwidths. For the radial oriented efficiency scores, Bădin et al. (2010) suggest adapting least squares cross validation techniques from the literature. However, Simar et al. (2016) show (see their Appendix B) that better monotonicity properties of the resulting efficiency estimates are achieved by searching for the optimal bandwidth when estimating the joint probability $H_{XY|Z}(x, y \mid z)$.[4] Here direct methods suggested by Li et al. (2013) can be used (Matlab code for this purpose is provided by Bădin et al. (2018). A detailed methodology on how to analyze the effect of Z on the production process has been proposed by Bădin et al. (2012, 2014). Simar et al. (2016) show how to adapt the approach when Z is latent and hence unobserved. This requires an additional model and an instrument to identify Z.

2.3 Partial frontiers: robust approaches

Nonparametric FDH and DEA estimators are envelopment estimators in the sense that the corresponding estimate of Ψ (or of Ψ^z) envelops the cloud of observed data points. Consequently, these estimators are highly sensitive to extreme data points and outliers. This provides the major interest in the robust version of these estimators developed for radial measures by Cazals et al. (2002), Aragon et al. (2005) and Daouia and Simar (2007). Simar and Vanhems (2012) extend these concepts to directional distances. In all cases, the idea is to define a less-extreme boundary to use as a benchmark, i.e. to define a *partial frontier* in contrast to the full frontier used above. By construction, some data points may lie outside the partial-frontier, but nonetheless the partial frontier provides a useful benchmark for evaluating efficiency. Two classes of partial frontiers have been suggested in the literature: the order-α quantile frontier and the order-m partial frontier. In this summary we give only some intuitive definitions for the case of one output and with the output orientation (e.g. $d_x = 0$ and $d_y = 1$) for the unconditional case where Z does not play a role. In the remaining sections of the paper we will derive expressions for the most general cases.

For any $\alpha \in (0, 1]$ the directional distance of order-α is given by

$$\beta_\alpha(x, y \mid 0, 1) = \sup\{\beta \mid S_{Y|X}(y + \beta \mid x) > 1 - \alpha\}, \tag{2.10}$$

where $S_{Y|X}(y \mid x) = \Pr(Y \geq y \mid X \leq x) = H_{XY}(x, y)/F_X(x)$ is the conditional survival function of Y given $X \leq x$. Note that if $\alpha \to 1$, we are back the usual full frontier measure [for $d = (0, 1)$]. So for $\alpha < 1$, the benchmark frontier for the unit (x, y) [i.e. where $\beta_\alpha(x, y \mid 0, 1) = 0$] corresponds to the α-quantile of the conditional distribution of the output among the population of units using less inputs than x.

Then the partial order-α frontier (or the "order-α quantile frontier") is given by

$$\varphi_\alpha(x) = y + \beta_\alpha(x, y \mid 0, 1), \tag{2.11}$$

where y can be any value in the support of $S_{Y|X}(\cdot \mid x)$. Note that $\beta_\alpha(x, y \mid 0, 1)$ can take negative values if y is large and hence this unit lies above the conditional quantile frontier of order-α.

[3] The inequality $||Z_i - z|| \leq h$ has to be understood component by component $|Z_i^{(j)} - z^{(j)}| \leq h^{(j)}$.

[4] See the discussion in Footnote 6 below.

The order-m frontier in the same (output) orientation can be defined for any integer m as

$$\varphi_m(x) = \mathbb{E}\left[\max(Y_1, \ldots, Y_m) \mid X \leq x\right], \tag{2.12}$$

where the Y_j are independent, identically distributed (iid) realization of the output Y, conditionally on $X \leq x$. Here, as $m \to \infty$, we are back to the usual full-frontier measure. So the benchmark frontier is the expected value of the maximum output among m peers drawn from the population of units using less inputs than x. It can be shown that when Y takes only positive values,

$$\varphi_m(x, y) = \int_0^\infty [1 - (1 - S_{Y|X}(y \mid x))^m] dy. \tag{2.13}$$

Note again that $\beta_m(x, y \mid 0, 1) = \varphi_m(x) - y$ can take negative values for large values of y.

Nonparametric estimators are obtained by plugging in the empirical versions of the conditional survival function. They share interesting properties, in particular, and by contrast to the full frontier estimates, they achieve the parametric $\sqrt{n}$-rate of convergence independently of the dimension of the problem $(p + q)$. The statistical properties of the order-m estimators have been established by Cazals et al. (2002) and for the order-α cases by Daouia and Simar (2007). These include the conditional to Z cases. Simar and Vanhems (2012) extend these to the directional distances cases.

We will provide below general expressions for evaluating the directional distance to these partial frontiers (conditional and unconditional to Z) and their estimators. Recall that their robustness properties rely on the fact that for large α (or m) we estimate a partial frontier not far from the full frontier, but for $\alpha < 1$ and finite m, the estimators will not envelop all the data points and so are robust to extreme data points and outliers. Comparisons of the two concepts from a robustness point of view can be found in Daouia and Ruiz-Gazen (2006) and Daouia and Gijbels (2011a). Daouia et al. (2010, 2012) show how these partial frontiers can be used for estimating the full frontier, letting $\alpha \to 1$ and $m \to \infty$ when $n \to \infty$ but at an appropriate rate.

2.4 Aim of the paper

After the summary in the preceding sections, we next focus on the computational issues of directional distance functions. Simar and Vanhems (2012) show the equivalence between directional distances and hyperbolic radial distances after a monotonic transformation of the coordinate space of the inputs and the outputs. To summarize, for the "active" variables (those with components in d being > 0), the transformation is defined as

$$X^* = \exp(X \oslash d_x) \quad \text{and} \quad Y^* = \exp(Y \oslash d_y), \tag{2.14}$$

where $\oslash$ is the Hadamard component-wise division of vectors. The "non-active" variables can remain as they are.

This transformation is useful for obtaining the theoretical properties of the resulting estimators but may create some numerical problems for their practical computations (Daraio and Simar 2014). The exponential transformation may provide huge numbers that have to be carefully handled to avoid numerical problems when handling ratios (which is typical in FDH approaches).[5] Also the log transformation at the end, for coming back to original units, may create other problems. Simar and Vanhems (2012) observe that this is particularly the

[5] Note e.g. that exp(50) is already of the order 5×10^{21}. So without rescaling the variables we may have numerical imprecision and overflow conditions on digital computers. Note that the corresponding elements of

case for the order-m estimators where the $\log(w)$ comes in an integral starting at $w = 0$. The integral is well defined but its numerical treatment can be difficult. So for the order-m estimators (either conditional on Z or unconditional on Z), only approximate solutions based on Monte-Carlo simulations have been proposed so far. These Monte-Carlo approximations are not easy to implement (especially for the conditional-on-Z case; see Daraio and Simar 2014 for discussion), and may involve substantial computational burden to achieve reasonable precision. In this paper we propose an alternative, but equivalent, formulation of the directional distances which avoids *all* of these drawbacks. In the same set up and for the different nonparametric robust conditional and unconditional cases covered by Simar and Vanhems (2012) and by Daraio and Simar (2005) we propose a fast and efficient formula for computing the directional distances, also in the most general cases where some components of d_x and of d_y might be equal to zero. In addition, we provide simple expressions for the exact computation of the order-m directional distances (both unconditional and conditional). The main difficulty is to provide a formulation capable of handling cases where some of the inputs or outputs are inactive (i.e., with d-elements equal to zero). This is important since it reflects one of the most interesting flexibility properties of the directional distance functions.

In the next sections we detail how our method can be applied in various scenarios, cases. "Appendix" provides the Matlab code implementing our methods.

3 Full frontier cases

3.1 Unconditional case

To fix the notation, and without loss of generality, let us partition $d_x = (d_{x_1}, d_{x_2})$, where $d_{x_2} = 0$ is of dimension $p_2 \geq 0$. Then $d_{x_1} > 0$ is of dimension $p_1 = p - p_2$. Note that d_{x_2} could be an empty vector with $p_2 = 0$. We use similar notational convention for the elements of $d_y = (d_{y_1}, d_{y_2})$ with $d_{y_2} = 0$ of dimension $q_2 \geq 0$. We partition all the inputs and outputs analogously, noting that X_2 and/or Y_2 could be empty vectors. So X_1 and Y_1 are the active variables in the optimization equations above.

Following Appendix B in Simar and Vanhems (2012), in the presence of inactive directions, the directional distance is defined as

$$\beta(x, y \mid d_x, d_y) = \sup\{\beta > 0 \mid H_{X_1Y_1|X_2Y_2}(x_1 - \beta d_{x_1}, y_1 + \beta d_{y_1} \mid x_2, y_2) > 0\}, \quad (3.1)$$

where $H_{X_1Y_1|X_2Y_2}(x_1, y_1 \mid x_2, y_2) = \Pr(X_1 \leq x_1, Y_1 \geq y_1 \mid X_2 \leq x_2, Y_2 \geq y_2)$ is the conditional probability of dominating (x_1, y_1) given that $X_2 \leq x_2, Y_2 \geq y_2$. This is in the spirit of the probabilistic characterization of the Farrell–Debreu concept of efficiency introduced by Cazals et al. (2002). For instance, in the pure output orientation $d_x = 0$ and $d_y > 0$, the efficient frontier for a unit (x, y) is given by the upper support of the conditional distribution of Y given $X \leq x$. Note also that for units where $H_{X_2Y_2}(x_2, y_2) > 0$, for the full frontier case, the directional distance may also be computed as

$$\beta(x, y \mid d_x, d_y) = \sup\{\beta > 0 \mid H_{XY}(x_1 - \beta d_{x_1}, x_2, y_1 + \beta d_{y_1}, y_2) > 0\}. \quad (3.2)$$

As explained below, the latter equivalence will not be valid for the robust versions of the frontiers where the conditioning on $X_2 \leq x_2, Y_2 \geq y_2$ has to be used, as in (3.1).

Footnote 5 continued
the direction vector have to be rescaled accordingly, to avoid misinterpretation of the resulting β. This creates additional potential for confusion or errors.

Directional distances are independent of the units of measurement as described in Färe et al. (2008) and formally proven in Appendix A in Simar and Vanhems (2012). The property can be stated as follows:

$$\beta(\theta\circ x, \lambda\circ y \mid \theta\circ d_x, \lambda\circ d_y) = \beta(x, y \mid d_x, d_y) \quad \text{for } \theta \in \mathbb{R}^p_+, \text{ and } \lambda \in \mathbb{R}^q_+, \tag{3.3}$$

where $\circ$ indicates the Hadamard product or component-wise multiplication of vectors. This property inspires the transformation of the variables that will make easy the characterization of directional distances and will facilitate the computation of their estimators. Consider first the case where all components of d are > 0. We have indeed as a consequence of (3.3) the following identity:

$$\beta(x, y \mid d_x, d_y) = \beta(x^\star, y^\star \mid i_p, i_q), \tag{3.4}$$

where i_k is a vector of ones of length k, $x^\star = x \oslash d_x$ and $y^\star = y \oslash d_y$. More generally, when some elements of d are zero, we consider the transformation

$$\begin{aligned} X_1^\star &= X_1 \oslash d_{x_1} \quad \text{and} \quad X_2^\star = X_2 \\ Y_1^\star &= Y_1 \oslash d_{y_1} \quad \text{and} \quad Y_2^\star = Y_2, \end{aligned} \tag{3.5}$$

which leads to

$$\beta(x, y \mid d_x, d_y) = \sup\{\beta > 0 \mid H_{X_1^\star Y_1^\star \mid X_2^\star Y_2^\star}(x_1^\star - \beta i_{p_1}, y_1^\star + \beta i_{q_1} \mid x_2^\star, y_2^\star) > 0\}, \tag{3.6}$$

where $H_{X_1^\star Y_1^\star \mid X_2^\star Y_2^\star}$ is the version of $H_{X_1 Y_1 \mid X_2 Y_2}$ in the new coordinate system.

Now the nonparametric estimator of the distance is obtained by plugging in the empirical version of $H_{X_1^\star Y_1^\star \mid X_2^\star Y_2^\star}$ in (3.6). This yields

$$\begin{aligned} \widehat{H}_{n, X_1^\star Y_1^\star \mid X_2^\star Y_2^\star}(x_1^\star, y_1^\star \mid x_2^\star, y_2^\star) &= \frac{\widehat{H}_{n, X_1^\star, X_2^\star Y_1^\star, Y_2^\star}(x_1^\star, x_2^\star, y_1^\star, y_2^\star)}{\widehat{H}_{n, X_2^\star Y_2^\star}(x_2^\star, y_2^\star)} \\ &= \frac{\sum_{i=1}^n \mathbb{1}(X_{1,i}^\star \le x_1^\star, X_{2,i}^\star \le x_2^\star, Y_{1,i}^\star \ge y_1^\star, Y_{2,i}^\star \ge y_2^\star)}{\sum_{i=1}^n \mathbb{1}(X_{2,i}^\star \le x_2^\star, Y_{2,i}^\star \ge y_2^\star)}. \end{aligned} \tag{3.7}$$

Some algebra leads to the following explicit formula for the FDH estimator of $\beta(x, y \mid d_x, d_y)$, namely

$$\begin{aligned} \widehat{\beta}(x, y \mid d_x, d_y) &= \sup\{\beta > 0 \mid \widehat{H}_{n, X_1^\star Y_1^\star \mid X_2^\star Y_2^\star}(x_1^\star - \beta i_{p_1}, y_1^\star + \beta i_{q_1} \mid x_2^\star, y_2^\star) > 0\}, \\ &= \max_{\{i \mid X_{2,i} \le x_2, Y_{2,i} \ge y_2\}} \left[\min_{\substack{k = 1 \ldots, p1 \\ \ell = 1, \ldots, q_1}} \left\{ x_1^{\star,k} - X_{1,i}^{\star,k}, Y_{1,i}^{\star,\ell} - y_1^{\star,\ell} \right\} \right], \end{aligned} \tag{3.8}$$

where for a vector a, a^j represents its jth component. The formulation (3.8) is easy to program in modern high-level languages like R or Matlab. The Matlab function `FDH_dirdist_new(x,y,dx,dy,X,Y)` in the Appendix computes $\widehat{\beta}(x, y \mid d_x, d_y)$ using the expression in (3.8).

3.2 Conditioning on environmental factors Z

Simar and Vanhems (2012) provide the conditional (on environmental factors $Z \in \mathbb{R}^r$), directional measure

$$\beta(x, y \mid d_x, d_y, z) = \sup\{\beta > 0 \mid H_{X_1Y_1|X_2Y_2Z}(x_1 - \beta d_{x_1}, y_1 + \beta d_{y_1} \mid x_2, y_2, z) > 0\}, \tag{3.9}$$

where $H_{X_1Y_1|X_2Y_2Z}(x_1, y_1 \mid x_2, y_2, z) = \Pr(X_1 \leq x_1, Y_1 \geq y_1 \mid X_2 \leq x_2, Y_2 \geq y_2, Z = z)$ is the conditional probability of dominating (x_1, y_1) given that $X_2 \leq x_2$, $Y_2 \geq y_2$ and $Z = z$, noting the difference in conditioning between the inactive inputs, the inactive outputs and the factors Z. This distribution is given by

$$H_{X_1Y_1|X_2Y_2Z}(x_1, y_1 \mid x_2, y_2, z) = \frac{H_{X_1,X_2,Y_1,Y_2|Z}(x_1, x_2, y_1, y_2 \mid z)}{H_{X_1,X_2,Y_1,Y_2|Z}(\infty, x_2, -\infty, y_2 \mid z)}. \tag{3.10}$$

The nonparametric estimator of the distribution in 3.10 requires smoothing in z using a kernel with compact support (see Daraio and Simar 2005). Following Simar et al. (2016), the optimal bandwidths h_z can be obtained through leave-one out cross validation for estimating the conditional distribution $H_{X_1,X_2,Y_1,Y_2|Z}$.[6] Then we have

$$\begin{aligned}&\widehat{H}_{n,X_1Y_1|X_2Y_2Z}(x_1, y_1 \mid x_2, y_2, z)\\ &= \frac{\sum_{i=1}^n \mathbb{1}(X_{1,i} \leq x_1, X_{2,i} \leq x_2, Y_{1,i} \geq y_1, Y_{2,i} \geq y_2) K\big((Z_i - z) \oslash h_z\big)}{\sum_{i=1}^n \mathbb{1}(X_{2,i} \leq x_2, Y_{2,i} \geq y_2) K\big((Z_i - z) \oslash h_z\big)},\end{aligned} \tag{3.11}$$

where, with some abuse of notations when Z is multivariate, $K(\cdot)$ is the chosen kernel function [for multivariate Z we use a product kernel, $h_z = (h_z^1, \ldots, h_z^r)$ and the division by h_z is component-wise]. The version for the transformed variables $(X^\star, Y^\star)$ is the same after adapting the notation. This estimator will be useful for the robust frontiers below. For the conditional full frontier, only the knowledge of the bandwidth h_z is needed. We know from Daraio and Simar (2007) and Jeong et al. (2010) that the conditional FDH estimator is a localized version of the FDH estimator, where "localizing" means using only observations (X_i, Y_i) such that $\|Z_i - z\| \leq h_z$ (component-wise). So the expression in (3.8) transforms as follows:

$$\begin{aligned}\widehat{\beta}(x, y \mid d_x, d_y, z) &= \sup\{\beta > 0 \mid \widehat{H}_{n,X_1^\star Y_1^\star|X_2^\star Y_2^\star Z}(x_1^\star - \beta i_{p_1}, y_1^\star + \beta i_{q_1} \mid x_2^\star, y_2^\star, z) > 0\},\\ &= \max_{\{i \mid X_{2,i} \leq x_2, Y_{2,i} \geq y_2, |Z_i - z| \leq h_z\}} \left[\min_{\substack{k=1\ldots,p1\\ \ell=1,\ldots,q_1}} \left\{ x_1^{\star,k} - X_{1,i}^{\star,k}, Y_{1,i}^{\star,\ell} - y_1^{\star,\ell} \right\} \right],\end{aligned} \tag{3.12}$$

[6] Note that when some elements of the direction vector d are zero, we condition on the inactive variables $(X_2 \leq x_2)$ and $(Y_2 \geq y_2)$, but for bandwidths selection, the argument from the Appendix A of Simar et al. (2016) remains valid. We select the optimal bandwidth (by cross-validation) for estimating $H_{XY|Z}(x, y \mid Z = z)$ rather the ones for estimating the conditional distributions $H_{X_1Y_1|X_2Y_2Z}(x_1, y_1 \mid X_2 \leq x_2, Y_2 \geq y_2, Z = z)$. It is easy to see that for fixed (x_1, y_1, z), the resulting $\widehat{H}_{n,X_1Y_1|X_2Y_2Z}(x_1, y_1 \mid X_2 \leq x_2, Y_2 \geq y_2, Z = z)$ defined in (3.10) cannot decrease with x_2 and cannot increase with y_2, and so for $\widehat{\beta}(x, y \mid d_x, d_y, z)$ which is required by the same economic reasoning. Everything else constant (i.e., producing the same level of outputs y_1 with the same level of inputs x_1 and under conditions z), the efficiency should increase when decreasing some of the inputs in x_2 or by increasing some of the outputs in y_2. This property is not guaranteed by using bandwidths $h_z(x_2, y_2)$ changing with the levels of the inactive variables (x_2, y_2).

where we see clearly after comparing with (3.8) that under the max operator, that conditional directional distance are localized versions of the unconditional FDH estimator.

The Matlab function `ZFDH_dirdist_new(hz,x,y,z,dx,dy,X,Y,Z)` in "Appendix" computes $\widehat{\beta}(x, y \mid d_x, d_y, z)$ as described in (3.12), where h_z has been determined in advance.

4 Robust version: order-α quantile frontiers

4.1 Unconditional case

Daouia and Simar (2007) introduced order-α quantile frontiers for radial measures in the multivariate case. These have been adapted to directional distances in Simar and Vanhems (2012). Their definition in the more general case where some elements of d_x and/or of d_y may be equal to zero can be presented as follows. For any $\alpha \in (0, 1]$, and for any (x_2, y_2) such that $H_{X_2Y_2}(x_2, y_2) > 0$,

$$\begin{aligned}\beta_\alpha(x, y \mid d_x, d_y) &= \sup\{\beta \mid H_{X_1Y_1|X_2Y_2}(x_1 - \beta d_{x_1}, y_1 + \beta d_{y_1} \mid x_2, y_2) > 1 - \alpha\},\\ &= \sup\{\beta \mid H_{X_1X_2Y_1Y_2}(x_1 - \beta d_{x_1}, x_2, y_1 + \beta d_{y_1}, y_2) > (1 - \alpha) H_{X_2Y_2}(x_2, y_2)\},\end{aligned} \tag{4.1}$$

noting that using the quantile of the complete joint distribution $H_{X_1X_2Y_1Y_2}$ [unconditional to (X_2, Y_2)] would give different objects unless both X_2 and Y_2 are empty and so $H_{X_2Y_2}(x_2, y_2) = 1$.[7] Note also that a negative value of $\beta_\alpha(x, y \mid d_x, d_y)$ indicates a unit (x, y) lying above the order-α frontier. In the transformed coordinate system this gives

$$\beta_\alpha(x, y \mid d_x, d_y) = \sup\{\beta \mid H_{X_1^\star Y_1^\star|X_2^\star Y_2^\star}(x_1^\star - \beta i_{p_1}, y_1^\star + \beta i_{q_1} \mid x_2^\star, y_2^\star) > 1 - \alpha\}. \tag{4.2}$$

Now Consider the random variable

$$W^{x,y}\left(X_1^\star, Y_1^\star\right) = \min_{\substack{k = 1 \ldots, p1\\ \ell = 1, \ldots, q_1}} \left\{x_1^{\star,k} - X_1^{\star,k}, Y_1^{\star,\ell} - y_1^{\star,\ell}\right\} \tag{4.3}$$

Clearly, the conditional survival function of $W^{x,y}\left(X_1^\star, Y_1^\star\right)$ is given by

$$\begin{aligned}S_W(w \mid x_2, y_2) &= \text{Prob}\left(W^{x,y}\left(X_1^\star, Y_1^\star\right) \geq w \mid X_2 \leq x_2, Y_2 \geq y_2\right)\\ &= H_{X_1^\star Y_1^\star|X_2^\star Y_2^\star}(x_1^\star - w i_{p_1}, y_1^\star + w i_{q_1} \mid x_2^\star, y_2^\star),\end{aligned} \tag{4.4}$$

which allows definition of $\beta_\alpha(x, y \mid d_x, d_y)$ through the quantiles of $W^{x,y}\left(X_1^\star, Y_1^\star\right)$. The nonparametric estimator is obtained by plugging the empirical version of $H_{X_1^\star Y_1^\star|X_2^\star Y_2^\star}$ into the last equation. Consider the sequence $W^{x,y}\left(X_{1,i}^\star, Y_{1,i}^\star\right)$ for $i = 1, \ldots, n$. Define $N_2^{x,y} = n\widehat{H}_{n,X_2^\star Y_2^\star}(x_2^\star, y_2^\star)$, the number of observations in the original sample with $X_{2,i} \leq x_2$ and $Y_{2,i} \geq y_2$ (note that $N_2^{x,y} = n$ if both X_2 and Y_2 are empty, with all the directions in d being > 0).

Next, define the order statistics

$$W_{(1)}^{x,y} \leq W_{(2)}^{x,y} \leq \cdots W_{(N_2^{x,y})}^{x,y} \tag{4.5}$$

[7] This would be more in the vein of the quantile frontier introduced by Daouia et al. (2017) and adapted to directional distances in Daraio and Simar (2005). This approach has the drawback of being defined only for quantiles $(1-\gamma)$ where $\gamma > 1 - H_{X_2Y_2}(x_2, y_2)$ [$\gamma = 1 - (1-\alpha)H_{X_2Y_2}(x_2, y_2)$, with $\alpha > 0$], so if the point (x_2, y_2) is on the edge of their possible values, only quantiles with very large values of γ would be available.

of the variables $W^{x,y}\left(X^{\star}_{1,i}, Y^{\star}_{1,i}\right)$ only for the $N_2^{x,y}$ observations with $X_{2,i} \leq x_2$ and $Y_{2,i} \geq y_2$: It immediately follows that

$$\widehat{\beta}_{\alpha}(x, y \mid d_x, d_y) = \begin{cases} W^{x,y}_{(\alpha N_2^{x,y})} & \text{if } \alpha N_2^{x,y} \in \mathbb{N} \\ W^{x,y}_{([\alpha N_2^{x,y}]+1)} & \text{otherwise,} \end{cases} \tag{4.6}$$

where $[a]$ denotes the integer part of a. Note that if $\alpha = 1$ we recover the full frontier estimate $\widehat{\beta}(x, y \mid d_x, d_y)$.

The Matlab function `OrderAlpha_dirdist_new(x,y,dx,dy,X,Y,alpha)` in "Appendix" computes $\widehat{\beta}_{\alpha}(x, y \mid d_x, d_y)$ using the expression in (4.6), where $\alpha \in (0, 1]$ is selected a priori.

4.2 Conditioning on environmental factors *Z*

The conditional (on $Z = z$) version of the order-α directional distance estimator is rather easy to derive using the conditional distribution and its nonparametric estimator described above in (3.10) and (3.11). The definition of the order-α conditional directional distance, for any $\alpha \in (0, 1]$, is given by

$$\beta_{\alpha}(x, y \mid d_x, d_y, z) = \sup\{\beta \mid H_{X_1Y_1|X_2Y_2Z}(x_1 - \beta d_{x_1}, y_1 + \beta d_{y_1} \mid x_2, y_2, z) > 1 - \alpha\}, \tag{4.7}$$

$$= \sup\{\beta \mid H_{X_1^{\star}Y_1^{\star}|X_2^{\star}Y_2^{\star}Z}(x_1^{\star} - \beta i_{p_1}, y_1^{\star} + \beta i_{q_1} \mid x_2^{\star}, y_2^{\star}, z) > 1 - \alpha\}. \tag{4.8}$$

The conditional (on Z) survival function of $W^{x,y}\left(X^{\star}_1, Y^{\star}_1\right)$ is given by

$$\begin{aligned} S_{W|Z}(w \mid x_2, y_2, z) &= \text{Prob}\left(W^{x,y}\left(X^{\star}_1, Y^{\star}_1\right) \geq w \mid X_2 \leq x_2, Y_2 \geq y_2, Z = z\right) \\ &= H_{X_1^{\star}Y_1^{\star}|X_2^{\star}Y_2^{\star}Z}(x_1^{\star} - wi_{p_1}, y_1^{\star} + wi_{q_1} \mid x_2^{\star}, y_2^{\star}, z). \end{aligned} \tag{4.9}$$

Its nonparametric estimator can be written as

$$\widehat{S}_{n,W|Z}(w \mid x_2, y_2, z) = \frac{\sum_{i=1}^{n} \mathbb{1}(W_i \geq w, X_{2,i} \leq x_2, Y_{2,i} \geq y_2) K\left((Z_i - z) \oslash h_z\right)}{\sum_{i=1}^{n} \mathbb{1}(X_{2,i} \leq x_2, Y_{2,i} \geq y_2) K\left((Z_i - z) \oslash h_z\right)}, \tag{4.10}$$

$$= \frac{\sum_{j=1}^{N_2^{x,y}} \mathbb{1}(W^{x,y}_{(j)} \geq w) K\left((Z^{x,y}_{[j]} - z) \oslash h_z\right)}{\sum_{i=1}^{n} \mathbb{1}(X_{2,i} \leq x_2, Y_{2,i} \geq y_2) K\left((Z_i - z) \oslash h_z\right)}, \tag{4.11}$$

where $Z^{x,y}_{[j]}$ is the observation Z_i corresponding to the jth order statistic $W^{x,y}_{(j)}$. Therefore

$$\widehat{S}_{n,W|Z}(w \mid x_2, y_2, z) = \begin{cases} 1 & \text{if } w \leq W^{x,y}_{(1)} \\ L_{k+1} & \text{if } W^{x,y}_{(k)} < w \leq W^{x,y}_{(k+1)}, k = 1, \ldots, N_2^{x,y} - 1 \\ 0 & \text{if } w > W^{x,y}_{(N_2^{x,y})}, \end{cases} \tag{4.12}$$

where for $k = 1, \ldots, N_2^{x,y} - 1$,

$$L_{k+1} = \frac{\sum_{j=k+1}^{N_2^{x,y}} K\left((Z^{x,y}_{[j]} - z) \oslash h_z\right)}{\sum_{i=1}^{n} \mathbb{1}(X_{2,i} \leq x_2, Y_{2,i} \geq y_2) K\left((Z_i - z) \oslash h_z\right)}. \tag{4.13}$$

Finally, the explicit expression of the conditional order-α directional distance is given by

$$\widehat{\beta}_\alpha(x, y \mid d_x, d_y, z) = \begin{cases} W^{x,y}_{(k)} & \text{if } L_{k+1} \leq 1-\alpha < L_k, k = 1, \ldots, N_2^{x,y} - 1 \\ W^{x,y}_{(N_2^{x,y})} & \text{if } 0 \leq 1-\alpha < L_{N_2^{x,y}}. \end{cases} \tag{4.14}$$

These formulae extend to the general directional distance case allowing some elements of d to equal zero, and hence include the expressions derived in Daouia and Simar (2007) for output radial distances as special cases.

The Matlab function `ZorderAlpha_dirdist_new(kernelz,hz,x,y,z,dx, dy,X,Y,Z,alpha)` in "Appendix" computes $\widehat{\beta}_\alpha(x, y \mid d_x, d_y, z)$ using (4.14), where a value $\alpha \in (0, 1]$ is passed to the function as an argument. The kernels for the components of Z can be Gaussian, Quartic, Epanechnikov or Uniform. The bandwidths h_z must be determined before this call.

5 Robust version: order-*m* partial frontiers

5.1 Unconditional case

Cazals et al. (2002) introduce order-m partial frontiers and corresponding radial efficiency measures, while Simar and Vanhems (2012) extend these to directional distances. Cazals et al. give an explicit, exact expression for computing the nonparametric estimator in the univariate case (e.g., $q = 1$ in the output orientation or $p = 1$ in the input orientation), but so far only Monte-Carlo approximations are available for more general cases (see Simar and Vanhems 2012 or Daraio and Simar 2014 for discussion). Below, we provide an exact expression for the estimators which is easy and fast to compute for any $p \geq 1$ or $q \geq 1$. The method also allows some elements of the direction vector d to be zero, and includes both the unconditional as well as the conditional-on-Z cases.

The order-m directional distance is defined as follows (see Simar and Vanhems 2012). For any integer $m \geq 1$ and for any (x_2, y_2) such that $H_{X_2Y_2}(x_2, y_2) > 0$, consider m iid variables $(X_{1,j}, Y_{1,j})$, $j = 1, \ldots, m$ drawn from the conditional distribution $H_{X_1Y_1|X_2Y_2}(x_1, y_1 \mid x_2, y_2)$ and define the random set $\widetilde{\Psi}_m^{x,y} = \bigcup_{j=1}^m \{(x_1, u, y_1, v) \in \Psi \mid x_1 \geq X_{1,j}, u \leq x_2, y_1 \leq Y_{1,j}, v \geq y_2\}$. Next, define the random measure

$$\widetilde{\beta}_m(x, y \mid d_x, d_y) = \sup\{\beta \mid (x_1 - \beta d_{x_1}, x_2, y_1 + \beta d_{y_1}, y_2) \in \widetilde{\Psi}_m^{x,y}\}. \tag{5.1}$$

Then the order-m directional distance is given by

$$\beta_m(x, y \mid d_x, d_y) = \mathbb{E}(\widetilde{\beta}_m(x, y \mid d_x, d_y) \mid X_2 \leq x_2, Y_2 \geq y_2). \tag{5.2}$$

Similar to the interpretation in Cazals et al. (2002), $\beta_m(x, y \mid d_x, d_y)$ in (5.2) benchmarks the unit (x, y) against the expectation of the "best" among m peers using less inactive inputs X_2 and producing more inactive outputs Y_2. A negative value of $\beta_m(x, y \mid d_x, d_y)$ indicates a unit at (x, y) operating above the order-m frontier.

We can now see what happens in the transformed coordinate system $(X^\star, Y^\star)$ defined above. We have

$$\widetilde{\beta}_m(x, y \mid d_x, d_y) = \sup\{\beta \mid \widetilde{H}_{m,X_1^\star Y_1^\star|X_2^\star Y_2^\star}(x_1^\star - \beta i_{p_1}, y_1^\star + \beta i_{q_1} \mid x_2^\star, y_2^\star) > 0\}, \tag{5.3}$$

where $\widetilde{H}_{m,X_1^\star Y_1^\star|X_2^\star Y_2^\star}$ is the empirical version (in the new coordinate system) of $H_{X_1^\star Y_1^\star|X_2^\star Y_2^\star}$ obtained from the random sample $\{(X_{1,j}, Y_{1,j})\}_{j=1}^m$. Hence by defining $W^{x,y}(X_1^\star, Y_1^\star)$, as above in (4.3), but now for the m transformed observations $\{(X_{1,j}^\star, Y_{1,j}^\star)\}_{j=1}^m$, it is clear that

$$\widetilde{\beta}_m(x, y \mid d_x, d_y) = \max_{j=1,\ldots,m} \left\{ W^{x,y}(X^\star_{1,j}, Y^\star_{1,j}) \right\} \tag{5.4}$$

where the $(X^\star_{1,j}, Y^\star_{1,j})$ are distributed according $H_{X^\star_1 Y^\star_1 | X^\star_2 Y^\star_2}$.

The survival function of $W^{x,y}(X^\star_1, Y^\star_1)$ is given by

$$\begin{aligned} S_W(w \mid x_2, y_2) &= \Pr(W^{x,y}(X^\star_1, Y^\star_1) \geq w \mid X_2 \leq x_2, Y_2 \geq y_2) \\ &= H_{X^\star_1 Y^\star_1 | X^\star_2 Y^\star_2}(x^\star_1 - w i_{p_1}, y^\star_1 + w i_{q_1} \mid x^\star_2, y^\star_2). \end{aligned} \tag{5.5}$$

Consequently, the distribution function of $\widetilde{\beta}_m(x, y \mid d_x, d_y)$ is given by

$$G^{x,y}_m(w) = \Pr(\widetilde{\beta}_m(x, y \mid d_x, d_y) \leq w) = [1 - S_W(w \mid x_2, y_2)]^m . \tag{5.6}$$

This leads to an equivalent expression for $\beta_m(x, y \mid d_x, d_y)$, namely

$$\beta_m(x, y \mid d_x, d_y) = \int_0^{\beta(x,y|d_x,d_y)} w \, dG^{x,y}_m(w). \tag{5.7}$$

It is straightforward to confirm that $G^{x,y}_m(0) = [1 - S_W(0 \mid x_2, y_2)]^m = 0$ and that $G^{x,y}_m(w) = 1$ for all $w \geq \beta(x, y \mid d_x, d_y)$. Integration by parts then gives the desired result,

$$\beta_m(x, y \mid d_x, d_y) = \beta(x, y \mid d_x, d_y) - \int_0^{\beta(x,y|d_x,d_y)} G^{x,y}_m(w) \, dw. \tag{5.8}$$

Nonparametric estimation is now easy. Plugging the empirical version $\widehat{H}_{n, X^\star_1 Y^\star_1 | X^\star_2 Y^\star_2}$ into (5.5) leads to

$$\widehat{\beta}_m(x, y \mid d_x, d_y) = \widehat{\beta}(x, y \mid d_x, d_y) - \int_0^{\widehat{\beta}(x,y|d_x,d_y)} \widehat{G}^{x,y}_m(w) \, dw, \tag{5.9}$$

where $\widehat{G}^{x,y}_m(w) = \left[1 - \widehat{S}_{n,W}(w \mid x_2, y_2)\right]^m$. As shown above, $\widehat{S}_{n,W}(w \mid x_2, y_2)$ is easily obtained from the order statistics of the $N^{x,y}_2$ variables $W^{x,y}\left(X^\star_{1,i}, Y^\star_{1,i}\right)$ defined in (4.3) such that $X_2 \leq x_2, Y_2 \geq y_2$, i.e., $W^{x,y}_{(1)} \leq \cdots \leq W^{x,y}_{(N^{x,y}_2)}$. Then an explicit expression for the order-m directional distance estimator is given by

$$\widehat{\beta}_m(x, y \mid d_x, d_y) = \sum_{j=1}^{N^{x,y}_2} W^{x,y}_{(j)} \left[\left(\frac{j}{N^{x,y}_2} \right)^m - \left(\frac{j-1}{N^{x,y}_2} \right)^m \right]. \tag{5.10}$$

The Matlab function `Orderm_dirdist_new(x,y,dx,dy,X,Y,m)` in "Appendix" organizes the computations for evaluating $\widehat{\beta}_m(x, y \mid d_x, d_y)$ as described in (5.10), where m is an integer ≥ 1 chosen a priori.

5.2 Conditioning on environmental factors *Z*

The conditional-on-Z case follows similar arguments. The definition of the conditional order-m directional distance is now based on the conditional (on Z) versions of the various distributions used above in Sect. 5.1.

For any integer $m \geq 1$ and for any (x_2, y_2) such that $H_{X_2 Y_2 | Z}(x_2, y_2 \mid z) > 0$, consider m iid variables $(X_{1,j}, Y_{1,j})$, $j = 1, \ldots, m$ drawn from the conditional distribution

$H_{X_1Y_1|X_2Y_2Z}(x_1, y_1 \mid x_2, y_2, z)$ and define the random set $\widetilde{\Psi}_m^{x,y,z} = \bigcup_{j=1}^m \{(x_1, u, y_1, v) \in \Psi \mid x_1 \geq X_{1,j}, u \leq x_2, y_1 \leq Y_{1,j}, v \geq y_2, Z = z\}$. Define the random measure

$$\widetilde{\beta}_m(x, y \mid d_x, d_y, z) = \sup\{\beta \mid (x_1 - \beta d_{x_1}, x_2, y_1 + \beta d_{y_1}, y_2) \in \widetilde{\Psi}_m^{x,y,z}\}. \tag{5.11}$$

Then the conditional order-m directional distance is defined by

$$\beta_m(x, y \mid d_x, d_y, z) = \mathbb{E}(\widetilde{\beta}_m(x, y \mid d_x, d_y) \mid X_2 \leq x_2, Y_2 \geq y_2, Z = z). \tag{5.12}$$

Following the arguments used in Sect. 5.1 for the unconditional case, we have

$$\begin{aligned}\widetilde{\beta}_m(x, y \mid d_x, d_y, z) &= \sup\{\beta \mid \widetilde{H}_{m,X_1^\star Y_1^\star|X_2^\star Y_2^\star Z}(x_1^\star - \beta i_{p_1}, y_1^\star + \beta i_{q_1} \mid x_2^\star, y_2^\star, z) > 0\}, \\ &= \max_{j=1,\ldots,m} \left\{ W^{x,y}(X_{1,j}^\star, Y_{1,j}^\star) \right\}, \end{aligned} \tag{5.13}$$

where now the $(X_{1,j}^\star, Y_{1,j}^\star)$ are distributed according to $H_{X_1^\star Y_1^\star|X_2^\star Y_2^\star Z}$. The conditional (on Z) survival function of $W^{x,y}(X_1^\star, Y_1^\star)$ is given by

$$\begin{aligned} S_{W|Z}(w \mid x_2, y_2, z) &= \Pr(W^{x,y}(X_1^\star, Y_1^\star) \geq w \mid X_2 \leq x_2, Y_2 \geq y_2, Z = z) \\ &= H_{X_1^\star Y_1^\star|X_2^\star Y_2^\star Z}(x_1^\star - w i_{p_1}, y_1^\star + w i_{q_1} \mid x_2^\star, y_2^\star, z). \end{aligned} \tag{5.14}$$

Finally, using reasoning analogous to that in Sect. 5.1, we have

$$\beta_m(x, y \mid d_x, d_y, z) = \beta(x, y \mid d_x, d_y, z) - \int_0^{\beta(x,y|d_x,d_y,z)} G_m^{x,y,z}(w)\, dw, \tag{5.15}$$

where $G_m^{x,y,z}(w) = \left[1 - S_{W|Z}(w \mid x_2, y_2, z)\right]^m$.

The nonparametric estimator is obtained by plugging the estimator of the survival function $S_{W|Z}$ described in (4.12) into (5.15), yielding

$$\widehat{\beta}_m(x, y \mid d_x, d_y, z) = \widehat{\beta}(x, y \mid d_x, d_y, z) - \int_0^{\widehat{\beta}(x,y|d_x,d_y,z)} \widehat{G}_m^{x,y,z}(w)\, dw, \tag{5.16}$$

where $\widehat{G}_m^{x,y,z}(w) = \left[1 - \widehat{S}_{n,W|Z}(w \mid x_2, y_2, z)\right]^m$. Simple analytical derivations reveal that the estimator can be computed by the explicit formula

$$\widehat{\beta}_m(x, y \mid d_x, d_y, z) = \sum_{k=1}^{N_2^{x,y}} W_{(k)}^{x,y} \left([1 - L_{k+1}]^m - [1 - L_k]^m \right), \tag{5.17}$$

where the L_{k+1} are defined in (4.13) for $k = 1, \ldots, N_2^{x,y} - 1$, noting that $L_1 = 1$ in (4.13), and defining $L_{N_2^{x,y}+1} = 0$.

The Matlab function `Zorderm_dirdist_new(kernelz,hz,x,y,z,dx,dy,X,YZ,m)` given in "Appendix" computes values for $\widehat{\beta}_m(x, y \mid d_x, d_y, z)$ using (5.17), where $m \geq 1$ is an integer, `kernelz` is the kernel chosen for Z and h_z is the vector of bandwidths.

6 Numerical illustrations

In this section we demonstrate how much is gained by the new computational methods introduced above, relative to the existing algorithms involving Monte Carlo methods proposed by Simar and Vanhems 2012 and Daraio and Simar (2005). For the full frontier estimates

and the order-α (both unconditional and conditional) the gain in computing speed is negligible. Nonetheless, the new methods avoid potential numerical problems related to ratios of exponentials and logarithms (recall footnote 5).

The new methods proposed for computing the *exact value* of the order-m directional distances (either unconditional or conditional) are remarkably faster than the existing Monte Carlo methods that have been used until now and that provide only (good) *approximations* of the exact values. The results presented above provide indeed exact expressions that are easy to implement in practical applications. The results presented below indicate that even if one is willing to accept precision to only two decimal places, the existing Monte Carlo-based algorithms require roughly 200 times the CPU time required by the new exact methods presented above. To obtain precision to 3 decimal places, the old methods require roughly 1000 times the CPU time required by the new methods.

It is important to note that implementation of the older Monte-Carlo algorithm is complicated when some elements of the direction vector d are zero. We used the algorithms suggested by Daraio and Simar (2005).[8] For the conditional-on-Z case, some complications arise due to step (1) of the algorithm where observations have to be sampled according to the weights given by the kernel function. These complications are avoided by the new methods developed above.

6.1 Simulated data

We first generate n frontier points in a $p+q=4$ dimensions as follows. We simulate $p=2$ inputs coordinates of frontier points $X^{\partial}_{j,i} \sim \text{Unif}(0,1)$, independently for $j=1,2$ and for $i=1,\ldots,n$. Then the coordinates of the $q=2$ efficient outputs are defined by

$$Y^{\partial}_{1,i} = X^{\partial,0.5}_{1,i} \times X^{\partial,0.5}_{2,i}$$

and

$$Y^{\partial}_{2,i} = X^{\partial,0.4}_{1,i} \times X^{\partial,0.3}_{2,i}. \tag{6.1}$$

The directional distance inefficiencies are drawn as $\beta_i \sim N^{+}(0, 0.75^2)$ for $i=1,\ldots,n$. Then we choose a common directional vector (d_x, d_y) as the population median point, say (x_0, y_0). The median input frontier point is fixed by $x^{\partial}_0 = (0.5, 0.5)'$, and the corresponding output frontier point is given by applying (6.1) to obtain $y^{\partial}_0 = (0.5, 0.6156)'$. Then the median of the distribution of the directional distances is $\beta_0 = 0.5059$. So the median point of our DGP is finally given by $x_0 = x^{\partial}_0/(1-\beta_0)$ and $y_0 = y^{\partial}_0/(1+\beta_0)$. This provides the chosen directional vector $d_x = x_0 = (1.0119, 1.0119)'$ and $d_y = y_0 = (0.3320, 0.4088)'$.

Now we can define the random sample of inputs and outputs as

$$X_i = X^{\partial}_i - \beta_i x_0,$$
$$Y_i = Y^{\partial}_i + \beta_i y_0.$$

Finally, we simulate independent $Z \sim N(1, 3^2)$. In this illustration, we are not interested interpreting the results, but rather in investigating how much computing time is saved by our new approach over the existing, previous method. Hence we choose the bandwidth by

[8] There is an unfortunate typo in Appendix B of Daraio and Simar (2005). The last line before step (1) of the algorithm appears as $D^*_{x,y} = \{(X^*_i, Y^*_i) \mid X^*_i \leq x^*, Y^*_i \geq y^*\}$ but should instead be defined as $D^*_{x,y} = \{(X^*_i, Y^*_i) \mid X^*_{2,i} \leq x^*_2, Y^*_{2,i} \geq y^*_2\}$, where only the inactive variables are concerned here.

Least Squares Cross Validation (LSCV) in estimating the conditional probability $H_{XY|Z}$ as motivated above in Sect. 2.2.

We have $p = q = 2$ and $r = 1$. For the numerical illustration we choose $n = 1000$ and $m = 100$ and $\alpha = 0.99$ (in the next example using real data we discuss how one should choose these robustness parameters in practice). With these two values the percentage of points above the partial frontiers are similar with 23% for order-α and 21% for order-m. Note that here with the one particular simulated sample providing Tables 1 and 2, the bandwidth obtained by LSCV is $h_z = 10.4502$. We know (see e.g Li et al. 2013) that in principle the "true" optimal bandwidth is $h_z = \inf$, because Z is independent of (X, Y), but in finite sample it may be smaller. Of course, as soon as the selected bandwidth is larger than the range of $(Z_1, \ldots, Z_n)$, the practical results are the same as if $h_z = \infty$.

The time required for computing the full frontier, the conditional full frontier, the order-α, the conditional order-α, the order-m and the conditional order-m for the $n = 1000$ units using the new methods presented above was only 2.11 s running on a 2.6 Ghz MacBook-Pro laptop computer running MacOS 10.13 with 8 GB of memory and using Matlab version R2015a. Using the Monte-Carlo algorithms for the order-m and conditional order-m with limited precision ($MC = 200$ replications for both cases as often recommended in the literature) requires 418.18 s of CPU time for producing the same full table of results (an increase in computation time by a factor of 198.2). Table 1 shows estimates obtained with the two approaches for the first 10 units, with subscripts "MC" identifying the order-m estimates computed via Monte Carlo methods. Comparing the MC estimates $\widehat{\beta}_{m,\mathrm{MC}}$ with the exact estimates $\widehat{\beta}_m$ gives an idea of the loss of precision due to the Monte-Carlo approximations when using the old method.

Increasing the number of Monte-Carlo trials to 1000 requires 2071.64 s of CPU time (a factor around 1000 compared to the CPU time required by our new method). The results are reported in Table 2. The Monte Carlo approximations are improved over those in Table 1, but still some error remains.

The statistical properties of these directional distance estimators have been established (see Simar and Vanhems 2012). For the unconditional full frontier case we know the order of the error of estimation is $O_p\left(n^{-1/(p+q)}\right)$, in our setup $O_p\left(n^{-1/4}\right)$, whereas it is of order $O_p\left(n^{-1/2}\right)$ for the partial efficiency measures. But it is interesting to see what this means in practice in our setup here by doing some Monte-Carlo simulations (1000 repeated random samples of size n, for various sample sizes) and look at the Monte-Carlo Bias and at the Monte-Carlo Root-Mean-Squared Errors (RMSE) of the estimators over 1000 generated samples by comparing the obtained estimators to the true values of the corresponding directional efficiency measures.

We do the exercise for the directional efficiencies computed at the median point (x_0, y_0) above defined. For achieving this, we have to know the true values of the directional efficiency. This is easy to obtain for the full frontier case, but not for the partial order frontier (see, e.g., Simar and Wilson 2013 for simplified examples where exact formula are available). But we can compute the true values by simulating a huge sample size. We did one estimation with a sample of size $n = 5 \times 10^7$ where the order of the errors is 1.4×10^{-4} for the partial order frontiers and 1.2×10^{-2} for the full frontier (in the latter case we know the true value by construction, see above). We obtain the true values appearing in the first row of Table 3, which are exact for the full frontier efficiencies and correct up to 4 decimal places for the partial efficiencies (confirmed by testing a couple of samples of this huge size). Since $h_z = \infty$, the true values of the conditional measures are identical to the unconditional ones. We select $z_0 = 1$.

Table 1 The first 10 units from the sample $n = 1000$, in the simulated example

Unit	$\widehat{\beta}(x, y)$	$\widehat{\beta}(x, y\|z)$	$\widehat{\beta}_\alpha(x, y)$	$\widehat{\beta}_\alpha(x, y\|z)$	$\widehat{\beta}_m(x, y)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y)$	$\widehat{\beta}_m(x, y\|z)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y\|z)$
1	0.3297	0.3297	0.1887	0.1910	0.2176	0.2119	0.2146	0.2143
2	0.2465	0.2465	0.1711	0.1759	0.1804	0.1744	0.1815	0.1749
3	0.1838	0.1838	0.1081	0.1086	0.1145	0.1185	0.1161	0.1154
4	0.5335	0.5335	0.3250	0.3250	0.3593	0.3680	0.3551	0.3670
5	0.0000	0.0000	− 0.1239	− 0.1237	− 0.1073	− 0.1155	− 0.1060	− 0.1139
6	0.7373	0.7373	0.6012	0.6050	0.6401	0.6365	0.6413	0.6417
7	0.8141	0.8141	0.7127	0.7050	0.7358	0.7404	0.7309	0.7350
8	0.0000	0.0000	− 0.2309	− 0.2309	− 0.1844	− 0.1933	− 0.1766	− 0.1713
9	0.2184	0.2184	0.1182	0.1184	0.1357	0.1338	0.1349	0.1284
10	0.2350	0.2350	0.0560	0.0560	0.0911	0.0825	0.0853	0.0785

We fixed $\alpha = 0.99$ and $m = 100$. Here $MC = 200$ have been used for the Monte-Carlo approximations of the order-m and conditional order-m measures

Table 2 The first 10 units from the sample $n = 1000$, in the simulated example

Unit	$\widehat{\beta}(x, y)$	$\widehat{\beta}(x, y\|z)$	$\widehat{\beta}_{\alpha}(x, y)$	$\widehat{\beta}_{\alpha}(x, y\|z)$	$\widehat{\beta}_{m}(x, y)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y)$	$\widehat{\beta}_{m}(x, y\|z)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y\|z)$
1	0.3297	0.3297	0.1887	0.1910	0.2176	0.2160	0.2146	0.2128
2	0.2465	0.2465	0.1711	0.1759	0.1804	0.1789	0.1815	0.1824
3	0.1838	0.1838	0.1081	0.1086	0.1145	0.1136	0.1161	0.1155
4	0.5335	0.5335	0.3250	0.3250	0.3593	0.3614	0.3551	0.3544
5	0.0000	0.0000	− 0.1239	− 0.1237	− 0.1073	− 0.1103	− 0.1060	− 0.1036
6	0.7373	0.7373	0.6012	0.6050	0.6401	0.6369	0.6413	0.6419
7	0.8141	0.8141	0.7127	0.7050	0.7358	0.7395	0.7309	0.7289
8	0.0000	0.0000	− 0.2309	− 0.2309	− 0.1844	− 0.1860	− 0.1766	− 0.1793
9	0.2184	0.2184	0.1182	0.1184	0.1357	0.1327	0.1349	0.1326
10	0.2350	0.2350	0.0560	0.0560	0.0911	0.0913	0.0853	0.0867

We fixed $\alpha = 0.99$ and $m = 100$. Here $MC = 1000$ have been used for the Monte-Carlo approximations of the order-m and conditional order-m measures

In the Monte-Carlo trials we select the bandwidths by LSCV over 10 pilot MC-samples and take the median of the results. In all the cases this value is much larger than the range of the sample in Z, so we select $h_z = 100$ in all the cases, giving, as expected, almost identical estimates (at least when using the exact new formulas developed in this paper).

Note also that in this exercise, to give an idea of the gain in computing time (with a gain in numerical precision for the order-m cases), in the case of $n = 10,000$, the computing time with the new exact formulas for the 1000 Monte-Carlo samples is 33 s, but 7100 s (almost 2 h) for the traditional Monte-Carlo approximations with only $MC = 200$ for the order-m cases (for $n = 20,000$, these computing times grow up respectively to 43 s and 12, 632 s = 3.5 h).

The results Table 3 confirm the theoretical results of Simar and Vanhems (2012). To summarize: (i) the full frontier estimates converge, but slowly as n increases (curse of dimensionality); (ii) the partial frontiers are much more reliable in small samples (due to the parametric $\sqrt{n}$-consistency); (iii) the conditional estimates here are identical to the unconditional ones (since Z is independent of the production process); and (iv) the Monte-Carlo approximations of the order-m, with $MC = 200$ replications, does not work so badly for small samples, but when the sample size increases the limited number of Monte-Carlo replications used to estimate the efficiencies introduces approximation error which becomes dominant (if MC is held constant) when n is large. So not only are our new formulas exact and faster, but also in these cases have better statistical behavior.

6.2 Real data on banks

In this section we illustrate our new methods using real data on US commercial banks observed in 2002. These data are also used by Simar and Wilson (2007), Bădin et al. (2012) and Florens et al. (2014) for analyzing the of environmental factors on the production process, while Daraio et al. (2018) use the same data to illustrate their test of the separability condition described by Simar and Wilson (2007).

As explained by Florens et al. (2014) and Daraio et al. (2018), the three inputs (purchased funds, core deposits and labor) can be aggregated into one input factor and the four outputs (consumer loans, business loans, real estate loans, and securities held) can be aggregated into a single output factor with minimal loss of information.[9] After the dimension-reduction, we have a sample of $n = 303$ banks with one input and one output factor and with two environmental factors Z_1, Z_2 which include a measure of the size of the banks (log of total assets) and a measure of diversity of the services offered by the banks (see Simar and Wilson (2007) for a detailed discussion of the variables). Here we again focus on the gain in computing time afforded by our new methods.

Since we want to use robust measures of efficiencies, we have to select values of the order m and the order α. As recommended in the literature (e.g., see Daraio and Simar 2014; Daouia and Gijbels 2011a, b; Simar 2003), the efficiency measures have to be computed for a number of values of m and α to determine sensible values. Simar (2003) proposes computing the order-m efficiency estimates for a grid of values of m, and reporting in a graph the corresponding percentage of points lying outside the corresponding frontiers in order to detect outliers.[10] This may involve substantial computational burden when Monte Carlos approximations are used to compute the order-m estimates.

[9] See also Wilson (2018) for detailed discussion of dimension reduction techniques for efficiency analysis.

[10] The idea is that in the absence of any outliers, this graph should show a steadily decreasing number of points outside the order-m partial frontier with increasing m (recall that as $m \to \infty$ the order-m estimates

Table 3 Statistical performances of the estimators evaluated over 1000 Monte-Carlo random samples of size $n \in \{100, 500, 1000, 5000, 10,000, 20,000\}$

	$\widehat{\beta}(x, y)$	$\widehat{\beta}(x, y\|z)$	$\widehat{\beta}_\alpha(x, y)$	$\widehat{\beta}_\alpha(x, y\|z)$	$\widehat{\beta}_m(x, y)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y)$	$\widehat{\beta}_m(x, y\|z)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y\|z)$
True Eff	0.5059	0.5059	0.2990	0.2990	0.3215	0.3215	0.3215	0.3215
$n = 100$								
BIAS	0.1876	0.1876	0.0438	0.0013	0.0344	0.0346	0.0344	0.0342
RMSE	0.1999	0.1999	0.0762	0.0748	0.0692	0.0691	0.0692	0.0693
$n = 500$								
BIAS	0.1056	0.1056	0.0096	0.0016	0.0060	0.0062	0.0060	0.0062
RMSE	0.1126	0.1126	0.0311	0.0310	0.0264	0.0267	0.0264	0.0272
$n = 1000$								
BIAS	0.0835	0.0835	0.0050	0.0012	0.0031	0.0031	0.0031	0.0034
RMSE	0.0890	0.0890	0.0224	0.0223	0.0186	0.0193	0.0186	0.0193
$n = 5000$								
BIAS	0.0492	0.0492	0.0011	0.0004	0.0008	0.0008	0.0008	0.0007
RMSE	0.0523	0.0523	0.0101	0.0101	0.0081	0.0092	0.0081	0.0094
$n = 10,000$								
BIAS	0.0381	0.0381	0.0008	0.0004	0.0004	0.0005	0.0004	0.0005
RMSE	0.0406	0.0406	0.0073	0.0073	0.0060	0.0076	0.0060	0.0076
$n = 20,000$								
BIAS	0.0301	0.0301	0.0003	0.0001	0.0002	−0.0002	0.0002	0.0005
RMSE	0.0321	0.0321	0.0049	0.0049	0.0040	0.0065	0.0040	0.0064

The measures are evaluated at the fixed median point (x_0, y_0) and when conditioning we fix $z_0 = 1$. The optimal bandwidth for each case is set at $h_z = 100$ (Z is independent of (X, Y)). For the order-m computed by Monte-Carlo approximations, we have $MC = 200$

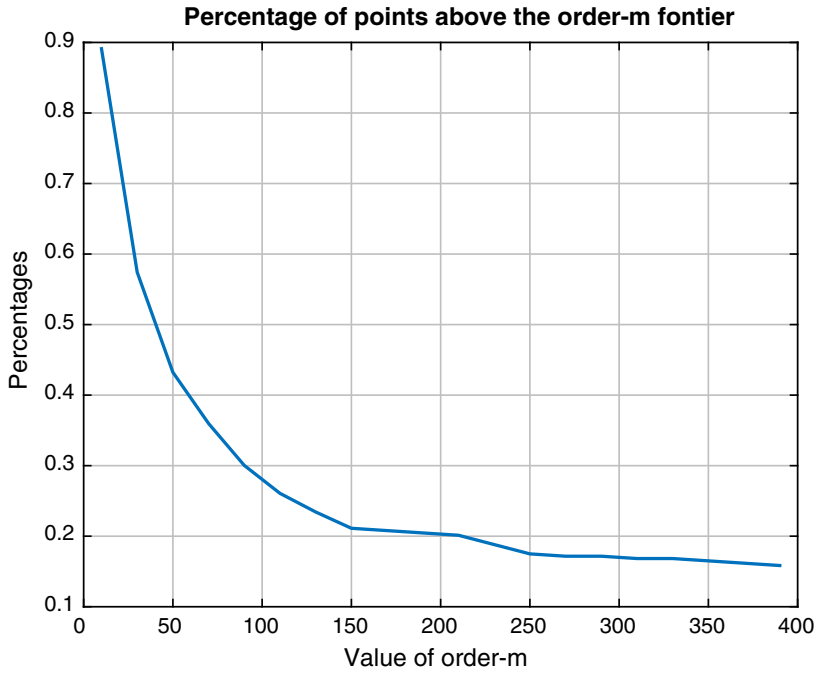

Fig. 1 Selection of m for order-m estimator

Figure 1 shows the percentage of observations lying above the order-m frontier for $m = 10, 30, 50, \ldots, 390$. Producing the estimates needed for this graph requires only 0.91 seconds using the new computational methods and the MacBook Pro described above. Using the older Monte Carlo approximations, and only $MC = 100$ Monte Carlo trials, consumes 444.0 seconds, an increase by a factor of about 488.

Figure 1 indicates that $m = 150$ may be reasonable value for the order m since the graph becomes nearly horizontal for larger values of m. To select α one could use the same approach, but for purposes of comparison, we select $\alpha = 0.995$ which results in the same proportion of observations lying outside the order-α quantile frontier as does $m = 150$ in the case of the order-m frontier. Note that this is close to the same value obtained by setting $\alpha = 0.5^{1/m} = 0.9954$ as discussed by Daouia and Gijbels (2011a, b), where the two partial frontiers coincide if the expectation operator $\mathbb{E}$ is replaced by the median operator in the definition of the order-m frontier.

Table 4 shows the various directional distance estimates for the first 20 units in the sample of $n = 303$ banks. The last four columns on the right allow comparison of the exact results for order-m and their counterparts obtained by Monte-Carlo approximation with $MC = 200$. Producing the full table for all observations required 0.44 s with the new method and 104 s with the Monte-Carlo approximations (again, a difference involving a factor of about 200, with roughly only 2 decimal digits of precision).[11]

Footnote 10 continued
converge to the corresponding FDH estimates). But if the graph exhibits a kink or "elbow" effect, then this is evidence that the remaining points outside the order-m frontier for the corresponding value of m are potential outliers.

[11] For the conditional-on-Z estimates with bivariate Z, the optimal vector of bandwidths was obtained by least-squares cross validation yielding $h_z = (0.0240,\ 0.2134)$.

Table 4 Directional Distances for the Bank Dataset

Unit	$\widehat{\beta}(x, y)$	$\widehat{\beta}(x, y\|z)$	$\widehat{\beta}_{\alpha}(x, y)$	$\widehat{\beta}_{\alpha}(x, y\|z)$	$\widehat{\beta}_{m}(x, y)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y)$	$\widehat{\beta}_{m}(x, y\|z)$	$\widehat{\beta}_{m,\mathrm{MC}}(x, y\|z)$
1	0.2612	0.2371	0.2371	0.2371	0.1705	0.1665	0.2206	0.2155
2	0.0326	0.0326	0.0102	0.0326	0.0120	0.0136	0.0325	0.0324
3	0.0975	0.0867	0.0867	0.0867	0.0703	0.0668	0.0867	0.0867
4	0.4638	0.4638	0.3885	0.4638	0.4081	0.4080	0.4629	0.4638
5	0.0000	0.0000	− 0.1749	0.0000	− 0.1608	− 0.1516	− 0.0000	0.0000
6	0.0379	0.0379	0.0373	0.0379	0.0318	0.0307	0.0379	0.0379
7	0.0254	0.0254	0.0197	0.0254	0.0154	0.0148	0.0254	0.0254
8	0.0012	0.0012	0.0000	0.0012	− 0.0060	− 0.0067	0.0007	0.0007
9	0.0000	0.0000	− 0.0028	0.0000	− 0.0037	− 0.0039	− 0.0000	0.0000
10	0.0579	0.0579	0.0547	0.0579	0.0477	0.0464	0.0579	0.0579
11	0.0871	0.0176	0.0384	0.0176	0.0461	0.0432	0.0168	0.0165
12	0.0174	0.0174	0.0000	0.0174	0.0005	− 0.0003	0.0168	0.0168
13	1.4536	1.4536	1.2835	1.4536	1.1000	1.1346	1.4536	1.4536
14	0.0650	0.0509	0.0509	0.0509	0.0420	0.0471	0.0509	0.0509
15	0.0000	0.0000	− 0.0074	0.0000	− 0.0174	− 0.0156	− 0.0000	0.0000
16	0.1250	0.1250	0.1008	0.1250	0.0789	0.0760	0.1245	0.1250
17	0.0000	0.0000	− 0.0695	0.0000	− 0.0497	− 0.0522	0.0000	0.0000
18	0.0000	0.0000	− 0.0013	0.0000	− 0.0442	− 0.0386	− 0.0000	0.0000
19	0.0151	0.0151	0.0000	0.0151	0.0006	− 0.0005	0.0151	0.0151
20	0.2475	0.2078	0.2078	0.2078	0.1954	0.2030	0.2078	0.2078

Here $n = 303$ and the Monte-Carlo approximations have been computed with $MC = 200$

7 Conclusions

This paper provides a new method for computing directional distance functions. The paper develops simple and easy-to-program expressions for computing full frontier, partial frontier (order-α and order-m cases) with their corresponding conditional-on-Z versions. The inputs, the outputs and the environmental factor can in principle be of any dimension, but we know that the curse of the dimensionality may in some cases jeopardize the quality of the estimators. In such cases, it is often possible to exploit multicollinearity in economic data using the dimension-reduction techniques analyzed by Wilson (2018). The direction vector $d \geq 0$ can have an arbitrary number of zero elements in either d_x or d_y (provided at least one element remains positive to preserve meaningfulness of the directional distance).

The new method for computing directional distances is much faster than the older Monte Carlo approximations that have been used to date as illustrated the examples in the preceding section. For the order-m cases (conditional and unconditional) to the best of our knowledge this is the first time an exact expression is provided. We have shown in the numerical illustrations, that even if low precision of the estimates is acceptable, the computing time is increased by a factor of 200–400 when the older Monte Carlo approximations are used in place of our new methods. As illustrated in the example using bank data, the new method is particularly useful when several order-m measures have to be evaluated, e.g., as in the outlier detection exercise discussed by Simar (2003) or when an appropriate value of m is chosen as discussed by Daraio and Simar (2014).

Appendix: Matlab codes

In the functions below, the reference data set when calling the functions is (X, Y) which are the matrices of inputs and outputs, with dimensions $(n \times p)$ and $(n \times q)$ respectively. When conditional measures are used, we add the matrix Z which is $(n \times d)$. All the vectors in the calling arguments must be column vectors. The point under evaluation is denoted by the vectors (xk, yk) with zk when conditioning on Z. Some elements of the direction vectors gx and gy may be zeros, but at least one element of the full vector (gx', gy') of dimension $((p+q) \times 1)$ must be strictly positive.

```
function [eff,Nx2y2] = FDH_dirdist_new(xk,yk,gx,gy,X,Y)
%  FDH Directional distance of (xk,yk) direction (gx,gy)
[n,p] = size(X); q = size(Y,2);
% Identify non-zeros in gx and gy
flagx = gx>zeros(p,1);  flagy = gy>zeros(q,1);
Xw=X; xw=xk';
Yw=Y; yw=yk';
% Changing coordinates only for elements with positive g
Xw(:,flagx') = X(:,flagx')/diag(gx(flagx)); Yw(:,flagy') = Y(:,flagy')/diag(gy(flagy));
xw(flagx') = xw(flagx')./(gx(flagx))'; yw(flagy') = yw(flagy')./(gy(flagy))';
Xw(:,~flagx') = X(:,~flagx'); Yw(:,~flagy') = Y(:,~flagy');
xw(~flagx') = xw(~flagx'); yw(~flagy') = yw(~flagy');
Ytilde=[-Xw  Yw]; yktilde=[-xw   yw];
yi = repmat(yktilde,n,1);
% Flag the ACTIVE columns of Ytilde
flagtilde=[flagx;flagy];
% Flagging the Dominating observations for the NON active variables
flagyw =Ytilde(:,~flagtilde') >=yi(:,~flagtilde');
flagw = all(flagyw,2);
ndi=sum(flagw);
if ndi == 0.00
```

```
    fprintf(' WARNING: no units with dominating inactive (x2k,y2k) \n')
    eff=-Inf;     Nx2y2=0; return
end
% Computing the DIFF only for the ACTIVE columns of Ytilde with positive g
diffyi=Ytilde(:,flagtilde') - yi(:,flagtilde');
Ydiff=min(diffyi,[],2); % n x 1 col vector with min over columns flagtilde
effi=max(Ydiff(flagw));% sort the W_i such that X2<=x2 and Y2>=y2 (flagw)
eff = effi;
Nx2y2=ndi;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Zeff,ZNxy] = ZFDH_dirdist_new(hz,xk,yk,zk,gx,gy,X,Y,Z)
%  CONDITIONAL to Z = zk, FDH  Directional distance of (xk,yk) direction (gx,gy)
%   hz is the vector of bandwidths and Z is (n x d)
n = size(X,1);
flagzk = all(abs( Z - repmat(zk',n,1)) <= repmat(hz',n,1),2);
Xzk = X(flagzk,:); Yzk = Y(flagzk,:);
[Zeff,ZNxy]= FDH_dirdist_new(xk,yk,gx,gy,Xzk,Yzk);


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [effm,Nx2y2] = Orderm_dirdist_new(xk,yk,gx,gy,X,Y,m)
%  ORDER-m FDH Directional distance of (xk,yk) direction (gx,gy)
[n,p] = size(X); q = size(Y,2);
% Identify non-zeros in gx and gy
flagx = gx>zeros(p,1); flagy = gy>zeros(q,1);
Xw=X; xw=xk';
Yw=Y; yw=yk';
% Changing coordinates only for elements with positive g
Xw(:,flagx') = X(:,flagx')/diag(gx(flagx)); Yw(:,flagy') = Y(:,flagy')/diag(gy(flagy));
xw(flagx') = xw(flagx')./(gx(flagx))'; yw(flagy') = yw(flagy')./(gy(flagy))';
Xw(:,~flagx') = X(:,~flagx'); Yw(:,~flagy') = Y(:,~flagy');
xw(~flagx') = xw(~flagx'); yw(~flagy') = yw(~flagy');
Ytilde=[-Xw  Yw]; yktilde=[-xw   yw];
yi = repmat(yktilde,n,1);
% Flag the ACTIVE columns of Ytilde
flagtilde=[flagx;flagy];
% Flaging the Dominating observations for the NON active variables
flagyw =Ytilde(:,~flagtilde') >=yi(:,~flagtilde');
flagw = all(flagyw,2);
ndi=sum(flagw);
if ndi == 0.00
    fprintf(' WARNING: no units with dominating inactive (x2k,y2k) \n')
    effm= -Inf;     Nx2y2=0; return
end
% Computing the DIFF only for the ACTIVE columns of Ytilde with positive g
diffyi=Ytilde(:,flagtilde') - yi(:,flagtilde');
Ydiff=min(diffyi,[],2);% n x 1 col vector with min over columns flagtilde
Wsort =sort(Ydiff(flagw));% sort the W_i such that X2<=x2 and Y2>=y2
indx=(1:ndi)';
p1=(indx/ndi).^m;
p2=((indx-1)/ndi).^m;
effm= sum(Wsort.*(p1-p2));
Nx2y2=ndi;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Zeffm,ZNx2y2] = Zorderm_dirdist_new(kernelz,hz,xk,yk,zk,gx,gy,X,Y,Z,m)
%  CONDITIONAL to Z = zk, ORDER-m  Directional distance of (xk,yk) direction (gx,gy)
% kernelz is the kernel function chosen for Z ('epan', 'quart', 'unif' or 'gauss')
% hz is the vector bandwidths
Kepan = @(u) (abs(u) <=1).*(1 - u.^2)*3/4; % |u| <= 1
Kgaus = @(u) exp(-u.^2/2)/sqrt(2*pi); % u\in R, cannot be used for ZFDH
Kquar = @(u) (abs(u) <=1).*(1 - u.^2).^2 *15/16;  % |u| <= 1
Kunif = @(u) 0.5*(abs(u) <=1);   % |u| <= 1
[n,p] = size(X); q = size(Y,2); d = size(Z,2);
```

```
% Kernel for Z
Dhz=diag(ones(d,1)./hz); % this is diag matrix d x d
tempz=(Z-repmat(zk',n,1)); % this is a (n x d) matrix
tempzh=tempz*Dhz;
switch lower(kernelz)
case ('gauss')
    kerzd= Kgaus(tempzh)*Dhz;
case ('quart')
    kerzd= Kquar(tempzh)*Dhz;
case ('epan')
    kerzd= Kepan(tempzh)*Dhz;
case ('unif')
    kerzd= Kunif(tempzh)*Dhz;
otherwise
    disp('Specify correct Kernel method for Z ''Epan'', ''Unif'' or ''Quart'''); return
end
kerz=prod(kerzd,2); % Product kernel: (n x 1) vector
% Identify non-zeros in gx and gy
flagx = gx>zeros(p,1); flagy = gy>zeros(q,1);
Xw=X; xw=xk';
Yw=Y; yw=yk';
% Changing coordinates only for elements with positive g
Xw(:,flagx') = X(:,flagx')/diag(gx(flagx)); Yw(:,flagy') = Y(:,flagy')/diag(gy(flagy));
xw(flagx') = xw(flagx')./(gx(flagx))'; yw(flagy') = yw(flagy')./(gy(flagy))';
Xw(:,~flagx') = X(:,~flagx'); Yw(:,~flagy') = Y(:,~flagy');
xw(~flagx') = xw(~flagx'); yw(~flagy') = yw(~flagy');
Ytilde=[-Xw  Yw];  yktilde=[-xw   yw];
yi = repmat(yktilde,n,1);
% Flag the ACTIVE columns of Ytilde
flagtilde=[flagx;flagy];
% Flaging the Dominating observations for the NON active variables
flagyw =Ytilde(:,~flagtilde') >=yi(:,~flagtilde');
flagw = all(flagyw,2);
ndi=sum(flagw);
if ndi == 0.00
    fprintf(' WARNING: no units with dominating inactive (x2k,y2k) \n')
    Zeffm= -Inf;  ZNx2y2=0; return
end
% Computing the DIFF only for the ACTIVE columns of Ytilde with positive g
diffyi=Ytilde(:,flagtilde') - yi(:,flagtilde');
% computation of L_k
Ydiff = min(diffyi,[],2);% n x 1 col vector with min over columns flagtilde
Wj = Ydiff(flagw);% keep only the W_i such that X2<=x2 and Y2>=y2 (flagw)
kerzj = kerz(flagw);
[Wsort,Is] =sort(Wj);% sort the W_j
kerzJ=kerzj(Is);
label=(ndi:-1:1);
kerzJrev=kerzJ(label);% reverse the order of the element of kerz to use cumsum below
Denom = sum(kerzj);
Lvec=cumsum(kerzJrev)/Denom;
Lvec=Lvec(label); % reorder to get the L_k
Lvec(ndi+1)=0;
indx=(2:ndi+1)';
p1=(ones(ndi,1)-Lvec(indx)).^m;
p2=(ones(ndi,1)-Lvec(indx-1)).^m;
Zeffm= sum(Wsort.*(p1-p2));
ZNx2y2=ndi;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [effa,Nx2y2] = OrderAlpha_dirdist_new(xk,yk,gx,gy,X,Y,alpha)
%  ORDER-alpha Directional distance of (xk,yk) direction (gx,gy)
[n,p] = size(X); q = size(Y,2);
% Identify non-zeros in gx and gy
flagx = gx>zeros(p,1);  flagy = gy>zeros(q,1);
Xw=X; xw=xk';
```

```
Yw=Y; yw=yk';
% Changing coordinates only for elements with positive g
Xw(:,flagx') = X(:,flagx')/diag(gx(flagx)); Yw(:,flagy') = Y(:,flagy')/diag(gy(flagy));
xw(flagx') = xw(flagx')./(gx(flagx))'; yw(flagy') = yw(flagy')./(gy(flagy))';
Xw(:,~flagx') = X(:,~flagx'); Yw(:,~flagy') = Y(:,~flagy');
xw(~flagx') = xw(~flagx'); yw(~flagy') = yw(~flagy');
Ytilde=[-Xw  Yw]; yktilde=[-xw   yw];
yi = repmat(yktilde,n,1);
% Flag the ACTIVE columns of Ytilde
flagtilde=[flagx;flagy];
% Flaging the Dominating observations for the NON active variables
flagyw =Ytilde(:,~flagtilde') >=yi(:,~flagtilde');
flagw = all(flagyw,2);
ndi=sum(flagw);
if ndi == 0.00
    fprintf(' WARNING: no units with dominating inactive (x2k,y2k) \n')
    effa=-Inf;     Nx2y2=0; return
end
% Computing the DIFF only for the ACTIVE columns of Ytilde with positive g
diffyi=Ytilde(:,flagtilde') - yi(:,flagtilde');
Ydiff=min(diffyi,[],2); % n x 1 col vector with min over columns flagtilde
Wsort=sort(Ydiff(flagw));
order=alpha*ndi;
if mod(order,floor(order))~=0
    order=floor(order)+1;
end
effai=Wsort(order);% sort the W_i such that X2<=x2 and Y2>=y2 (flagw)
effa = effai;
Nx2y2=ndi;


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Zeffa,ZNx2y2] = ZorderAlpha_dirdist_new(kernelz,hz,xk,yk,zk,gx,gy,X,Y,Z,alpha)
%  CONDITIONAL to Z = zk, ORDER-alpha Directional distance of (xk,yk) direction (gx,gy)
% kernelz is the kernel function chosen for Z ('epan', 'quart', 'unif' or 'gauss')
% hz is the vector bandwidths
Kepan = @(u) (abs(u) <=1).*(1 - u.^2)*3/4; % |u| <= 1
Kgaus = @(u) exp(-u.^2/2)/sqrt(2*pi); % u\in R, cannot be used for ZFDH
Kquar = @(u) (abs(u) <=1).*(1 - u.^2).^2 *15/16;  % |u| <= 1
Kunif = @(u) 0.5*(abs(u) <=1);   % |u| <= 1
[n,p] = size(X);
q = size(Y,2);
d = size(Z,2);
% Kernel for Z
Dhz=diag(ones(d,1)./hz); % this is diag matrix d x d
tempz=(Z-repmat(zk',n,1)); % this is a (n x d) matrix
tempzh=tempz*Dhz;
switch lower(kernelz)
case ('gauss')
    kerzd= Kgaus(tempzh)*Dhz;
case ('quart')
    kerzd= Kquar(tempzh)*Dhz;
case ('epan')
    kerzd= Kepan(tempzh)*Dhz;
case ('unif')
    kerzd= Kunif(tempzh)*Dhz;
otherwise
    disp('Specify corect Kernel method for Z ''Epan'', ''Unif'' or ''Quart'''); return
end
kerz=prod(kerzd,2); % Product kernel: (n x 1) vector
% Identify non-zeros in gx and gy
flagx = gx>zeros(p,1); flagy = gy>zeros(q,1);
Xw=X; xw=xk';
Yw=Y; yw=yk';
% Changing coordinates only for elements with positive g
```

```
Xw(:,flagx') = X(:,flagx')/diag(gx(flagx)); Yw(:,flagy') = Y(:,flagy')/diag(gy(flagy));
xw(flagx') = xw(flagx')./(gx(flagx))'; yw(flagy') = yw(flagy')./(gy(flagy))';
Xw(:,~flagx') = X(:,~flagx'); Yw(:,~flagy') = Y(:,~flagy');
xw(~flagx') = xw(~flagx'); yw(~flagy') = yw(~flagy');
Ytilde=[-Xw  Yw]; yktilde=[-xw   yw];
yi = repmat(yktilde,n,1);
% Flag the ACTIVE columns of Ytilde
flagtilde=[flagx;flagy];
% Flaging the Dominating observations for the NON active variables
flagyw =Ytilde(:,~flagtilde') >=yi(:,~flagtilde');
flagw = all(flagyw,2);
ndi=sum(flagw);
if ndi == 0.00
    fprintf(' WARNING: no units with dominating inactive (x2k,y2k) \n')
    Zeffa= -Inf;     ZNx2y2=0; return
end
% Computing the DIFF only for the ACTIVE columns of Ytilde with positive g
diffyi=Ytilde(:,flagtilde') - yi(:,flagtilde');
% computation of L_k
Ydiff = min(diffyi,[],2);% n x 1 col vector with min over columns flagtilde
Wj = Ydiff(flagw);% keep only the W_i such that X2<=x2 and Y2>=y2 (flagw)
kerzj = kerz(flagw);
[Wsort,Is] =sort(Wj);% sort the W_j
kerzJ=kerzj(Is);
label=(ndi:-1:1);
kerzJrev=kerzJ(label);% reverse the order of the element of kerz to use cumsum below
Denom = sum(kerzj);
Lvec=cumsum(kerzJrev)/Denom;
Lvec=Lvec(label); % reorder to get the L_k
% computation of the order k
beta=ones(ndi,1)*(1-alpha);
order=sum(beta < Lvec);
Zeffa= Wsort(order);
ZNx2y2=ndi;
```

References

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

Bădin, L., Daraio, C., & Simar, L. (2010). Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. *European Journal of Operational Research*, *201*(2), 633–640.

Bădin, L., Daraio, C., & Simar, L. (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research*, *223*(3), 818–833.

Bădin, L., Daraio, C., & Simar, L. (2014). Explaining inefficiency in nonparametric production models: The state of the art. *Annals of Operations Research*, *214*(1), 5–30.

Bădin, L., Daraio, C.,& Simar, L. (2018). A bootstrap approach for bandwidth selection in estimating conditional efficiency, TR n. 02 2018, DIAG, Sapienza university of Rome.

Baležentis, T., & De Witte, K. (2015). One- and multi-directional conditional efficiency measurement: Efficiency in Lithuanian family farms. *European Journal of Operational Research*, *245*(2), 612–622.

Broekel, T. (2012). Collaboration intensity and regional innovation efficiency in Germany: A conditional efficiency approach. *Industry and Innovation*, *19*(2), 155–179.

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

Chambers, R. G., Chung, Y. H., & Färe, R. (1996). Benefit and distance functions. *Journal of Economic Theory*, *70*, 407–419.

Chambers, R. G., Chung, Y. H., & Färe, R. (1998). Profit, directional distance functions and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, *98*, 351–364.

Charnes, A. W., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *Journal of Operational Research*, *2*, 429–444.

Cordero, J. M., Pedraja-Chaparro, F., Pisaflores, E. C., & Polo, C. (2017). Efficiency assessment of Portuguese municipalities using a conditional nonparametric approach. *Journal of Productivity Analysis*, *48*(1), 1–24.

Cordero, J. M., Santin, D., & Simancas, R. (2017). Assessing European primary school performance through a conditional nonparametric model. *Journal of the Operational Research Society*, *68*(4), 364–376.
Daouia, A., Florens, J. P., & Simar, L. (2010). Frontier estimation and extreme values theory. *Bernoulli*, *16*(4), 1039–1063.
Daouia, A., Florens, J. P., & Simar, L. (2012). Regularization of non-parametric frontier estimators. *Journal of Econometrics*, *168*, 285–299.
Daouia, A., & Gijbels, I. (2011). Robustness and inference in nonparametric partial frontier modeling. *Journal of Econometrics*, *161*, 147–165.
Daouia, A., & Gijbels, I. (2011). Estimating frontier cost models using extremiles. In I. Van Keilegom & P. W. Wilson (Eds.), *Exploring research frontiers in contemorary statistics and econometrics* (pp. 65–81). Berlin: Sprinter.
Daouia, A., & Ruiz-Gazen, A. (2006). Robust nonparametric frontier estimators: Qualitative robustness and influence function. *Statistica Sinica*, *16*, 1233–1253.
Daouia, A., & Simar, L. (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics*, *140*, 375–400.
Daouia, A., Simar, L., & Wilson, P. W. (2017). Measuring firm performance using nonparametric quantile-type distances. *Econometric Review*, *36*(1–3), 156–181.
Daraio, C., & Simar, L. (2005). Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis*, *24*(1), 93–121.
Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric nethods in efficiency analysis: Methodology and applications*. New-York: Springer.
Daraio, C., & Simar, L. (2014). Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research*, *237*, 358–369.
Daraio, C., Simar, L., & Wilson, P. W. (2018). Central limit theorems for conditional efficiency measures and tests of the "separability" condition in nonparametric, two-stage models of production. in press, *The Econometrics Journal*.
Debreu, G. (1951). The coefficient of resource utilization. *Econometrica*, *19*(3), 273–292.
Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor inefficiency in post offices. In M. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurements* (pp. 243–267). Amsterdam: North-Holland.
De Witte, K., & Geys, B. (2011). Evaluating efficient public good provision: Theory and evidence from a generalised conditional efficiency model for public libraries. *Journal of urban economics*, *69*(3), 319–327.
Färe, R., & Grosskopf, S. (2004). *Efficiency and productivity: New directions*. Boston, MA: Kluwer Academic Publishers.
Färe, R., Grosskopf, S., & Margaritis, D. (2008). Efficiency and productivity: Malmquist and more. In H. Fried, C. A. Knox Lovell, & S. Schmidt (Eds.), *The measurement of productive efficiency* (Vol. 2). Oxford: Oxford University Press.
Farrell, M. J. (1957). The measurement of the productive efficiency. *Journal of the Royal Statistical Society, Series A, CXX, Part*, *3*, 253–290.
Ferreira, D. C., Marques, R. C., & Nunes, A. M. (2018). Economies of scope in the health sector: The case of Portuguese hospitals. *European Journal of Operational Research*, *266*(2), 716–735.
Florens, J. P., Simar, L., & Van Keilegom, I. (2014). Frontier estimation in nonparametric location-scale models. *Journal of Econometrics*, *178*, 456–470.
Fuentes, R., Torregrosa, T., & Ballenilla, E. (2015). Conditional order-*m* efficiency of wastewater treatment plants: The role of environmental factors. *Water*, *7*(10), 5503–5524.
Guerrini, A., Romano, G., Mancuso, F., & Carosi, L. (2016). Identifying the performance drivers of wastewater treatment plants through conditional order-m efficiency analysis. *Utilities Policy*, *42*, 20–31.
Haelermans, C., & De Witte, K. (2012). The role of innovations in secondary school performance: Evidence from a conditional efficiency model. *European Journal of Operational Research*, *223*(2), 541–549.
Halkos, G. E., & Managi, S. (2016). Measuring the effect of economic growth on countries an environmental efficiency: A conditional directional distance function approach. *Environmental and Resource Economics,* 1–23.
Halkos, G. E., Stern, D. I., & Tzeremes, N. G. (2016). Population, economic growth and regional environmental inefficiency: Evidence from US states. *Journal of cleaner production*, *112*, 4288–4295.
Halkos, G. E., Sundström, A., & Tzeremes, N. G. (2015). Regional environmental performance and governance quality: A nonparametric analysis. *Environmental Economics and Policy Studies*, *17*(4), 621–644.
Halkos, G. E., & Tzeremes, N. G. (2013a). A conditional directional distance function approach for measuring regional environmental efficiency: Evidence from UK regions. *European Journal of Operational Research*, *227*(1), 182–189.

Halkos, G. E., & Tzeremes, N. G. (2013b). National culture and eco-efficiency: An application of conditional partial nonparametric frontiers. *Environmental Economics and Policy Studies*, *15*(4), 423–441.
Halkos, G. E., & Tzeremes, N. G. (2014). Public sector transparency and countries: Environmental performance: A nonparametric analysis. *Resource and Energy Economics*, *38*, 19–37.
Jeong, S. O., Park, B. U., & Simar, L. (2010). Nonparametric conditional efficiency measures: Asymptotic properties. *Annals of Operations Research*, *173*, 105–122.
Kneip, A., Simar, L., & Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. *Econometric Theory*, *24*, 1663–1697.
Li, Q., Lin, J., & Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, *31*(1), 57–65.
Mallick, S., Matousek, R., & Tzeremes, N. G. (2016). Financial development and productive inefficiency: A robust conditional directional distance function approach. *Economics Letters*, *145*, 196–201.
Manello, A. (2017). Productivity growth, environmental regulation and win-win opportunities: The case of chemical industry in Italy and Germany. *European journal of operational research*, *262*(2), 733–743.
Matousek, R., & Tzeremes, N. G. (2016). CEO compensation and bank efficiency: An application of conditional nonparametric frontiers. *European Journal of Operational Research*, *251*(1), 264–273.
Minviel, J. J., & De Witte, K. (2017). The influence of public subsidies on farm technical efficiency: A robust conditional nonparametric approach. *European Journal of Operational Research*, *259*(3), 1112–1120.
Park, B. U., Jeong, S.-O., & Simar, L. (2010). Asymptotic distribution of conical-hull estimators of directional edges. *Annals of Statistics*, *38*(6), 1320–1340.
Park, B. U., Simar, L., & Weiner, C. (2000). The FDH estimator for productivity efficiency scores: Asymptotic properties. *Econometric Theory*, *16*, 855–877.
Serra, T., & Lansink, A. O. (2014). Measuring the impacts of production risk on technical efficiency: A state-contingent conditional order-m approach. *European Journal of Operational Research*, *239*(1), 237–242.
Shephard, R. W. (1970). *Theory of cost and production function*. Princeton, NJ: Princeton University Press.
Simar, L. (2003). Detecting outliers in frontiers models: A simple approach. *Journal of Productivity Analysis*, *20*, 391–424.
Simar, L., & Vanhems, A. (2012). Probabilistic characterization of directional distances and their robust versions. *Journal of Econometrics*, *166*, 342–354.
Simar, L., Vanhems, A., & Van Keilegom, I. (2016). Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *Journal of Econometrics*, *190*, 360–373.
Simar, L., Vanhems, A., & Wilson, P. W. (2012). Statistical inference with DEA estimators of directional distances. *European Journal of Operational Research*, *220*, 853–864.
Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, *136*(1), 31–64.
Simar, L., & Wilson, P. W. (2011). Inference by the *m* out of *n* bootstrap in nonparametric frontier models. *Journal of Productivity Analysis*, *36*, 33–53.
Simar, L., & Wilson, P. W. (2013). Estimation and inference in nonparametric frontier models: Recent developements and perspectives. *Foundations and Trends in Econometrics*, *5*, 183–337.
Simar, L., & Wilson, P. W. (2015). Statistical approaches for nonparametric frontier models: A guided tour. *International Statistical Review*, *83*, 77–110.
Tzeremes, N. G. (2015). Efficiency dynamics in Indian banking: A conditional directional distance approach. *European Journal of Operational Research*, *240*(3), 807–818.
Varabyova, Y., Blankart, C. R., Torbica, A., & Schreyögg, J. (2016). Comparing the efficiency of hospitals in Italy and Germany: Nonparametric conditional approach based on partial frontier. *Health Care Management Science*. https://doi.org/10.1007/s10729-016-9359-1.
Varabyova, Y., & Schreyögg, J. (2017). Integrating quality into nonparametric analysis of efficiency: A simulation comparison of popular methods. *Annals of Operations Reseacrh*. https://doi.org/10.1007/s10479-017-2628-7.
Verschelde, M., & Rogge, N. (2012). An environment-adjusted evaluation of citizen satisfaction with local police effectiveness: Evidence from a conditional data envelopment analysis approach. *European Journal of Operational Research*, *223*(1), 214–225.
Wilson, P. W. (2018). Dimension reduction in nonparametric models of production. *European Journal of Operational Research*, *267*, 349–367.
Zschille, M. (2015). Consolidating the water industry: An analysis of the potential gains from horizontal integration in a conditional efficiency framework. *Journal of Productivity Analysis*, *44*(1), 97–114.

Affiliations

Cinzia Daraio[1] · Léopold Simar[1,2] · Paul W. Wilson[3]

✉ Paul W. Wilson
pww@clemson.edu

Cinzia Daraio
daraio@diag.uniroma1.it

Léopold Simar
leopold.simar@uclouvain.be

1 Department of Computer, Control and Management Engineering A. Ruberti (DIAG), Sapienza University of Rome, Rome, Italy

2 Institut de Statistique, Biostatistique, et Sciences Actuarielles, Université Catholique de Louvain-la-Neuve, Louvain-la-Neuve, Belgium

3 Department of Economics and School of Computing, Division of Computer Science, Clemson University, Clemson, SC, USA