Estimating Nonparametric Conditional Frontiers and Efficiencies: A New Approach

Camilla Mastromarco
University of Calabria, CESifo and RCEA
Léopold Simar
Institut de Statistique, Université Catholique de Louvain.
Ingrid Van Keilegom
ORSTAT, KULeuven, Leuven, Belgium and ISBA, UCLouvain, Louvain-la-Neuve

77th European meeting of the Econometric Society - ESEM 2024
Erasmus School of Economics, Rotterdam, Netherlands
August 26-30, 2024

Motivation

$\rightarrow$ In production theory, conditional frontiers and conditional efficiency measures are a flexible and appealing approaches to consider the role of environmental variables on the production process.

$\rightarrow$ **Effects of external factors on efficiency of units:**

- Direct Approach: estimate non-parametrically conditional distribution functions requiring smoothing techniques and the use of selected bandwidths.
  - the statistical literature produces way to derive bandwidths of optimal order, by using e.g. least-squares-cross-validation techniques.
  - it has been shown that the resulting order may not be optimal when estimating the boundary of the distribution function.

Motivation

- Indirect Approach: a full nonparametric approach which avoids the problem of estimating these bandwidths
  - by eliminating in a first step the influence of the environmental factors on the inputs and the outputs. By doing this we produce "pure" inputs and outputs which allow to estimate a "pure" measure of efficiency, more reliable for ranking the firms, since the influence of the external factors have been eliminated.
  - Our approach can be viewed as an extension of the use of location-scale models (implying some semi-parametric structure) to full nonparametric models, based on nonseparable, nonparametric models. We are also able to recover the frontier and efficiencies in original units.

$\rightarrow$ **We describe the method, its statistical properties and we show in some Monte-Carlo simulations, how our new method dominates the traditional direct approach and the location scale model.**

Direct Approach: Conditional Frontier and Efficiency

$\rightarrow$ (Cazals et al. 2002, Daraio and Simar 2005):

1. We consider a probabilistic formulation of the production process where the random variables (X, Y, Z) are defined on an appropriate probability space. The conditional distribution of the input-outputs (X, Y) given particular values of Z can be described by the conditional survival function

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

   where $S_{X|Y,Z}(x|y, z) = \mathbb{P}(X \geq x|Y \geq y, Z = z)$ and $S_{Y|Z}(y|z) = \mathbb{P}(Y \geq y|Z = z)$. Note the unusual conditioning on Y in $S_{X|Y,Z}(x|y, z)$, because Y is an output.

2. The conditional minimum input (or cost) frontier is then defined as the minimal achievable input level x for units producing at least the level y of outputs, facing the environmental conditions z:

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

Direct Approach: Conditional Frontier and Efficiency

(Cazals et al. 2002, Daraio and Simar 2005):

1. Partial frontiers have also been introduced, considering less extreme benchmarks, and so providing estimators more robust to outliers and extreme data points. The order-α quantile frontier introduced by Aragon et al. (2005) and Daouia and Simar (2007). Their conditional version are defined for any $\alpha \in [0, 1]$

$$\tau_\alpha(y, z) = \inf\{x | S_{X|Y,Z}(x|y, z) < \alpha\},$$

where we see clearly that when $\alpha \to 1$, $\tau_\alpha(y, z) \to \tau(y, z)$.

2. the order-m conditional frontier introduced by Cazals et al. (2002). Here for a given integer $m \geq 1$ we have

$$\begin{aligned} \tau_m(y, z) &= \mathbb{E}[\min(X_1, \ldots, X_m) | Y \geq y, Z = z] \\ &= \int_0^\infty S^m_{X|Y,Z}(x|y, z)\, dx, \end{aligned} \tag{2}$$

where as $m \to \infty$, $\tau_m(y, z) \to \tau(y, z)$. → **The idea here is to benchmark the input of a unit producing the value y of outputs and facing environmental conditions z not against the minimal possible input of such firms but the average of m peers facing the same conditions z and producing at least the value y of outputs**

3. Suppose we have data (x_i, y_i, z_i) for the cross section unit $i = 1, \ldots, n$, the unconditional and conditional attainable sets can be estimated and nonparametric estimator of the conditional survival function $S_{X,Y|Z}(x, y|z)$ and of $\tau_\alpha(y, z)$ could be obtained (as e.g. in Badin et al. , 2010).

Direct Approach: Conditional Frontier and Efficiency

- The statistical properties of the resulting frontier estimators are well established: see Park et al. (2000), Cazals et al. (2002).

$$n^{1/(1+q)}\big(\widehat{\tau}(y) - \tau(y)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_y^{(1+q)}, 1+q) \tag{3}$$

$$\sqrt{n}\big(\widehat{\tau}_m(y) - \tau_m(y)\big) \xrightarrow{\mathcal{L}} \text{N}(0, \sigma_y^2), \tag{4}$$

  where exact expressions for the parameters of the limiting distributions have been derived.

- For the conditional to Z frontiers, it has been proven (see Cazals et al., 2002 and Jeong et al., 2010) that under mild regularity conditions, we have similar results where the sample size n has to be replaced by its effective number of observations in a neighborhood of z, namely $n\bar{h}_z$ where $\bar{h}_z = \prod_{j=1}^{d} h_{z_j}$.

$$(n\bar{h}_z)^{1/(1+q)}\big(\widehat{\tau}(y,z) - \tau(y,z)\big) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_{y,z}^{(1+q)}, 1+q) \tag{5}$$

$$\sqrt{n\bar{h}_z}\big(\widehat{\tau}_m(y,z) - \tau_m(y,z)\big) \xrightarrow{\mathcal{L}} \text{N}(0, \sigma_{y,z}^2), \tag{6}$$

- Similar results hold for the order-α and conditional order-α frontiers to their equivalent order-m (see Daouia and Simar, 2006, for details).

Direct Approach: Conditional Frontier and Efficiency

- In this traditional or "direct" approach for estimating conditional frontiers, the statistical properties rely on the properties of the bandwidths h_z used for estimating the conditional survival function (1).
- Least squares cross-validation (LSCV) techniques are available providing bandwidths with the optimal order, i.e. $h_{z_j} \propto n^{-1/(d+4)}$ (see Li et al., 2013) which deteriorates the convergence rates since n is replaced by $n^{4/(d+4)}$, in particular when d increases.
- As pointed in Bădin et al.(2019), these bandwidths might not be optimal when the objective is to estimate the lower bound of the support of $S_{X,Y|Z}(x, y|z)$ in the x direction. The problem, already noticed by Jeong et al. (2010), is that for a given h_z, the conditional FDH estimator does not target $\tau(y, z)$ but rather $\tau^{h_z}(y, z)$ defined as

$$\tau^{h_z}(y, z) = \inf \left\{ x | S^{h_z}_{X|Y,Z}(x|y, z) = \mathbb{P}\big(X \geq x \mid Y \geq y, |Z - z| \leq h_z\big) < 1 \right\} \quad (7)$$

  where $|Z - z| \leq h_z$ has to be understood component by component, i.e. $|Z_j - z_j| \leq h_{z_j}$, $j = 1, \ldots, d$.
- This introduces an additional error, similar to a bias of localization, which under mild regularity condition (smoothness of the frontier as a function of z) is of order $||h||$, unless the separability condition holds. Since in practice, with real data, we do not know if this separability condition holds, we may have problem and the LSCV might not be optimal.

Indirect Approach: Location-scale model

$\rightarrow$ Florens et al. (2014): clear I/O variables by flexible Location-Scale Models:

$$\begin{cases} X_i &= \mu_x(Z_i) + \sigma_x(Z_i)\varepsilon_{x,i} \\ Y_i &= \mu_y(Z_i) + \sigma_y(Z_i)\varepsilon_{y,i} \end{cases}, \tag{8}$$

μ_x, σ_x and ε_x have each p components and, for ease of notations, the product of vectors is componentwise. They assume that each element of ε_x and ε_y have mean zero and standard deviation equal to 1. The model also assume that $(\varepsilon_x, \varepsilon_y)$ is independent of (Z).

1. estimation of the location functions $\mu_\ell(z_i)$;
2. estimation of the variance functions $\sigma^2_\ell(z_i)$ by regressing the resulting square residuals of the first step on (z_i). For the first step they use local linear and for the second step local constant to avoid negative values of the estimated variances.

Location-scale model, contd.

→ Florens, Simar and van Keilegom (JoE2014):

- From this first analysis they obtain the residuals

$$\widehat{\varepsilon}_{1,i} = \frac{X_i - \mu_1(z_i)}{\widehat{\sigma}_1(z_i)},$$
$$\widehat{\varepsilon}_{2,it} = \frac{Y_i - \mu_2(z_i)}{\widehat{\sigma}_2(z_i)}$$

- $\widehat{\varepsilon}_{1,i}$ and $\widehat{\varepsilon}_{2,i} \perp (z_i)$ → can be tested see Florens et al. (2014)
- These are the whitened inputs and output obtained by eliminating the influence of the external and other environmental variables as common factors.
- our method can be viewed as an extension of the use of location-scale models (implying some semi-parametric structure) to full nonparametric models, based on nonseparable, nonparametric models.

Our approach:

- The location model is quite flexible, semiparametric but still assume this additive form for the input and each output and more importantly the joint independence between $(\varepsilon_x, \varepsilon_y)$ and Z;
- we extend the location-scale model to a more general structure where the link between the input and outputs and the environmental factors Z, is described by fully nonparametric models.
- the location-scale models of Florens et al. (2014) can be viewed as a particular semiparametric case of our models. We will show that by doing this extension we do not lose the nice properties of the location-scale approach.
- the resulting estimators of the pure efficiency measures are:
  1. free of the curse of dimensionality due to the dimension of Z and converge with rate $\sqrt{n}$ to a Gaussian process;
  2. for the robust frontier estimates, we keep the $\sqrt{n\bar{h}_z}$-convergence to a Gaussian process, as is the case for the estimators derived by the direct approach. However, in our case here, we do not have to rely on the location-scale assumption.

Our approach:

We suppose the vector (X, Y, Z) follows the following model:

$$\begin{cases} X = \varphi_x(Z, U_x) \\ Y_j = \varphi_{y_j}(Z, U_{y_j}) \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{9}$$

where the functions $\varphi_\ell(\cdot, U_\ell)$ are nonseparable and monotone increasing in U_ℓ, $\ell = x, y_1, \ldots, y_q$.

- we assume that the U_ℓ are uniformly distributed on $[0, 1]$ then φ_ℓ can be interpreted as a quantile function. The choice of the uniform is a matter of rescaling the U_ℓ to get this nice interpretation;
- U_ℓ are independent of Z, this assumption is part of the model and is needed to identify each individual equation in (9).
- The variables U are constructed as being the part of the input (respectively outputs) which is independent on Z. In other words they are whitened versions of X and Y defined in a set of general nonparametric nonseparable equations.
- In the same lines of Florens et al. (2014), we interpret (U_x, U_y) as "pure" input and outputs which remain monotone transformations of the original measures. It can also be seen as the part of the input and outputs not dependent on Z. So, here again we will be able to build the efficient frontier in the pure input-output space, allowing to define an efficiency score, or a "managerial" efficiency, that will be independent of the environmental conditions.

Our approach:

It is known that under the above assumptions, the U_ℓ are identified by the conditional distribution of the input and the outputs given Z.

$$\begin{aligned} F_{X|Z}(x|z) &= \mathbb{P}(X \leq x|Z = z) \\ &= \mathbb{P}(\varphi_x(Z, U_x) \leq x|Z = z) \\ &= \mathbb{P}(U_x \leq \varphi_x^{-1}(Z, x)|Z = z) \\ &= \mathbb{P}(U_x \leq \varphi_x^{-1}(z, x)) \\ &= \varphi_x^{-1}(z, x), \end{aligned}$$

where the last line is obtained because we assume $U_x \sim \mathrm{Unif}([0, 1])$. Since this is true for all (x, z), we have $U_x = F_{X|Z}(X|Z)$ with probability one. So, more generally we have:

$$\begin{cases} U_x = F_{X|Z}(X|Z) \\ U_{y_j} = F_{Y_j|Z}(Y_j|Z) \quad \text{for } j = 1, \ldots, q, \end{cases} \tag{10}$$

So we see that $\varphi_x(Z, U_x) = F^{-1}_{X|Z}(U_x|Z)$, i.e. the conditional quantile of X given Z evaluated at $U_x \in [0, 1]$. The same is true for $Y_1, \ldots, Y_q$, each function $\varphi_{y_j}(Z, U_{y_j}) = F^{-1}_{Y_j|Z}(U_{y_j}|Z)$, $j = 1, \ldots, q$ has the same conditional quantile interpretation.

Our approach:

Since the functions φ_ℓ are unknown, the values $U_{\ell,i}$ are not observed but they can be estimated by nonparametric methods by estimating the appropriate conditional distribution functions:

$$\widehat{U}_{x,i} = \widehat{F}_{X|Z}(X_i|Z_i) = \frac{\sum_{k=1}^{n} G_{h_x}(X_k - X_i)K_{h_z}(Z_k - Z_i)}{\sum_{k=1}^{n} K_{h_z}(Z_k - Z_i)}, \tag{11}$$

$$\widehat{U}_{y_j,i} = \widehat{F}_{Y_j|Z}(Y_{j,i}|Z_i) = \frac{\sum_{k=1}^{n} G_{h_{y_j}}(Y_{j,k} - Y_{j,i})K_{h_z}(Z_k - Z_i)}{\sum_{k=1}^{n} K_{h_z}(Z_k - Z_i)}, \tag{12}$$

where now optimal bandwidths h_z, h_x and h_{y_j} can be obtained for each equation, by the LSCV techniques described in Li et al. (2013) and we avoid the problem of selecting optimal bandwidths h_z when estimating the boundary of some conditional distribution function and all the issues mentioned above.
Here the kernels used are standard: $K_{h_z}(\cdot)$ are usual kernels for estimating densities and $G_{h_\ell}(\cdot)$ are cumulative kernels used for estimating distribution functions (cdf).

Our approach:

Having these input and outputs in pure units, we can estimate the minimal cost frontier and its order-m robust version, by usual techniques. For the full frontier, it is defined in pure units by

$$\phi(u_y) = \inf\{u_x | S_{U_x|U_y}(u_x|u_y) < 1\}, \tag{13}$$

where $S_{U_x|U_y}(u_x|u_y) = \mathbb{P}(U_x \geq u_x | U_y \geq u_y\}$. So $\phi(u_y)$ is the minimal achievable level of input in pure units, for units producing at least the level of output u_y in pure units.
For the order-m frontier, we have for a given m

$$\begin{aligned}\phi_m(u_y) &= \mathbb{E}\left[\min(U_{x,1}, \ldots, U_{x,m}) | U_y \geq u_y\right], \\ &= \int_0^1 S^m_{U_x|U_y}(u_x|u_y)\, du_x, \end{aligned} \tag{14}$$

which provides also, for finite m, a less extreme benchmark than the full frontier $\phi(u_y)$.
Here, as shown in Cazals et al. (2002), as $m \to \infty$, we have $\phi_m(u_y) \to \phi(u_y)$.
Note that from the frontiers in the pure units we can recover the frontiers in the original units.

Numerical Illustration

We conducted a series of Monte Carlo simulations to assess the performance of our new nonparametric method for estimating conditional frontiers and efficiencies. These simulations compare our approach with the traditional direct nonparametric method and the location-scale method under various scenarios. We first illustrate results from three straightforward examples (columns three to five in Table 1), where our method is evaluated against the direct nonparametric approach and the location-scale method. Additionally, a more complex and realistic example is considered (column six of Table 1). The performance of order-m estimators and the impact of outliers are further explored.
In all the numerical examples, the optimal bandwidths have been computed for each sample by least-squares cross validation, for each nonparametric regressions in the location-scale approach and for each estimation of conditional distribution functions.

- The first case, referred to as a "Toy" example (third column in Table 1), involves a simple classic frontier model where the external factor Z only influences the density of inefficiencies (see e.g. Kimbhakar and Loovell, 2000).
- Next, we examine a scenario where Z is fully independent of X and Y (fourth column in Table 1), meaning Z does not influence the production process.
- In a more general model fitting the location-scale assumptions (fifth column in Table 1), we aim to explore the performance of our new method under conditions favourable to the location-scale model. The model is a bit artificial but indeed the location-scale assumptions are less natural in a frontier model setup, even if these models can be viewed as flexible approximations of the DGP.
- Inspired by the DGPs proposed by Badin et al. (2019) and Simar and Wiilson (2011), we further illustrate the performance of various approaches under different conditions where the external variable Z influences the production process. Specifically, the effect of Z may be observed on the level of the frontier, on the distribution of inefficiencies, on both, or on neither. Here, we focus on one specific case, where Z is bivariate and affects both the frontier level and the inefficiency distribution (this is "Complex Scenario" in the sixth column of table 1). We analyze the impact of introducing three outliers into the dataset. We also examine the behavior of the order-m partial frontiers under these conditions. In the paper, we explore these four cases across various scenarios involving one input X (where X represents a cost) and one output Y, considering both univariate and bivariate Z.

Numerical Illustration

Now the idea is to see if we can correctly estimate the true values $\tau(Y_i, Z_i)$ in each case. Note that the true value is given by the chosen scenario, so we can look at the estimate of the *ISE* (Integrated Squared Error) given by

$$ISE = n^{-1} \sum_{i=1}^{n} (\widehat{\tau}(Y_i, Z_i) - \tau(Y_i, Z_i))^2 \tag{15}$$

We will do this for $n = 100, 200$ and 500. We estimate the Mean Integrated Squared Error (*MISE*) by doing 500 Monte-Carlo repetitions and averaging the *ISE* over the 500 trials. To check if some differences are significant we give also the estimate of the standard deviation of the Monte-Carlo estimator of the *MISE*.
We first illustrate (table 1) on 3 simple examples how our method performs compared to the direct nonparametric method and compared to the location-scale method. Then (table 1) in a more realistic example we will also consider order-m estimators and investigate the effect of the presence of outliers. In all the numerical examples, the optimal bandwidths have been computed for each sample by least-squares cross validation, for each nonparametric regressions in the location-scale approach and for each estimation of conditional distribution functions.

Numerical Illustration

Table: Performance comparison of different methods *MISE* (standard deviations in parentheses).

Sample Size	Method	'Toy" Example	Full Independence	Location-Scale	Complex Scenario
100	Direct	0.2459 (0.0080)	1.6502 (0.0762)	0.0410 (0.0013)	1.4624 (0.1510)
100	New Method	0.1837 (0.0045)	0.7281 (0.0227)	0.0181 (0.0004)	0.7701 (0.0264)
100	Loc-Scale	0.3187 (0.0081)	1.1930 (0.0606)	0.0240 (0.0008)	2.1966 (0.1836)
200	Direct	0.1749 (0.0058)	0.9712 (0.0408)	0.0289 (0.0009)	1.0943 (0.0569)
200	New Method	0.1407 (0.0030)	0.4480 (0.0151)	0.0109 (0.0003)	0.5825 (0.0161)
200	Loc-Scale	0.2491 (0.0065)	0.6333 (0.0287)	0.0141 (0.0004)	1.6307 (0.0761)
500	Direct	0.1001 (0.0026)	0.5651 (0.0271)	0.0176 (0.0004)	0.7965 (0.0372)
500	New Method	0.1098 (0.0022)	0.2339 (0.0078)	0.0054 (0.0001)	0.5089 (0.0129)
500	Loc-Scale	0.1948 (0.0040)	0.3178 (0.0147)	0.0063 (0.0002)	1.3869 (0.0608)

Numerical Illustration

- The first case, referred to as a "Toy" example (third column in Table 1): our method consistently outperformed the other methods, displaying lower MISE values across different sample sizes (n = 100, 200, 500). The direct traditional approach suffers from issues mentioned and as noted by Florens et al. (2014). Although the performance of the direct approach improves with larger sample sizes, it still cannot match our new method. Notably, the location-scale model underperforms in this scenario, as it does not fit the necessary assumptions.
- Full Independence: (fourth column in Table 1): Despite this independence, our method again outperformed both the traditional and location-scale methods. Interestingly, even though the location-scale model is appropriate in this case, our method still performs better than the location-scale approach. This comes probably from the fact that in the location-scale approach, we first estimate the location by a local linear estimator then in a second step, we regress on Z, in a nonparametric way, the squares of the residuals obtained from estimating the location function. Squaring the residuals introduces some instability in the estimation process. But still, the location-scale approach dominates the traditional direct approach. The Monte Carlo standard deviations indicate that these differences are statistically significant.
- Location-Scale: Here again, and surprisingly, our new approach outperforms the location-scale approach, likely due to the reasons mentioned earlier. However, as expected, the location-scale method outperforms the traditional direct approach in this scenario.
- Complex Scenario: The results demonstrate that our new method consistently outperforms both the location-scale method and the direct approach, particularly in complex cases where Z affects both the frontier and efficiency distribution. We also observe that even without outliers, the order-m estimators provide more accurate estimates of the full frontier (as discussed in the paper). This is because they are less sensitive to extreme data points that can jeopardize the full frontier estimates.

Numerical Illustration: Conclusion

In summary (including all the cases examined in the paper), our new method based on the control functions approach consistently outperforms, as expected, the traditional direct approach and also, the location-scale model. The latter is too restrictive, as it assumes Z only influences the first two moments of the input and outputs. Even when the location-scale model is true, our new method shows better *MISE* performance, likely due to the instability introduced in the second-stage nonparametric regression for estimating scale functions. Additionally, our method generally exhibits lower Monte Carlo standard deviations of *MISE*, indicating greater statistical stability. These observations also apply to the order-m estimates.

Real Data: Banking Sector

- banking sector, from Simar and Wilson (2007); also used in Bãdin et al. (2012) using the direct traditional approach and in Florens et al. (2014), hereafter FSVK, using location scale.
- Sample of 303 banks which contains 3 inputs (purchased funds, core deposits and labor) and 4 outputs (consumer loans, business loans, real estate loans and securities held) for banks. Two environmental factors are considered, the size of the banks Z_1 (the log of the total assets) and a measure of the diversity of the services proposed by the banks Z_2.
- Daraio et al. (2018) using the same data set rejected the separability condition, advocating the use of conditional efficiency measures.
- Following the other papers, by using the methodology described in Daraio and Simar (2007), the inputs can be aggregated in a one- dimensional input measure, without losing much information and the same is true for the outputs. The final output Y is highly correlated (more than 0.93) with the 4 original outputs and the same is true for the final input X (correlation with the original inputs more than 0.97). This allows to illustrate the results in low-dimensional pictures.

Banking Sector: Results

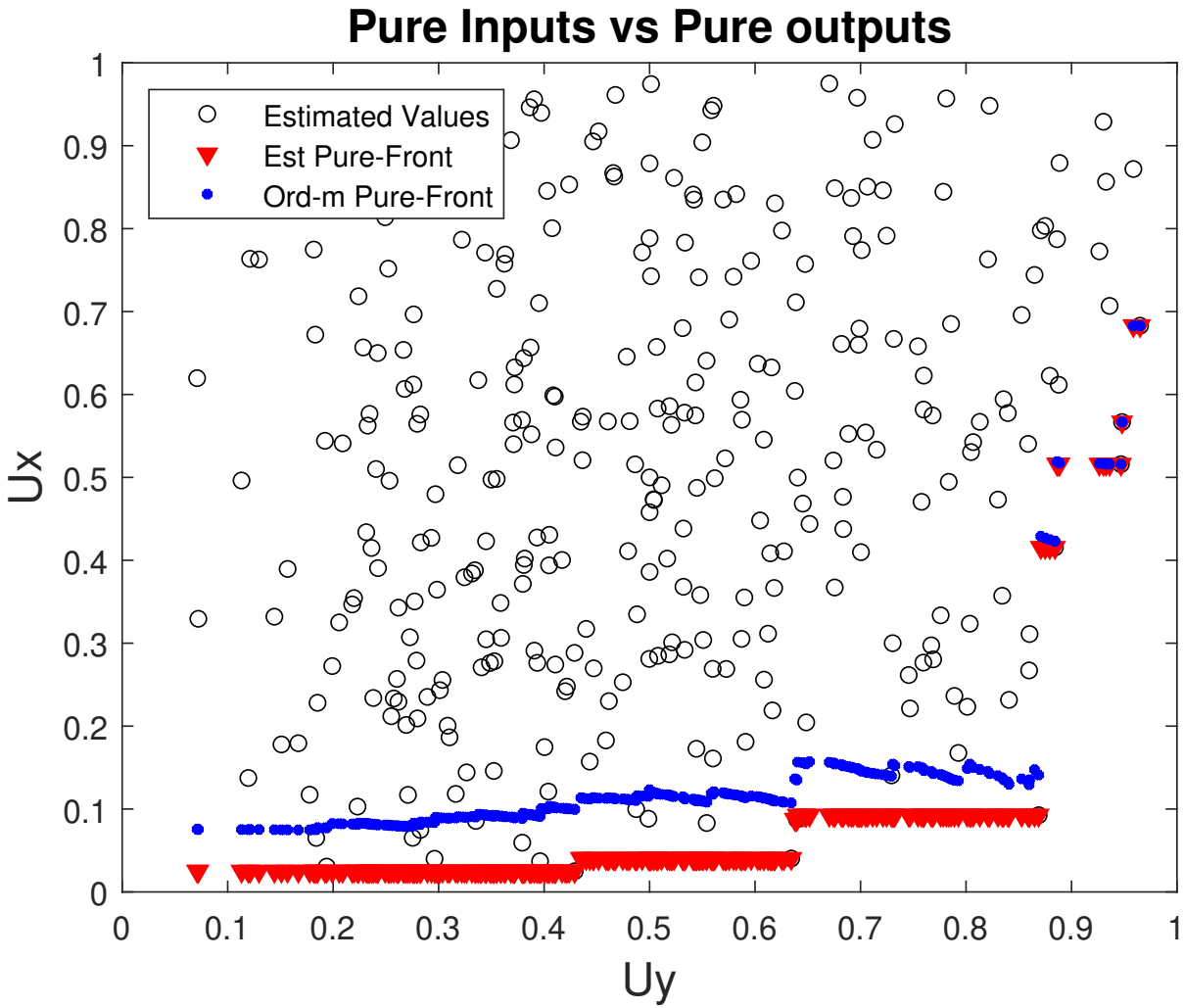

Figure: Bank example: Estimated "pure" inputs and outputs and the estimated efficient frontiers $\widehat{\phi}(\hat{u}_y)$ and $\widehat{\phi}_m(\hat{u}_y)$, here $m = 30$.

Banking Sector: Results

We can compare this figure with Figure 10 in FSVK:

- we see that the full frontier in Figure 4 envelops nicely the cloud of data in pure units, which is less the case in Figure 10 of FSVK, with the same qualitative remark for the order-m frontier.
- Our approach is fully nonparametric, whereas FSVK is based on a semiparametric model, imposing the location-scale models for cleaning the data.
- We know also from our Monte-Carlo experiments that our approach is statistically and numerically very stable.
- The FSVK location-scale approach involves in a second step, for the scale model, a nonparametric smoothing of the squares of the residuals from the nonparametric location estimation, and this is more sensitive to some extreme values.

Banking Sector: Results

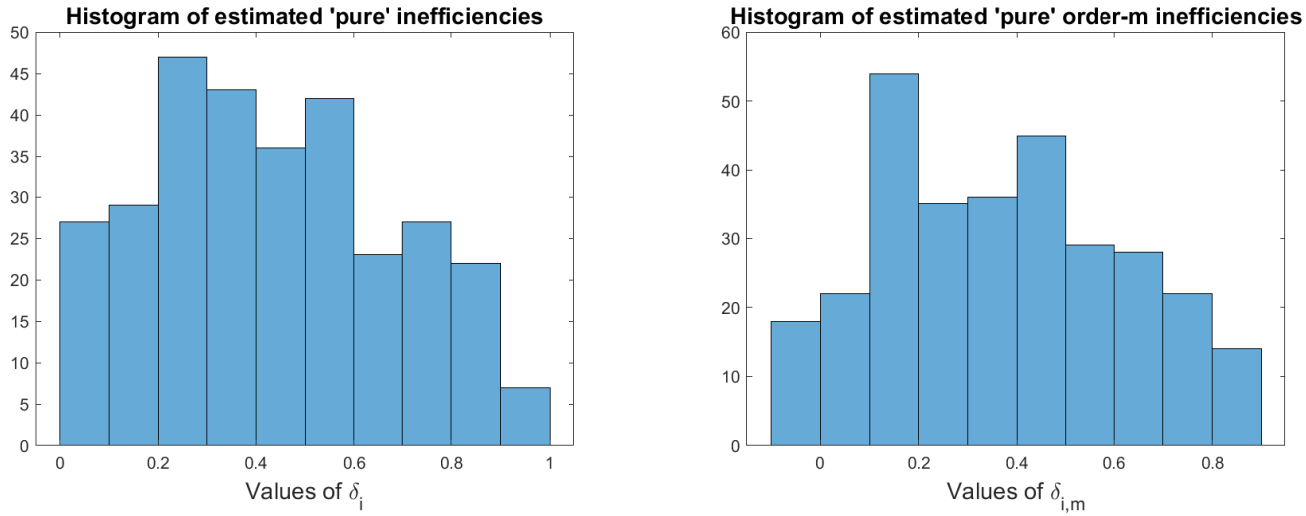

Figure: Bank example: distribution of the estimated "pure" inefficiencies, relative to the frontier estimates (full and order-m).

"Pure" efficiency scores, i.e. $\delta_i = \hat{u}_{x_i} - \widehat{\phi}(\hat{u}_{y_i})$ and $\delta_{i,m} = \hat{u}_{x_i} - \widehat{\phi}_m(\hat{u}_{y_i})$. This can be compared with Figure 11 in FSVK, but the comparison is difficult since our pure units and the "pure" frontiers are somewhat different as explained above. Still our approach is more general and does not rely on some semiparametric assumptions.

Banking Sector: Results

Table: Results for 15 randomly selected banks, as in FSVK. Data values and estimates of the frontier levels at the data points. Full and order-m with $m = 30$: "old" is for the traditional direct method and "new" is for our new method.

Unit i	X_i	Y_i	Z_{1i}	Z_{2i}	$\widehat{\tau}_{old}(Y_i, Z_i)$	$\widehat{\tau}_{new}(Y_i, Z_i)$	$\widehat{\tau}_{m,old}(Y_i, Z_i)$	$\widehat{\tau}_{m,new}(Y_i, Z_i)$
259	7.2986	8.5127	1.1778	1.1388	7.2986	6.2393	7.2986	6.2432
237	0.3505	0.5186	0.9409	0.8371	0.3505	0.3505	0.3505	0.4047
258	0.1998	0.2958	0.8700	1.2650	0.1998	0.1604	0.1998	0.1830
1	1.1985	1.3999	1.0199	0.8539	1.1038	1.0134	1.1038	1.0565
241	0.8693	0.8173	0.9863	0.8737	0.8693	0.6903	0.8693	0.7254
66	0.3421	0.3546	0.9028	0.8407	0.2874	0.2396	0.2876	0.2592
164	1.8694	2.1868	1.0551	1.1069	1.8694	1.6086	1.8694	1.6930
274	0.4026	0.4185	0.9152	1.0590	0.3589	0.2715	0.3589	0.2990
303	0.2969	0.5053	0.9024	1.1611	0.2969	0.2625	0.2969	0.2725
199	2.6751	2.9132	1.0786	0.9124	2.6751	2.5152	2.6751	2.5158
216	7.2741	6.6494	1.1670	1.0387	7.2741	5.7449	7.2741	5.8366
125	1.0559	1.1407	1.0111	0.6974	1.0559	0.8559	1.0559	0.9117
239	1.3945	1.7164	1.0330	1.2376	1.2875	1.1799	1.2875	1.2690
170	2.9572	2.4389	1.0807	0.8205	2.4650	1.9072	2.4650	2.1360
242	1.8388	2.1842	1.0735	0.8725	1.8388	1.8138	1.8388	1.9190

Banking Sector: Results

Table: Pure and conditional efficiency scores of order-m, for the same 15 units, as computed in FSVK, i.e. $\widehat{\delta}_{m,i} = \widehat{U}_{x,i} - \widehat{\phi}_m(\widehat{U}_{y,i})$ and $\widehat{\theta}_m(Y_i, Z_i) = \widehat{\tau}(Y_i, Z_i)/X_i$. The ranks are computed relative to the pure order-m efficiency scores, $m = 30$.

Unit i	Rank$(\widehat{\delta}_{m,i})$	$\widehat{\delta}_{m,i}$	$\widehat{\theta}_{m,old}(Y_i, Z_i)$	$\widehat{\theta}_{m,new}(Y_i, Z_i)$
259	41	0.1017	1.0000	0.8554
237	1	-0.0745	1.0000	1.1548
258	37	0.0743	1.0000	0.9157
1	120	0.2848	0.9210	0.8816
241	150	0.3514	1.0000	0.8344
66	218	0.5225	0.8408	0.7577
164	104	0.2478	1.0000	0.9056
274	206	0.4826	0.8916	0.7428
303	53	0.1378	1.0000	0.9177
199	107	0.2558	1.0000	0.9404
216	271	0.7154	1.0000	0.8024
125	108	0.2562	1.0000	0.8634
239	145	0.3354	0.9233	0.9100
170	297	0.8274	0.8335	0.7223
242	7	-0.0438	1.0000	1.0436

Banking Sector: Results

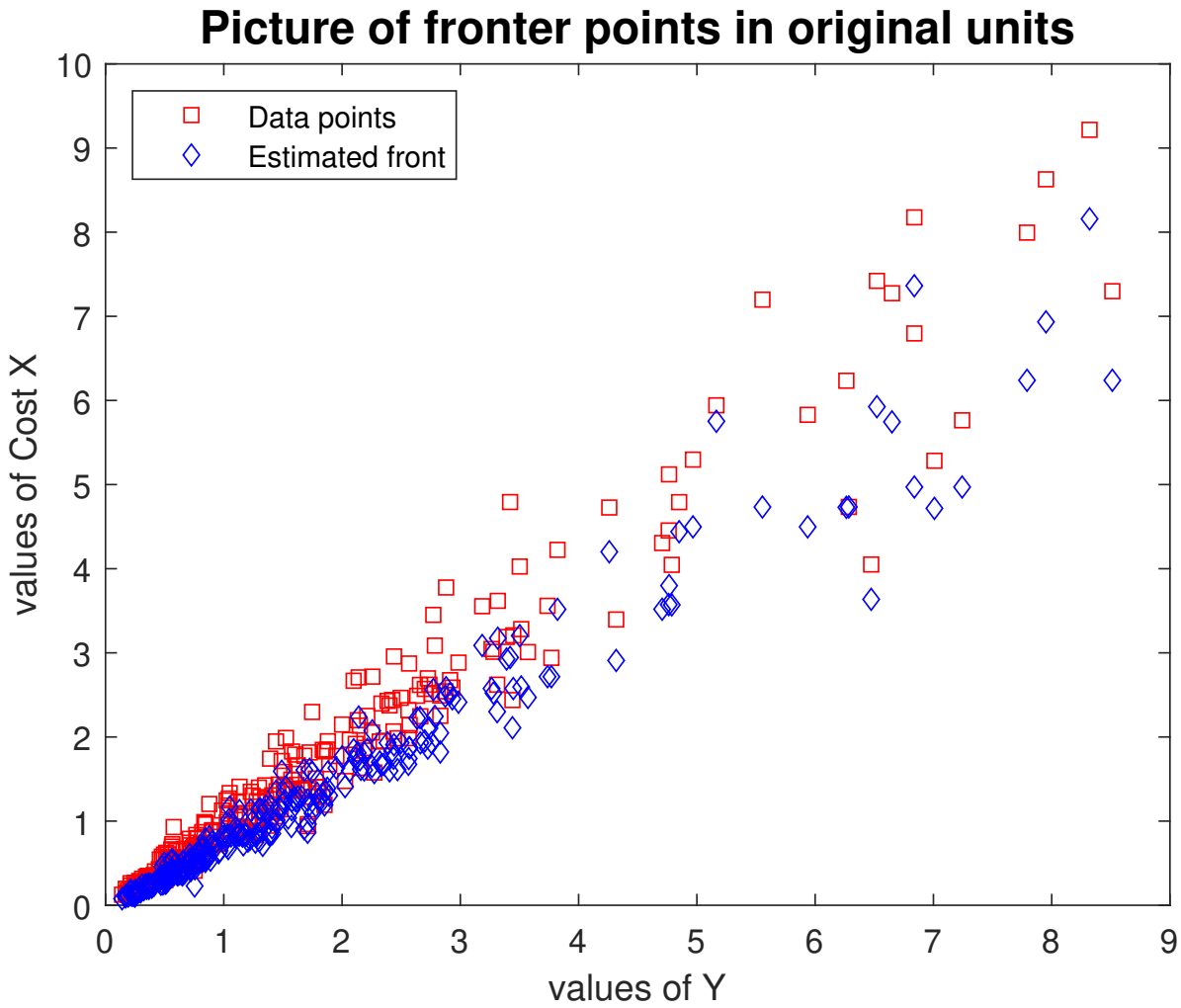

Figure: Bank example: data points in original units and estimated $\widehat{\tau}(Y_i, Z_i)$.

Banking Sector: Results

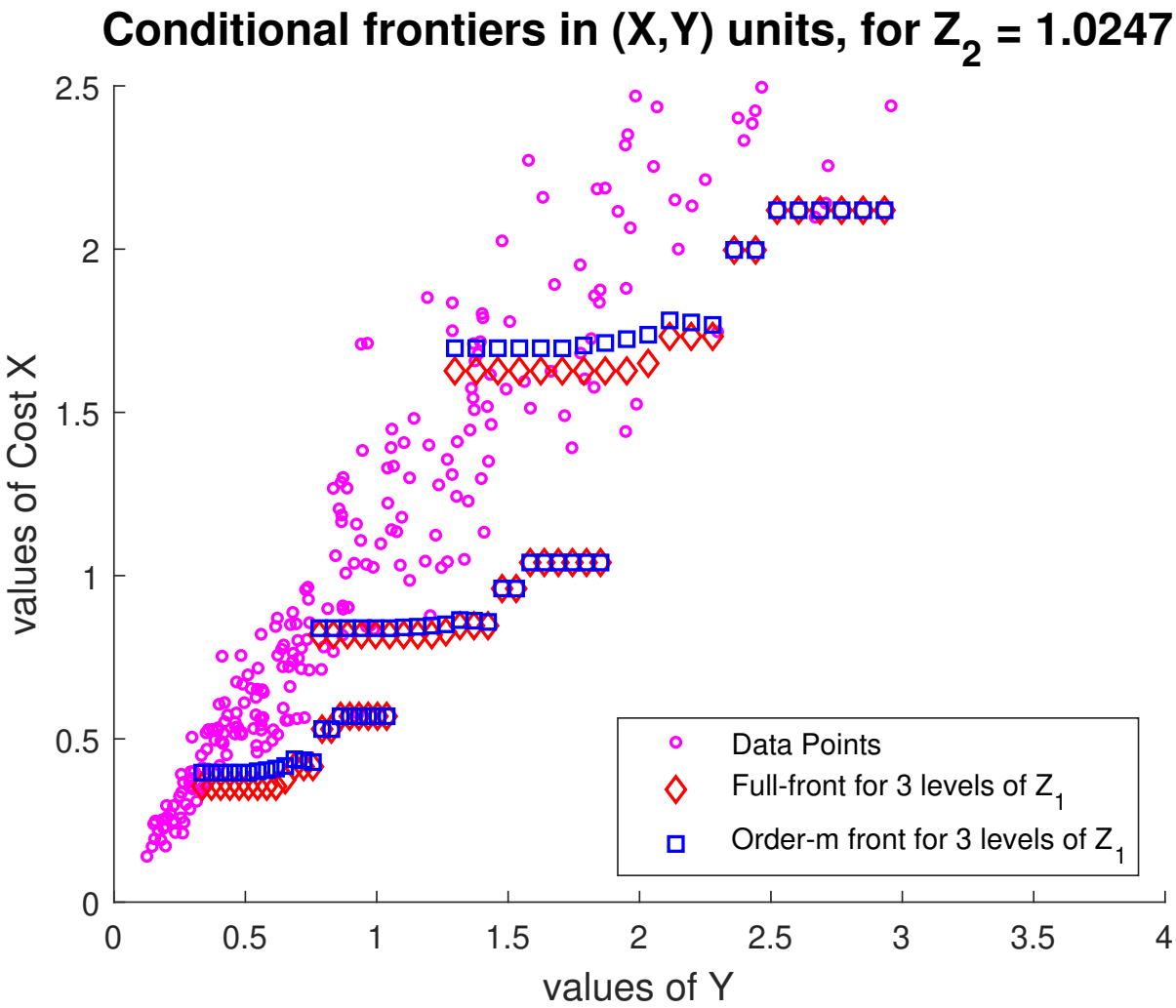

Figure: Bank example: data points in original units and frontier estimates when fixing the level of Z. Here Z_2 is fixed at its median value, and Z_1 is fixed at its 3 quartiles (from the left to the right).

Banking Sector: Results

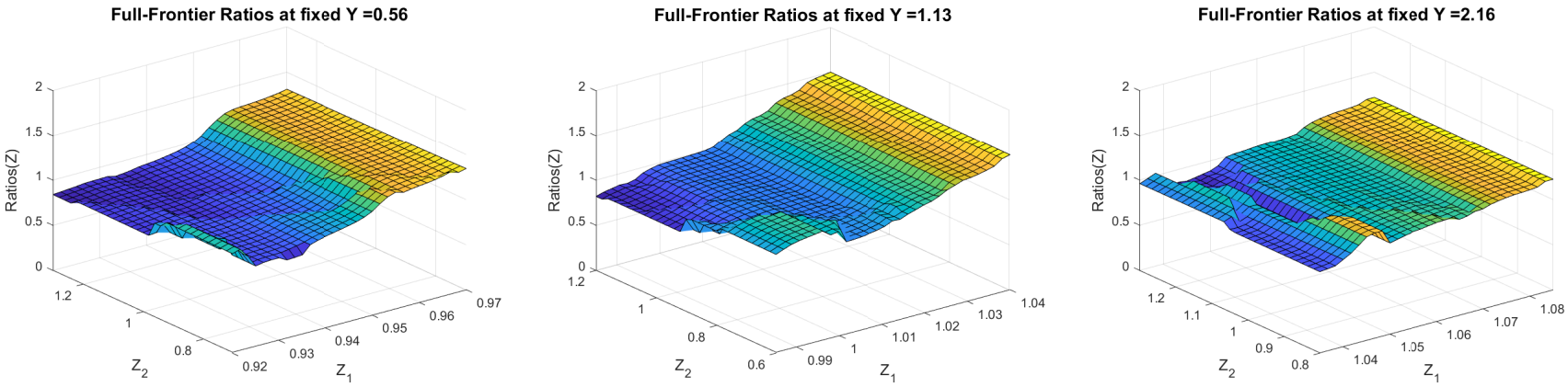

Figure: Bank example: Analysis of the ratios $\widehat{\tau}(y,z)/\widehat{\tau}(y)$ for fixed values of y. From left to right the 3 quartiles of Y.

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

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

Florens, J., Simar, L. and van Keilegom, I.: 2014, Frontier estimation in nonparametric location-scale models, *Journal of Econometrics* **178**, 456–470.