INSTITUT DE
STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

0828

OPTIMAL BANDWIDTH SELECTION FOR CONDITIONAL EFFICIENCY MEASURES : A DATA-DRIVEN APPROACH

BADIN, L., DARAIO, C. and L. SIMAR

This file can be downloaded from
http://www.stat.ucl.ac.be/ISpub

Optimal Bandwidth Selection for Conditional Efficiency Measures: a Data-driven Approach

Luiza Bădin
Department of Mathematics, Academy of Economic Studies
G.Mihoc-C. Iacob Institute of Mathematical Statistics and Applied Mathematics
Bucharest, Romania
luizab@ase.ro

Cinzia Daraio
CIEG Department of Management
University of Bologna, Italy
cinzia.daraio@unibo.it

Léopold Simar*
Institute of Statistics
Université Catholique de Louvain-la-Neuve, Belgium
leopold.simar@uclouvain.be

Revised version
February 04, 2009

[*] The authors acknowledge support from the "Interuniversity Attraction Pole", Phase VI (No. P6/03) from the Belgian Science Policy.

Abstract

In productivity analysis an important issue is to detect how external (environmental) factors, exogenous to the production process and not under the control of the producer, might influence the production process and the resulting efficiency of the firms. Most of the traditional approaches proposed in the literature have serious drawbacks. An alternative approach is to describe the production process as being conditioned by a given value of the environmental variables (Cazals, Florens and Simar, 2002, Daraio and Simar, 2005). This defines conditional efficiency measures where the production set in the input × output space may depend on the value of the external variables. The statistical properties of nonparametric estimators of these conditional measures are now established (Jeong, Park and Simar, 2008). These involve the estimation of a nonstandard conditional distribution function which requires the specification of a smoothing parameter (a bandwidth). So far, only the asymptotic optimal order of this bandwidth has been established. This is of little interest for the practitioner. In this paper we fill this gap and we propose a data-driven technique for selecting this parameter in practice. The approach, based on a Least Squares Cross Validation procedure (LSCV), provides an optimal bandwidth that minimizes an appropriate (weighted) integrated Squared Error (ISE). The method is carefully described and exemplified with some simulated data with univariate and multivariate environmental factors. An application on real data (performances of Mutual Funds) illustrates how this new optimal method of bandwidth selection works in practice.

Keywords: nonparametric efficiency estimation, conditional efficiency measures, environmental factors, conditional distribution function, bandwidth.

JEL Classification: C14, C40, C60, D20

1 Introduction

The efficiency analysis literature was originally developed towards ranking the economic producers with respect to their technical efficiency scores rather than explaining the differences in the performances of the analyzed units. During the last decades, the efficiency literature has become more concerned with connecting the efficiency measures to environmental factors that cannot be controlled by the producers, but might influence the production process.

So far, two main approaches have been proposed by the efficiency literature. The traditional one-stage approach (see *e.g.* Banker and Morey, 1986, Reinhard, Lovell and Thijssen, 2000), is based on including the environmental factors (denoted below by Z) either as (non-discretionary) inputs or outputs in the model, providing an "augmented" attainable set. This approach has the shortcomings that first, it requires the *a priori* specification of the role of these exogenous factors, favorable (as a free disposal input) or unfavorable (as an undesired free disposal output) and second, that this influence is in the same direction for all values of Z. In addition, restrictive assumptions on the free disposability and on the convexity of the resulting augmented attainable set are also required. Finally, the appropriate linear programs used to estimate the resulting efficiency scores depend also on the returns to scale assumption made on this non-discretionary input or output. This makes a lot of assumptions or restrictions when often, at the start, the researcher has not a clear view on the influence of Z on the production process.

The other traditional approach is a two-stage procedure, where the efficiency scores are nonparametrically estimated in a first stage, in the input-output space. Then, in a second stage the estimated efficiency scores are regressed (mainly using parametric models) on the environmental variables (see Simar and Wilson, 2007 and the dozens of references quoted there and even more recently McDonald, 2008). However, as pointed out by Simar and Wilson (2007), usual inference on the obtained estimates of the regression coefficients is not available in this framework and if used, it is wrong. So, they propose a bootstrap-based procedure to obtain more reliable results. Recently, Park, Simar and Zelenyuk (2008), suggested to use a nonparametric model for the second stage regression. Still, these two-stage approaches require a restrictive separability condition between the input-output space and the space of external, environmental factors, assuming that these factors have no influence on the attainable set, affecting only the probability of being more or less efficient. This is often unrealistic.

A more general and appealing approach is to consider the probabilistic formulation of the production process proposed by Cazals, Florens and Simar (2002). Here the production set is the support of some probability measure in the input-output space and the traditional

Debreu–Farrell efficiency scores can be defined in terms of some nonstandard conditional distribution function (see details below). This approach allows to extend quite naturally the model in the presence of environmental factors leading to conditional Debreu–Farrell efficiency scores. Nonparametric estimators are then easily obtained by estimating, at some stage, a nonstandard conditional distribution and so providing conditional FDH estimators (Free Disposal Hull) as in Daraio and Simar (2005) or conditional DEA estimators (Data Envelopment Analysis), as in Daraio and Simar (2007b). Asymptotic properties of these estimators are now established (see Jeong, Park and Simar, 2008) and both estimators require smoothing techniques for the environmental variables (i.e., using a kernel function and a bandwidth). So far only the asymptotic order of this bandwidth has been determined (see below) but this result is of little interest for practitioners who only handle finite samples whereas it is well known that the choice of the bandwidth may be crucial for the quality of the resulting estimates.

Daraio and Simar (2005, 2007a), pointing that the bandwidth selection is an open issue in this context, suggest to use bandwidth selection methods for the kernel density estimation of the external variables Z, such as cross-validation and plug-in rules. However, these methods do not take into account the influence of the environmental variables on the production process while determining the window size, since they are based on marginal properties of Z. So there is certainly room for improving the method.

This paper is intended to fill this gap by proposing a data driven method for selecting an optimal bandwidth in practice, where optimality will be defined with respect to an integrated squared error criterion. We propose to adapt the procedure based on the theoretical results from Hall, Racine and Li (2004) on estimation of conditional probability density function (pdf) and those of Li and Racine (2007, 2008) for conditional cumulative distribution function (cdf) to our particular setup of a production process. The procedure is also useful to identify external variables components that have no influence on the production process.

The paper is organized as follows. Section 2 summarizes the theoretical results available so far on conditional efficiency measures and their nonparametric estimates. Section 3 is our main contribution: we describe the method for selecting an optimal bandwidth. Section 4 is dedicated to numerical examples. We consider simulation scenarios already used in the literature and we also exemplify our methodology on real data using the same sample of US Mutual Funds. These examples illustrate the superiority of our optimal method for selecting the bandwidth over the former methods based as in Daraio and Simar (2005) on marginal characteristics of Z. A last section concludes.

2 Conditional Efficiency Scores and Nonparametric Estimators

According to Cazals et al. (2002), the production process can be described by the joint distribution of the input-output pairs (X, Y) on $\mathbb{R}_+^p \times \mathbb{R}_+^q$, whose support is Ψ, the attainable set. For the presentation, to save space, we only focus on the output orientation and on the FDH version of the estimators, but of course, the same could be done for the input orientation and for the DEA version of the estimators (as proposed in Daraio and Simar, 2007b), since in all the cases, the problem rests on the estimation of a nonstandard conditional distribution function.

2.1 Conditional efficiency scores

An output-oriented technical efficiency for a fixed point $(x, y) \in \Psi$ can be defined in terms of the support of the q-variate survival function $S_{Y|X}(y|x) = \text{Prob}(Y \geq y|X \leq x)$. This support can indeed be interpreted as the attainable set of output values Y for a producer using the input level x. For instance, if $q = 1$, for any given x, the upper boundary of this support provides the production (frontier) function:

$$\varphi(x) = \sup\{y \mid S_{Y|X}(y|x) > 0\}. \tag{1}$$

When $q \geq 1$, we can use (as in the Debreu–Farrell approach), radial distances to evaluate the efficiency level of a point of interest. So, the (radial) output measure of efficiency of a unit operating at the level (x, y) is the maximal radial expansion of y that is attainable:

$$\lambda(x, y) = \sup\{\lambda \mid S_{Y|X}(\lambda y|x) > 0\}. \tag{2}$$

Under free disposability of the inputs and of the outputs, Cazals et al. (2002) have shown that this is equivalent to the Debreu-Farrell output efficiency score. Moreover, if we consider the probability of dominating a unit operating at level (x, y)

$$H(x, y) = \text{Prob}(X \leq x, Y \geq y), \tag{3}$$

then the following decomposition is possible:

$$H(x, y) = \text{Prob}(Y \geq y \mid X \leq x)\ \text{Prob}(X \leq x) = S_{Y|X}(y|x)F_X(x), \tag{4}$$

where $F_X(x) = \text{Prob}(X \leq x)$ is the marginal cdf of X. Consequently, for all x with $F_X(x) > 0$, the output oriented technical efficiency measure admits also the following representation:

$$\lambda(x, y) = \sup\{\lambda \mid H(x, \lambda y) > 0\}. \tag{5}$$

This probabilistic formulation of the production process allows to introduce quite easily external, environmental factors, which are exogenous to the production process itself, but may influence the process. Denote by $Z \in \mathbb{R}^r$ these factors. When an environmental variable is generated, we include the random variable Z in the model and consider the triple $(X, Y, Z) \in \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r$. For instance the following distribution function $H(x, y, z) = \text{Prob}(X \leq x, Y \geq y, Z \leq z)$ completely characterizes the process. In fact, for defining conditional measures, conditional to a level z of the external factors Z, we are mainly interested in:

$$H(x, y|z) = \text{Prob}(X \leq x, Y \geq y|Z = z) = \frac{\partial_z H(x, y, z)}{\partial_z H(\infty, 0, z)},$$

where ∂_z denotes the operator of derivative with respect to z: $\partial_z = \partial/\partial z$. Note that here also we have the following decomposition:

$$\begin{aligned} H(x, y|z) &= \text{Prob}(Y \geq y \mid X \leq x, Z = z) \, \text{Prob}(X \leq x|Z = z) \\ &= S_{Y|X,Z}(y \mid x, z) F_{X|Z}(x|z), \end{aligned}$$

where $S_{Y|X,Z}(y|x, z) = H(x, y|z)/H(x, 0|z)$. Daraio and Simar (2005), by analogy with the output Farrell efficiency score, define the conditional output efficiency measure:

$$\lambda(x, y|z) = \sup\{\lambda \mid S_{Y|X,Z}(\lambda y|x, z) > 0\} = \sup\{\lambda \mid H(x, \lambda y|z) > 0\}. \tag{6}$$

2.2 Nonparametric estimators

For sake of simplicity, we will first introduce the estimators for univariate Z, then we will summarize the main features for the multivariate case. From a random sample of i.i.d. observations $\mathcal{X} = \{(X_i, Y_i, Z_i) \mid i = 1, \ldots, n\}$, natural nonparametric estimators of the conditional survival functions introduced above are given by

$$\widehat{S}_{Y|X}(y|x) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i \geq y, X_i \leq x)}{\sum_{i=1}^n \mathbb{1}(X_i \leq x)}, \tag{7}$$

$$\widehat{S}_{Y|X,Z}(y|x, z) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i \geq y, X_i \leq x) K_h(Z_i, z)}{\sum_{i=1}^n \mathbb{1}(X_i \leq x) K_h(Z_i, z)}, \tag{8}$$

where $K_h(Z_i, z) = h^{-1} K\big((Z_i - z)/h\big)$ with $K(\cdot)$ being an univariate kernel defined on a compact support and h is the bandwidth. Note that it is the equality in the conditioning on Z that requires some smoothing techniques, whereas for inputs and outputs only inequalities are involved.

The nonparametric estimators $\widehat{\lambda}_n(x, y)$ and $\widehat{\lambda}_n(x, y|z)$ are then obtained by plugging $\widehat{S}_{Y|X}(y|x)$ and $\widehat{S}_{Y|X,Z}(y|x, z)$ in the formulas (2) and (6) above. So, we have for instance

$$\widehat{\lambda}_n(x, y) = \sup\{\lambda | \widehat{S}_{Y|X}(\lambda y|x) > 0\}. \tag{9}$$

As pointed by Cazals et al. (2002), this estimator coincides with the FDH estimator: $\widehat{\lambda}_{FDH}(x,y) = \max_{i|X_i \leq x} \left\{ \min_{j=1,\ldots,q} \frac{Y_i^j}{y^j} \right\}$, where a^j denotes the jth component of a vector a. Following the same lines, the conditional output efficiency estimator, suggested by Daraio and Simar (2005) is defined as:

$$\widehat{\lambda}_n(x,y|z) = \sup\{\lambda | \widehat{S}_{Y|X,Z}(\lambda y|x,z) > 0\} = \max_{\{i|X_i \leq x, |Z_i - z| \leq h\}} \left\{ \min_{j=1,\ldots,q} \frac{Y_i^j}{y^j} \right\}. \tag{10}$$

Looking carefully to (10) we see that the conditional efficiency scores heavily depend on the value of the bandwidth h. As noted in Daraio and Simar (2005) only kernel functions with compact support can be used.

The multivariate case

For multivariate environmental variables Z, we must select a kernel function $K(u)$ where $u \in \mathbb{R}^r$, such that $K(u) \geq 0$ and $\int_{\mathbb{R}^r} K(u)\,du = 1$. Then we have to select a bandwidth matrix H that has to be a $(r \times r)$ symmetric positive definite matrix. The scaled kernel function can then be written as $K_H(u) = |H|^{-1} K(H^{-1}u)$ where $|H|$ stands for the determinant of the matrix H. In our setup here (remember we need kernels with compact support), two possibilities will be considered: (i) As described in Daraio and Simar (2007a, Section 5.3.2), a multivariate Gaussian kernel truncated on a sphere of radius one and rescaled in order to obtain a continuous density over $\mathbb{R}^r$, with a bandwidth matrix H scaled by S, the empirical covariance matrix of the r components of Z and (ii) a product kernel of r univariate kernel functions each with its own bandwidth.

For determining optimal bandwidths, we prefer to focus on the second approach[1] which requires more computational effort (optimization in r components in place of only one) but which has, as described in Hall et al. (2004), the merit of producing asymptotically optimal smoothing for the relevant components of Z having the appropriate rate, while eliminating irrelevant components by oversmoothing (the bandwidths for these irrelevant components converge to infinity).

We define the product kernel as $K(u) = \prod_{j=1}^r K_j(u_j)$ and $H = \text{diag}(h_1, \ldots, h_r)$, where the $K_j(u_j)$ are univariate kernels with compact support and h_j are the individual bandwidths. The advantage here is that we can select an optimal bandwidth for each component of Z and so, as explained in the next section, detect irrelevant components in Z.

The conditional output efficiency estimator is computed by:

$$\widehat{\lambda}_n(x,y|z) = \sup\{\lambda | \widehat{S}_{Y|X,Z}(\lambda y|x,z) > 0\} = \max_{\{i|X_i \leq x, ||Z_i - z|| \leq h\}} \left\{ \min_{j=1,\ldots,q} \frac{Y_i^j}{y^j} \right\}, \tag{11}$$

[1] We thank two anonymous referees who stressed the main advantages of using multivariate bandwidths for Z.

where $||Z_i - z|| \leq h$ has to be understood component by component $|Z_i^j - z^j| \leq h_j$ for the product kernel with $H = \text{diag}(h_1, \dots, h_r)$.

Detecting the effect of Z on the production process

Daraio and Simar (2005) developed a useful methodology allowing to detect the effect (positive or negative) of these environmental factors on the performance of the firms. The idea is to analyze the behavior of the ratio of the conditional efficiency scores over the unconditional scores as a function of the conditioning factor Z. They show that the shape of a nonparametric regression of these ratios over the conditioning factor allows to detect positive, negative, neutral or even variable effect of the environmental factor on the production process. We will illustrate this tool in the examples below.

3 Bandwidth Selection: an optimal data driven method

The statistical properties of the nonparametric estimators are well known, see Park, Simar and Weiner (2000) for the FDH efficiency scores, Kneip, Simar and Wilson (2008) for the DEA scores and Jeong et al. (2008) for their conditional versions. To summarize, the estimators are consistent estimators (they converge to the corresponding unknown measures when the sample size $n \to \infty$), the sampling distribution of the (appropriated scaled) error of estimation converges to some nondegenerate distribution but the rate of convergence is deteriorated as the dimension in the input/output space increases (the "curse of dimensionality" of most of the nonparametric estimators).

For instance for the FDH case, with an univariate baseline bandwidth h, we have $\big(\widehat{\lambda}_n(x,y) - \lambda(x,y)\big) = O_p\big(n^{-1/(p+q)}\big)$ and $\big(\widehat{\lambda}_n(x,y|z) - \lambda(x,y|z)\big) = O_p\big((nh^r)^{-1/(p+q)}\big)$, as far as $h \to 0$ with $nh^r \to \infty$ when $n \to \infty$. We will see below that the optimal order for h is $n^{-1/(r+4)}$, so that, as pointed by Jeong et al. (2008), the rate of convergence of the FDH conditional efficiency estimators is deteriorated to $(n^{4/(r+4)})^{-1/(p+q)}$. This was expected since for conditional measures, smoothing in r dimensions is required. Similar results exist for the DEA version of the estimators. For details, see the references above.

The optimal order of the bandwidths for estimating the conditional survival function $S_{Y|X,Z}(y|x,z)$ with the kernel estimate $\widehat{S}_{Y|X,Z}(y|x,z)$ is $h_j \approx n^{-1/(r+4)}$ (see Li and Racine, 2007, p.186). Still, for practical purposes, in finite samples, this does not help to select the bandwidths. Daraio and Simar (2005, 2007a) suggest a cross-validation rule based on the estimation of the marginal density of Z, using some nearest-neighbor method. This provides a bandwidth of appropriate order, but with no particular optimality properties in finite samples. In addition, it does not take into account the way Z may influence the behavior of

Y given that $X \leq x$. So, we would prefer an adaptive data-driven method which optimizes the estimation of the conditional cdf (survival) $S(y|x,z)$, where the dependence between Y and Z, for $X \leq x$ is implicit.

We will adapt the approach developed in Hall et al. (2004) and Li and Racine (2007, 2008) who suggest for continuous Y to use Least Squares Cross Validation (LSCV) for selecting the best bandwidth when estimating the conditional pdf of Y given that $X \leq x$ and $Z = z$. We define this density as $g(y|X \leq x, z) = f(y, X \leq x, z)/m(X \leq x, z)$ where f and m are densities in y and z that can be defined as

$$f(y, X \leq x, z) = \frac{\partial^{(q+1)}}{\partial y\,\partial z} H(x,y,z) \tag{12}$$

$$m(X \leq x, z) = \frac{\partial}{\partial z} H(x,0,z). \tag{13}$$

Since we are now estimating the pdf of Y we also have to smooth the Y variables using an appropriate kernel and a bandwidth vector h_y. The estimate of the conditional density could then be written as

$$\begin{aligned}\widehat{g}(y|X \leq x, Z = z) &= \frac{\widehat{f}(y, X \leq x, z)}{\widehat{m}(X \leq x, z)} \\ &= \frac{n^{-1}\sum_{i=1}^{n} 1\!\!1(X_i \leq x) K_h(Z_i, z) L_{h_y}(Y_i, y)}{n^{-1}\sum_{i=1}^{n} 1\!\!1(X_i \leq x) K_h(Z_i, z)},\end{aligned} \tag{14}$$

where

$$L_{h_y}(Y_i, y) = \prod_{j=1}^{q} \frac{1}{h_{y_j}} L\left(\frac{Y_{ij} - y_j}{h_{y_j}}\right), \tag{15}$$

$L(\cdot)$ being a univariate kernel. We use here, to simplify the presentation, a product kernel for the Ys but any other multivariate kernel could be used. The idea is to find by LSCV the optimal values for (h_y, h), even if we will not use the resulting values of h_y when estimating the conditional cdf $S(y|x,z)$. Our final objective is indeed to find only a reasonable value for h, where h is the vector of bandwidths $h = (h_1, \dots, h_r)$.

Remark 1. *Conditionally on $X \leq x$, we are exactly in the same situation as in Li and Racine (2007, 2008) and Hall et al. (2004), but the number of observations really used in the estimation above is in fact $N_x = \sum_{i=1}^{n} 1\!\!1(X_i \leq x)$, i.e. the number of observations in the sample such that $X_i \leq x$. We know that $N_x = n\widehat{F}_X(x)$ will converge to infinity when $n \to \infty$, but in practice, and as for the FDH estimator, we will lose accuracy for small values of x, because N_x will be small. So clearly in all what follows, the results and the selected bandwidths will depend on the current value of x, but to avoid the notational complexity we will keep the notation (h_y, h) for denoting the bandwidths, even if they are determined for a specific value of x.*

The criterion is thus based on a weighted integrated squared error (ISE). We have:

$$ISE = \int \{\widehat{g}(y|X \leq x, Z = z) - g(y|X \leq x, Z = z)\}^2 m(X \leq x, z) dW(z) dy, \tag{16}$$

where the integral is over (y, z). Note that, as pointed in Hall et al. (2004), the presence of $dW(z)$ serves only to avoid difficulties caused by dividing by 0, or by numbers close to 0, in the ratio $\widehat{f}(y, X \leq x, z)/\widehat{m}(X \leq x, z)$ in (14), since we are dealing with continuous Z. By straightforward developments, it can be seen that the part of ISE that depends on the bandwidths (h_y, h) can be expressed as $I_{1n} - 2I_{2n}$ where:

$$I_{1n} = \int \widehat{g}^2(y|X \leq x, Z = z) m(X \leq x, z) dW(z) dy, \tag{17}$$

$$I_{2n} = \int \widehat{g}(y|X \leq x, Z = z) f(y, X \leq x, z) dW(z) dy. \tag{18}$$

We observe that

$$I_{1n} = \int \widehat{G}(x, z) \frac{m(X \leq x, z)}{\widehat{m}^2(X \leq x, z)} dW(z), \tag{19}$$

where $\widehat{G}(x, z) = \int \widehat{f}^2(y, X \leq x, z) dy$ can be expressed as

$$\widehat{G}(x, z) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} K_h(Z_{i_1}, z) K_h(Z_{i_2}, z) \mathbb{1}(X_{i_1} \leq x) \mathbb{1}(X_{i_2} \leq x) \int L_{h_y}(Y_{i_1}, y) L_{h_y}(Y_{i_2}, y) dy. \tag{20}$$

Finally, the cross-validation approximations $\widehat{I}_{1n}$ and $\widehat{I}_{2n}$ are obtained by

$$\widehat{I}_{1n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{G}_{(i)}(x, Z_i) \mathbb{1}(X_i \leq x) w(Z_i)}{\widehat{m}^2_{(i)}(X \leq x, Z_i)} \tag{21}$$

$$\widehat{I}_{2n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{f}_{(i)}(Y_i, X \leq x, Z_i) \mathbb{1}(X_i \leq x) w(Z_i)}{\widehat{m}_{(i)}(X \leq x, Z_i)} \tag{22}$$

where the subscript (i) indicates that the respective quantity is computed based on a sample of $(n-1)$ observations obtained from $\{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$ by deleting the ith observation (X_i, Y_i, Z_i) (leave-one-out). Here $w(Z_i)$ is a weight function that could be set equal to 1 unless Z_i is such that $\widehat{m}_{(i)}(X \leq x, Z_i)$ is close to 0, in which case, $w(Z_i) = 0$.

Note that in the computation of $\widehat{G}(x, z)$, the integral can be solved analytically, since $\int L_{h_y}(Y_{i_1}, y) L_{h_y}(Y_{i_2}, y) dy$ is the convolution of $L_{h_y}(Y_{i_1}, y)$ with itself (see Silverman, 1986, page 50).

Remark 2. *If we choose for Y a product kernel as in (15), then we have for each component of y (we drop the component index for notational convenience, so the integrals in the next equalities are univariate):*

$$\begin{aligned}\int L_{h_y}(Y_{i_1},y)L_{h_y}(Y_{i_2},y)dy &= \frac{1}{h_y^2}\int L\Big(\frac{Y_{i_1}-y}{h_y}\Big)L\Big(\frac{y-Y_{i_2}}{h_y}\Big)dy \\ &= \frac{1}{h_y}\int L(h_y^{-1}Y_{i_1}-u)L(u-h_y^{-1}Y_{i_2})du \\ &= \frac{1}{h_y}\int L(h_y^{-1}(Y_{i_1}-Y_{i_2})-v)L(v)dv \\ &= \frac{1}{h_y}L^{(2)}\Big(\frac{Y_{i_1}-Y_{i_2}}{h_y}\Big),\end{aligned}$$

where $L^{(2)}$ is the convolution of L with itself. If $L(\cdot)$ is the standard normal, $L^{(2)}(\cdot)$ will be a normal with mean 0 and variance equal to 2.

Coming back to the full multivariate integral with Gaussian product kernels we have:

$$\int L_{h_y}(Y_{i_1},y)L_{h_y}(Y_{i_2},y)dy = (2\pi)^{-q/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(Y_{i_1}-Y_{i_2})'\Sigma^{-1}(Y_{i_1}-Y_{i_2})\right\}, \qquad (23)$$

where $\Sigma = 2\text{diag}(h_{y_1}^2,\ldots,h_{y_q}^2)$. This is the density of a multivariate normal of dimension q with mean 0 and covariance matrix Σ, evaluated at the point $(Y_{i_1}-Y_{i_2})$. The same kind of result would be obtained with more general multivariate gaussian kernels.

The cross-validation criterion CV is obtained by computing[2] for the current value of x (generally this is done for the observed values of X_i) and for selected values of (h_y,h),

$$CV(h_y,h) = \widehat{I}_{1n} - 2\widehat{I}_{2n}. \qquad (24)$$

We look for minimal value of CV. As recommended in Hall et al. (2004), if there are two or more local minima, we select the second smallest of these turning points, instead of the smallest, to prevent of using too small bandwidths.

It is shown in Hall et al. (2004) and Li and Racine (2007, 2008) that this criterion (LSCV) leads to bandwidths of optimal size for the relevant components of Z. In addition, if some components of Z are irrelevant, Hall et al. (2004) show that the corresponding asymptotic value of h_j converges to infinity. In practical situations with finite samples, this will result in appropriate smoothing for the relevant components of Z and oversmoothing for the eventual irrelevant components. This will be illustrated in the next section.

At the end, as pointed in Li and Racine (2007, page 183)[3], we have also to correct the resulting h by an appropriate scaling factor, because we are considering the estimation of

[2] See the Appendix for more details on the practical implementation.

[3] We thank Qi Li and Jeff Racine for helping us in correcting a typo in the correction factor given in their book on page 183, where, in their notations, the factor $n^{(5+q)/(4+q)}$ has to be read $n^{-\frac{1}{(5+q)(4+q)}}$.

the conditional (survival) cdf and not the pdf. This scaling factor would be $n^{-\frac{q}{(4+q+r)(4+r)}}$ in our case, where q is the dimension of Y and r is the dimension of Z.

4 Numerical Illustrations

In this section we present some examples using simulated and real data, to illustrate the proposed method. For the Monte Carlo simulations, we considered one output-oriented model of production process with multiple inputs and multiple outputs already used in the literature. We introduce the dependency on the environmental factor Z and we study the cases with univariate Z and with bivariate Z. The section ends with an application on a sample of 129 US Mutual Funds analyzing a cross-section of US Aggressive-Growth (AG) Mutual Funds. These are mutual funds that may invest in emerging market growth companies, focusing on growth of capital gains and not so much on size of dividends paid.

Daraio and Simar (2005) suggest a procedure bringing light on the effect of the environmental factors on the efficiencies. They propose to analyze how the ratio of the conditional measure over the unconditional measure behaves as a function of Z . From a practical point of view, they use nonparametric smoothed regression of the ratios $\widehat{Q}^z = \frac{\widehat{\lambda}_n(x,y|z)}{\widehat{\lambda}_n(x,y)}$ over the values of z to investigate the effect of Z on the efficiencies. This regression is estimated from the observed ratios at the data point (X_i, Y_i, Z_i) , for $i = 1, \ldots, n$. Of course in this regression, we eliminate the data points with spurious one for $\widehat{Q}^z$ i.e. corresponding to original FDH-efficient points. Indeed if $\widehat{\lambda}_n(x, y) = 1$, by construction, $\widehat{\lambda}_n(x, y|z)$ cannot be different from one. These data points do not bring any information on how Z may affect the efficiency.

As explained and illustrated in Daraio and Simar (2005, 2007a), for an output oriented case, a decreasing regression line indicates an average negative effect of the variable Z on the performances (as if Z would act as an undesired output), an increasing regression would indicate an average positive effect (as if Z would act as an additional free disposal input) and an horizontal line would indicate a neutral effect. For an input oriented case, as in the Mutual Funds example below, it is just the contrary.

In all the examples we used for Z a multiplicative quartic kernel with a vector of bandwidths $h = (h_1, \ldots, h_r)$. Since we are less interested in the smoothing for Y, we used for Y product gaussian kernels with an univariate baseline bandwidth h_0 with $h_y = h_0 s_y$, s_y being the vector of empirical standard deviations of Y.

4.1 Simulated examples

We simulate our data according to a convex technology with $p = q = 2$ and with additive output, described in Park et al. (2000) and further adapted in Simar (2007) and Daraio and Simar (2007a). Here, the efficient frontier can be described by:

$$y^{(2)} = 1.0845(x^{(1)})^{0.3}(x^{(2)})^{0.4} - y^{(1)} \tag{25}$$

where $y^{(1)}$, $y^{(2)}$, and $x^{(1)}$, $x^{(2)}$ denote the components of y and x, respectively. We generate independent uniform random variables $X_i^{(j)} \sim U(1,2)$ and also independent $\tilde{Y}_i^{(j)} \sim U(0.2,5)$ for $j = 1,2$.

The output efficient random points on the frontier can be defined by:

$$Y_{i,eff}^{(1)} = \frac{1.0845(X_i^{(1)})^{0.3}(X_i^{(2)})^{0.4}}{S_i + 1} \tag{26}$$

$$Y_{i,eff}^{(2)} = 1.0845(X_i^{(1)})^{0.3}(X_i^{(2)})^{0.4} - Y_{i,eff}^{(1)}. \tag{27}$$

where $S_i = \tilde{Y}_i^{(2)}/\tilde{Y}_i^{(1)}$ represent the slopes which characterize the generated random rays in the output space for $j = 1,2$.

The efficiencies are then simulated according to $\exp(-U_i)$ and finally, the output variables are defined by $Y_i = Y_{i,eff} * \exp(-U_i)$. We generated $U_i \sim \text{Expo}(\text{mean} = 1/3)$ in the case of univariate Z and $U_i \sim \text{Expo}(\text{mean} = 1/2)$ in the multivariate case.

4.1.1 Univariate Z

In this first example we introduce an environmental factor Z generated from the Uniform distribution, $Z \sim U(1,4)$ independently of X and Y so, with no influence on the production process. We simulate a sample of $n = 100$ observations according to the following scenario:

$$Y_i^{(1)} = Y_{i,eff}^{(1)} * \exp(-U_i) \tag{28}$$

$$Y_i^{(2)} = Y_{i,eff}^{(2)} * \exp(-U_i). \tag{29}$$

Table 1 presents the FDH and conditional FDH measures of efficiency computed on this simulated data set for 10 randomly selected units. For the nonparametric estimation, we used a quartic kernel but we noted that the results remain stable when other kernels with compact support are used. As it appears by inspecting the column with the value of h in Table 1, the data-driven method we propose in this paper is able to detect the non-influence of Z by providing very large values of h (the 3 quartiles of the values of h over the 100 observations are 16.28, 17.57 and 18.61 well beyond the range of Z, which varies between 1 and 4). Thus the method allows to identify Z as irrelevant external factor by oversmoothing.

The neutral effect of Z on the production process is confirmed when looking to Figure 1 that depicts the ratios of conditional to unconditional FDH efficiency scores with the nonparametric regression of the observed scores on the values of Z. As expected the estimated regression (top panel) and the estimate of the derivative of the ratios with respect to z (bottom panel) are both, as they should be, absolutely flat.

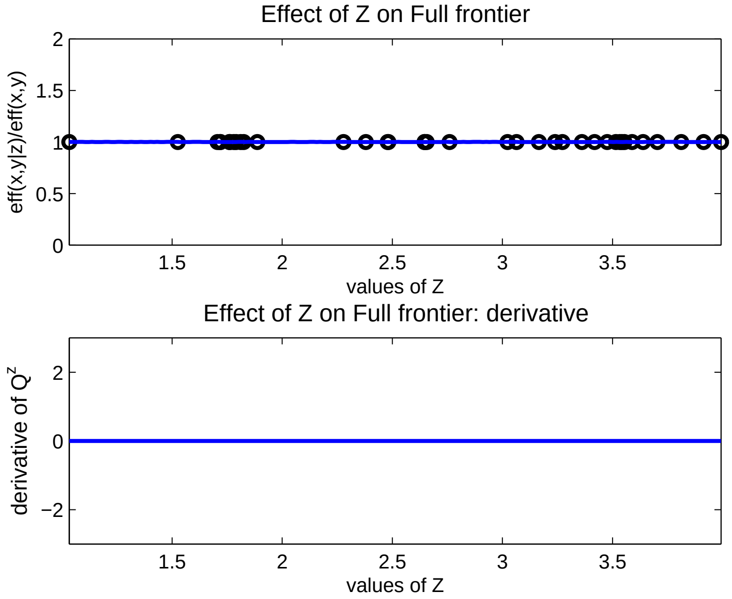

Figure 1: *Simulated example with univariate Z. Scatterplot and smoothed nonparametric regression of the ratios $Q^z = \widehat{\lambda}_n(x,y|z)/\widehat{\lambda}_n(x,y)$ on z (top panel) and the estimate of the derivative of Q^z with respect to z (bottom panel).*

4.1.2 Multivariate Z

For the second example we generated two independent uniform variables $Z_j \sim U(1,4)$ to build the bivariate variable $Z = (Z_1, Z_2)$. The influence of Z on the production process is described by:

$$Y_i^{(1)} = (1 + 2 * |Z_1 - 2.5|^3) * Y_{i,eff}^{(1)} * \exp(-U_i)$$
$$Y_i^{(2)} = (1 + 2 * |Z_1 - 2.5|^3) * Y_{i,eff}^{(2)} * \exp(-U_i).$$

Here we see that Z_1 pushes the efficient frontier above when far from 2.5, in both directions, with a cubic effect, while Z_2 has no effect on the frontier and so is irrelevant for the conditional measures. Note that there is no interaction between Z_1 and Z_2, the two variables being independent.

Units	N_D	$\hat{\lambda}_n(x,y)$	$\hat{\lambda}_n(x,y\|z)$	N_z	h
4	4	1.2204	1.2204	48	16.8015
12	2	1.1491	1.1491	7	18.7554
21	0	1.0000	1.0000	1	0.2878
25	1	1.1310	1.1310	5	19.1195
36	0	1.0000	1.0000	14	18.0273
45	6	1.2743	1.2743	57	2547.9574
67	0	1.0000	1.0000	43	16.9074
79	7	2.0093	2.0093	12	6871.8840
91	6	1.5215	1.5215	14	2.6379
98	1	1.0724	1.0724	14	18.0270
mean	1.6	1.1367	1.1367	25.3	823.5181

Table 1: *Results for 10 selected units from the simulated data in the case of univariate Z. N_D is the number of observations dominating the corresponding unit and N_z represents the number points used to estimate the conditional distribution given $Z = z$.*

Again, we simulate $n = 100$ observations according to this scenario. The mean value of the unconditional measure is 3.3032 compared to 1.6065, the mean value of the conditional output oriented measure, i.e. as expected, a global increase in efficiency.

The following Table summarizes the distribution of the obtained optimal bandwidths, the min, the 3 quartiles and the max values. Clearly, as expected, most of the values for h_2 produce oversmoothing (the range of Z_2 is between 1 and 4).

h_j	$\min(h_j)$	$Q_1(h_j)$	$Q_2(h_j)$	$Q_3(h_j)$	$\max(h_j)$
h_1	0.1366	0.5051	0.7209	0.8819	260.07
h_2	0.1319	2.6744	5.8308	1862.5	6925.4

Table 2: *Distribution of the optimal bandwidths.*

Figure 2 provides an even more detailed information on the impact of Z on the production process. It plots the ratios $\widehat{\lambda}_n(x,y|z)/\widehat{\lambda}_n(x,y)$ against z_1 and z_2. As we expected, we see a positive cubic effect of $|Z_1 - 2.5|$ and a flat behavior of the surface with respect to Z_2. The marginal effects can better be viewed in Figure 3 which shows the preceding surface regression evaluated at the observations (X_i, Y_i, Z_i) but viewed marginally as a function of each component Z_1 and Z_2 separately: here the cubic effect for Z_1 and the neutral effect for Z_2 clearly appears. The lines in the picture represent just a local average line of the points to stress the global effect.

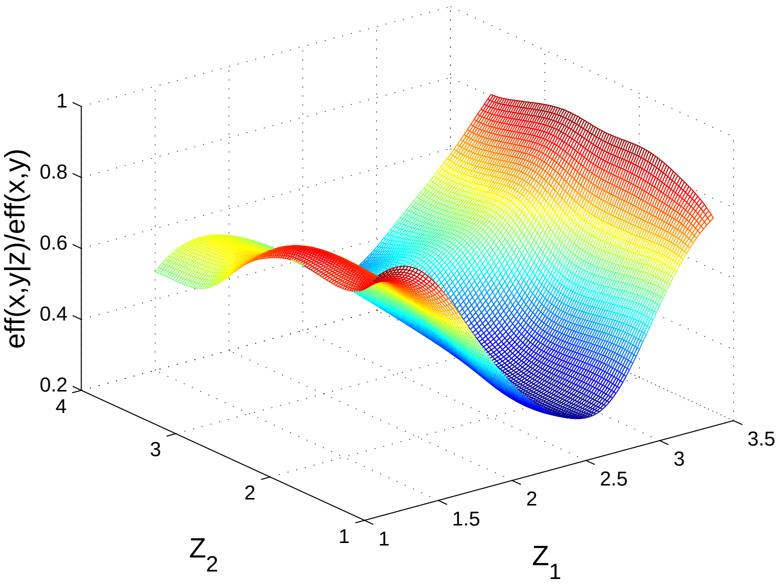

Figure 2: *Simulated example with multivariate Z. Smoothed nonparametric surface regression of $\widehat{\lambda}_n(x,y|z)/\widehat{\lambda}_n(x,y)$ on Z_1 and Z_2.*

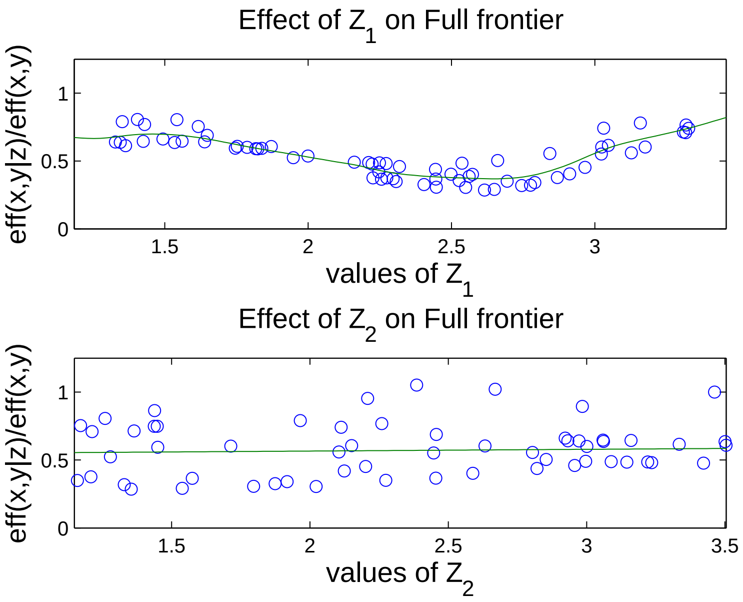

Figure 3: *Simulated example with multivariate Z. Marginal views of the surface regression of $\widehat{\lambda}_n(x,y|z)/\widehat{\lambda}_n(x,y)$ on z at the observed points (X_i,Y_i,Z_i), viewed as a function of Z_1 (top panel) and as a function of Z_2 (bottom panel).*

4.2 An illustration on real data

In the following we illustrate the performances of our method on real data consisting of a sample of 129 US Aggressive-Growth (AG) Mutual Funds for the year 2002. As in previous studies, Daraio and Simar (2005, 2006) and Jeong et al. (2008), an input oriented analysis was performed in order to evaluate the performance of Mutual Funds in terms of their risk (as expressed by standard deviation of Return) and transaction costs (measured in terms of Expense Ratio, Loads and Turnover) management. The traditional output to be considered in this framework is the Total Return of funds (the annual return expressed in percentage terms). In our illustration we use the Market Risks as environmental variable, to investigate its effect on our data, i.e. if it is favorable or unfavorable to the performance of mutual funds over the period under analysis. Hence, our analysis uses 3 inputs, 1 output, 1 environmental factor and 129 observations.

As far as the individual results are concerned, the bandwidths here are much smaller than those obtained in the previous studies. Of course, here we use a quartic kernel without any scaling. In the previous studies (using truncated Gaussian kernel), the bandwidths should be multiplied by the standard deviation of Z (=0.157) to make things comparable. For instance, the median value of the optimal bandwidth computed with truncated Gaussian kernel is 0.0591 (slightly smaller than the non-optimal value in Jeong et al, 2008). In our case here with quartic kernel we obtain a median value of 0.0117. The average value over the 129 observations of the input efficiency scores is 0.6083; for the conditional measures we reach the average value of 0.9595 (0.8442 in Jeong et al. , 2008). Table 3 gives a view of the distribution of the resulting optimal bandwidths.

h_j	$\min(h_j)$	$Q_1(h_j)$	$Q_2(h_j)$	$Q_3(h_j)$	$\max(h_j)$
h	0.0056	0.0093	0.0117	0.0305	6.0525

Table 3: *Distribution of the optimal bandwidths for the Mutual Funds example.*

Except for very few isolated data points, the optimal bandwidths are indeed very small with respect to the range of Z ($Z_i \in [0.0573, 0.9962]$), indicating certainly an influence of Z on the production process.

In order to detect a direction of the effect of the risk factor Z on the performance of the Mutual Funds we analyze again the observed values of the ratios $\widehat{\lambda}_n(X_i, Y_i|Z_i)/\widehat{\lambda}_n(X_i, Y_i)$ against Z_i. Figure 4 illustrates the influence of Z on the production process showing a global slight positive effect of the risk factor Z on the performance of mutual funds (remember here we are in an input orientation). We notice that this impact is more clear than the one detected in Daraio and Simar (2005) and Jeong et al. (2008), where the regression line was rather flat or even slightly increasing for small values of Z. By choosing an optimal bandwidth by the method proposed in this paper, we are able to detect a slight positive effect for all the values of Z, even if, all of this has to be taken with caution because the number of observations is too small to have definite conclusions (we have only 129 data points starting from a space of $p + q + d = 5$ dimensions). This is also more in accordance with traditional approaches where Market Risk is used as an input, underlying that the effect of market risks is conducive for mutual funds performance (Sengupta, 2000). The big difference here is that we do not impose such behavior by *a priori* assumptions, but it appears as a result from our methodology. Of course, the investigation on the statistical significance of this slight positive effect as it appears from Figure 4 is still an open question and stress the need for tools allowing inference and tests of hypothesis in this framework.

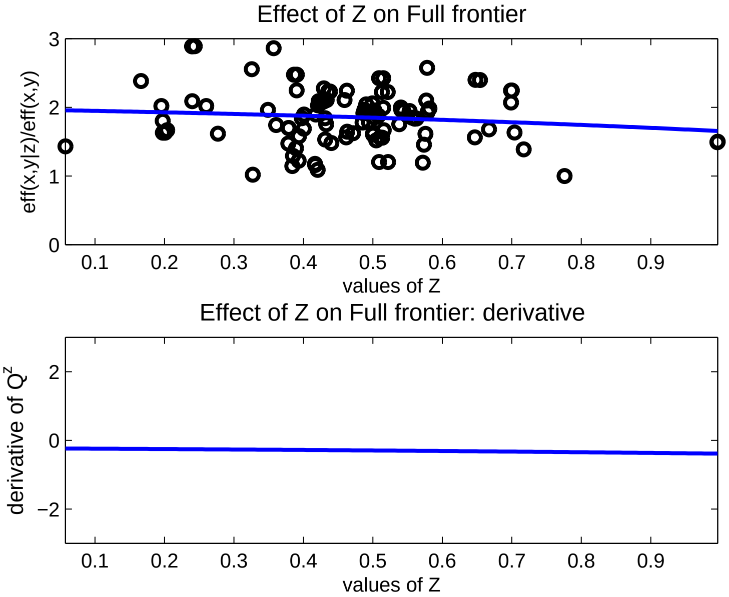

Figure 4: *Aggressive Growth US Mutual Funds. Scatterplot and smoothed nonparametric regression of the ratios* $Q^z = \widehat{\lambda}_n(x, y|z)/\widehat{\lambda}_n(x, y)$ *on* z *(top panel) and the estimate of the derivative of* Q^z *with respect to* z *(bottom panel).*

5 Conclusions

Conditional efficiency measures are very important for the analysis of the influence of external, environmental factors on the production process. As Daraio and Simar (2005, 2007a) and Jeong et al. (2008) pointed out, the selection of the bandwidth in this complex framework is a relevant open issue. In this paper we fill this gap in the literature and we propose a procedure for selecting the optimal bandwidth involved in the estimation of a non-standard distribution function on which nonparametric conditional efficiency estimates are based. The method is based on the extension of theoretical results obtained by Hall et al. (2004) and Li and Racine (2007, 2008) to the estimation of non-standard conditional distributions defining the conditional efficiency scores. The procedure allows separating the influential from the irrelevant factors, by assigning to the irrelevant ones large smoothing parameters.

We considered in this paper only the case where the environmental variables Z are continuous. The extension to the case of vectors Z having categorical and continuous components is straightforward and left as an exercise for the readers. The only difference is the use of appropriate kernels handling either ordered discrete variables or qualitative variables, the formula for computing the Least Squares Cross Validation criterion being the same as above. All the details on these special kernels can be found in Hall et al. (2004), Li and Racine (2007, 2008). As shown in our illustration with real data, a relevant direction for future research would be to develop tools for testing if some factors are significantly influential or not. This would involve bootstrapping procedures. We know that for FDH estimators subsampling is a consistent bootstrap procedure (see Jeong and Simar, 2006), but so far no data-driven procedures have been established to select the appropriate subsample size in this particular setup.

References

[1] Banker, R.D. and R.C. Morey (1986), Efficiency analysis for exogenously fixed inputs and outputs, *Operations Research*, 34(4), 513–521.

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

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

[4] Daraio, C. and L. Simar (2006), A robust nonparametric approach to evaluate and explain the performance of mutual funds, *European Journal of Operational Research*, 175(1), 516-542.

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

[6] Daraio, C. and L. Simar (2007b), Conditional nonparametric frontier models for convex and non convex technologies: a unifying approach, *Journal of Productivity Analysis*, vol 28, 13–32.

[7] Hall, P., Racine, J.S. and Q. Li (2004), Cross-Validation and the Estimation of Conditional Probability Densities, *Journal of the American Statistical Association*, vol 99, 486, 1015–1026.

[8] Jeong, S.O. , B.U. Park and L. Simar (2008), Nonparametric conditional efficiency measures: asymptotic properties, *Annals of Operations Research*, doi: 10.1007/s10479-008-0359-5.

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

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

[11] Li, Q. and J.S. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[12] Li, Q. and J.S. Racine (2008), Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data, forthcoming in *Journal of Business and Economic Statistics*.

[13] McDonald, J. (2008), Using least squares and tobit in second stage DEA efficiency analyses, *European Journal of Operational Research*, doi:10.1016/j.ejor.2008.07.039.

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

[15] Park, B., Simar,L. and V. Zelenyuk (2008), Local Likelihood Estimation of Truncated Regression and its Partial Derivatives: Theory and Application, *Journal of Econometrics*, vol 146, 1, 185–198.

[16] Reinhard, S. , C. A. Knox Lovell, G. J. Thijssen (2000), Environmental efficiency with multiple environmentally detrimental variables; estimated with SFA and DEA, *European Journal of Operational Research*, 121(2), 287-303.

[17] Sengupta, J. K. (2000), *Dynamic and Stochastic Efficiency Analysis, Economics of Data Envelopment Analysis*, World Scientific, Singapore.

[18] Silverman, B.W. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.

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

[20] Simar, L. (2007), How to Improve the Performances of DEA/FDH Estimators in the Presence of Noise?, *Journal of Productivity Analysis*, vol. 28(3), 183-201.

Appendix: Practical Implementation

In this appendix we provide a Matlab routine that evaluates the LSCV criterion (CV as given in equation (24)). For the computation of the convolution involved in the estimation of $\widehat{G}(x,z)$, we used Gaussian product kernels for Y. For the Z variable we use the multiplicative quartic kernel with a vector $h=(h_1,\ldots,h_r)$ of bandwidths, as described above. To avoid the computational burden, since optimal smoothing for Y is not of central interest, we suggest to define the bandwidths for the components in Y according to $h_{y_j}=h_0 s_{y_j}$, where s_{y_j} is the empirical standard deviation of the jth component of Y, and h_0 is a baseline bandwidth to select. This would simplify the search of the optimal bandwidths (h_y,h) to a $(1+r)$-dimensional search: (h_0,h). We present below the output oriented version, the notation in the routine is self-explanatory.

```
function  CV=Ker_LSCV_OUT(h,x,X,Y,Z,n,q,d)
%
%  Evaluate the LSCV criterion for estimating a conditional pdf
%  of (y |X<=x, Z=z) for baseline bandwidths hby and vector hz
%  OUTPUT ORIENTATION
%
hby=h(1);
hy=hby*std(Y); % this is a (1 x q) vector but hby is one-dimensional
wz=ones(n,1);
Q2x=zeros(n,1);
Q1x=zeros(n,1);
Cst1=(4*pi)^(q/2)* prod(hy);
Cst=1/Cst1;
Yh=Y*diag(ones(1,q)./hy);
YY=Yh*Yh';
DYY=diag(YY);
DYY1=kron(DYY,ones(1,n));
Convol=Cst*exp(-(1/4) * (DYY1 + DYY1' - 2*YY));
Xv=ones(n,1)*x;FlagXv=(X <= Xv);FlagX=all(FlagXv,2);
```

```
xv=ones(n-1,1)*x;
for i=1:n
    zi=Z(i,:);
    yi=Y(i,:);
    %     leave-one out sample
    Xi=X([(1:i-1)'; (i+1:n)'],:);
    Yi=Y([(1:i-1)'; (i+1:n)'],:);
    Zi=Z([(1:i-1)'; (i+1:n)'],:);

    flagx=(Xi<=xv);flagx1=all(flagx,2);
    tempz=(Zi-repmat(zi,n-1,1)); % this is a (n-1) x d matrix
    tempy=(Yi-repmat(yi,n-1,1)); % this is a (n-1) x q matrix
    tempyh=tempy*diag(ones(1,q)./hy);
    keryi=(exp(-(tempyh.^2)/2)/sqrt(2*pi))*diag(ones(1,q)./hy);% Individual Gaussian Kernel
    kery=prod(keryi,2); % Gaussian product kernel: kery is a (n-1) x 1 vector

    hz=h(2:d+1);
    tempzh=tempz*diag(ones(1,d)./(hz'));
    kerzi=(15/16)*((ones(size(Zi))-tempzh.^2).^2).*(abs(tempzh)<=1)*diag(ones(1,d)./(hz'));%
    kerz=prod(kerzi,2); % Quartic product kernel: (n-1) x 1 vector

    mxzi=mean(flagx1.*kerz);
    fyxzi=mean(flagx1.*kerz.*kery);
    if(mxzi <= 1.0e-06);
        Q2x(i)=0; Q1x(i)=0;
    else
    Q2x(i)= fyxzi*wz(i)/mxzi;

    integ=Convol([(1:i-1)'; (i+1:n)'],[(1:i-1)'; (i+1:n)']);
    Bigi1i2=kron(flagx1,flagx1').*kron(kerz,kerz').*integ;

    Gxzi=mean(mean(Bigi1i2));
    Q1x(i)=Gxzi*wz(i)/(mxzi^2);
    end

end

I1x=mean(Q1x.*FlagX);
I2x=mean(Q2x.*FlagX);
CV=I1x - 2*I2x ;
```