Robust Alternatives to Estimate Benchmark Frontiers

KEI - september 2006, KUL

Léopold Simar
Institut de Statistique
Université Catholique de Louvain, Belgium

Contents

Dominance and Probabilistic Formulation -1-

- Reformulates the production process and introduces in a **natural** way
  - The **Farell-Debreu** benchmark frontier
  - Its various nonparametric estimators **FDH and DEA**
- Extensions allows for some noise and **Robustness** to outliers and extremes
- The formulation allows to introduces easily **Environmental Factors**

Dominance and Probabilistic Formulation -2-

- The **Production process** generates inputs X and outputs Y such that $(X, Y) \in \Psi \subset \mathbb{R}_+^p \times \mathbb{R}_+^q$
  - The attainable set is $\Psi = \{(x, y) \in \mathbb{R}_+^p \times \mathbb{R}_+^q \mid x \text{ can produce } y\}$
  - The DGP (data generating process) according a probabilty model completely characterized by the knowledge of

$$H_{XY}(x, y) = \text{Prob}(X \leq x, Y \geq y),$$

    the probability for a unit operating at the level (x, y) to be **dominated**.
  - **The support of $H_{XY}(\cdot, \cdot)$ is Ψ.**
  - Decomposition

$$\begin{aligned} H_{XY}(x, y) &= \text{Prob}(X \leq x \,|\, Y \geq y)\, \text{Prob}(Y \geq y) = F_{X|Y}(x|y)\, S_Y(y) \\ &= \text{Prob}(Y \geq y \,|\, X \leq x)\, \text{Prob}(X \leq x) = S_{Y|X}(y|x)\, F_X(x), \end{aligned}$$

- **All the relevant information is there!**

Dominance and Probabilistic Formulation -4-

Farell-Debreu Efficiency

- input orientation

$$\theta(x,y) = \inf\{\theta \,|\, F_{X|Y}(\theta x|y) > 0\} = \inf\{\theta \,|\, H_{XY}(\theta x, y) > 0\}.$$

- output orientation

$$\lambda(x,y) = \sup\{\lambda \,|\, S_{Y|X}(\lambda y|x) > 0\} = \sup\{\lambda \,|\, H_{XY}(x, \lambda y) > 0\}.$$

If Ψ is Free-disposal nothing new!

$$\theta(x,y) = \inf\{\theta \,|\, (\theta x, y) \in \Psi\}$$
$$\lambda(x,y) = \sup\{\lambda \,|\, (x, \lambda y) \in \Psi\}$$

Dominance and Probabilistic Formulation -3-

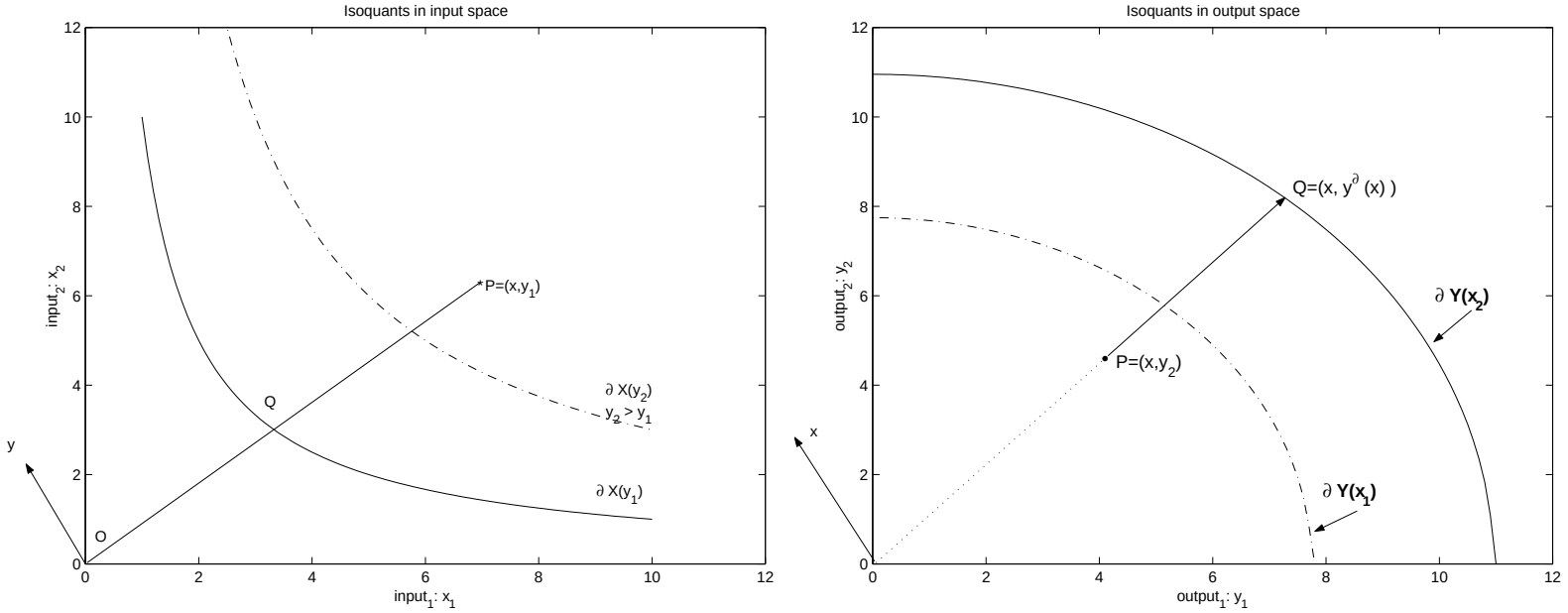

Figure 1: *Isoquants and input efficiency measure: (left)* $\theta_P = |OQ|/|OP| \leq 1$ *and (right)* $\lambda_P = |OQ|/|OP| \geq 1$.

Dominance and Probabilistic Formulation -5-

Ψ is unknown: has to be estimated from $\mathcal{X}_n = \{(x_i, y_i) | i = 1, \ldots, n\}$.

- **Nonparametric estimators**: **plug-in the empirical version** of $H_{XY}(x, y)$

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

So that:

$$\widehat{F}_{X|Y,n}(x|y) = \frac{\widehat{H}_{XY,n}(x, y)}{\widehat{H}_{XY,n}(\infty, y)} \quad \text{and} \quad \widehat{S}_{Y|X,n}(y|x) = \frac{\widehat{H}_{XY,n}(x, y)}{\widehat{H}_{XY,n}(x, 0)}$$

- **Efficiency estimators**:

$$\widehat{\theta}(x, y) = \inf\{\theta \,|\, \widehat{F}_{X|Y,n}(\theta x|y) > 0\} \quad \text{and} \quad \widehat{\lambda}(x, y) = \sup\{\lambda \,|\, \widehat{S}_{Y|X,n}(\lambda y|x) > 0\}$$

These are the FDH estimators (DEA by convexifying the FDH).

Dominance and Probabilistic Formulation -6-

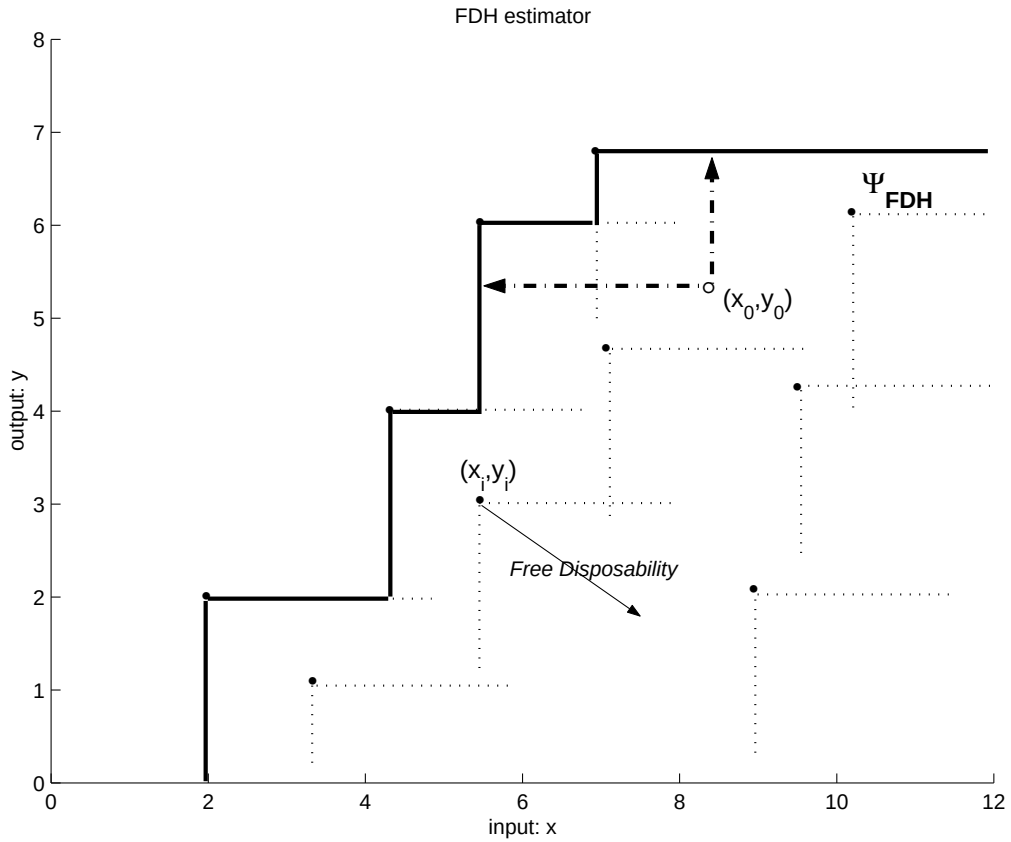

Figure 2: *FDH estimator* $\widehat{\Psi}_{FDH}$ *of the production set* Ψ*: the* • *are the observations.*

Dominance and Probabilistic Formulation -7-

Properties

- FDH are consistent estimators (if Ψ is convex, DEA is also consistent)
- Asymptotic theory is available
- Bootstrap has to be used in practice
- Drawbacks:
  - **sensitivity to extreme and outliers**
  - **curse of dimensionality**: not $n^{-1/2}$ but rather, *e.g.* for FDH, $n^{-1/(p+q)}$

Robust Benchmark Frontier -1-

Basics: Presentation for output orientation and one output $y \in \mathbb{R}_+$.

- **The production set** is the set:

$$\Psi = \{(x, y) \in \mathbb{R}_+^p \times \mathbb{R}_+ \mid x \text{ can produce } y\}.$$

- **The production process** is defined by the joint cdf of (X, Y) on $\mathbb{R}_+^p \times \mathbb{R}_+$:

$$\begin{aligned} F(x, y) = \text{Prob}(X \leq x, Y \leq y) &= \text{Prob}(Y \leq y | X \leq x)\text{Prob}(X \leq x) \\ &= F_{Y|X}(y|x)F_X(x), \end{aligned}$$

  where here $F_{Y|X}(y|x) = 1 - S_{Y|X}(y|x)$ is a nonstandard conditional cdf (conditionned on $X \leq x$).

- If Ψ is free disposal, the **Farrell-Debreu benchmark frontier** function is:

$$\varphi(x) = \{y | (x, \lambda y) \notin \Psi, \forall \lambda > 1\} \equiv \inf\{y | F_{Y|X}(y|x) = 1\}.$$

Robust Benchmark Frontier -2-

Partial order frontiers: Economic interpretation

a new benchmark frontier less extreme than the "full frontier".

- **Order-m**
  - a unit (x, y) is benchmarked against the average maximal output reached by m peers randomly drawn from the population of units using less input than x.
  - As $m \to \infty$, order-m frontier converges to the **full-frontier**.
- **Order-α**: quantile-type
  - a unit (x, y) is benchmarked against the output level not exceeded by $100(1-\alpha)\%$ of firms in the population of units using less input than x.
  - As $\alpha \to 1$, order-α frontier converges to the **full-frontier**.

Robust Benchmark Frontier -3-

Partial order frontiers: Mathematical definition

- In place of looking for **full frontier** $\varphi(x) = \inf\{y|F_{Y|X}(y|x) = 1\}$ define a **less extreme** concept:
  - **Order-m frontier**

$$\begin{aligned}\varphi_m(x) &= E\left[\max(Y^1,\ldots,Y^m)|X \leq x\right] \\ &= \int_0^\infty (1 - [F_{Y|X}(y|x)]^m)\, dy\end{aligned}$$

  - **Order-α quantile frontier**

$$\begin{aligned}\varphi_\alpha(x) &= F_{Y|X}^{-1}(\alpha|x) \\ &= \inf\{y \in \mathbb{R}_+|F_{Y|X}(y|x) \geq \alpha\}\end{aligned}$$

- **Properties**

$$\text{as } m \to \infty,\ \varphi_m(x) \to \varphi(x) \quad \text{and as} \quad \alpha \to 1,\ \varphi_\alpha(x) \to \varphi(x)$$

Robust Benchmark Frontier -4-

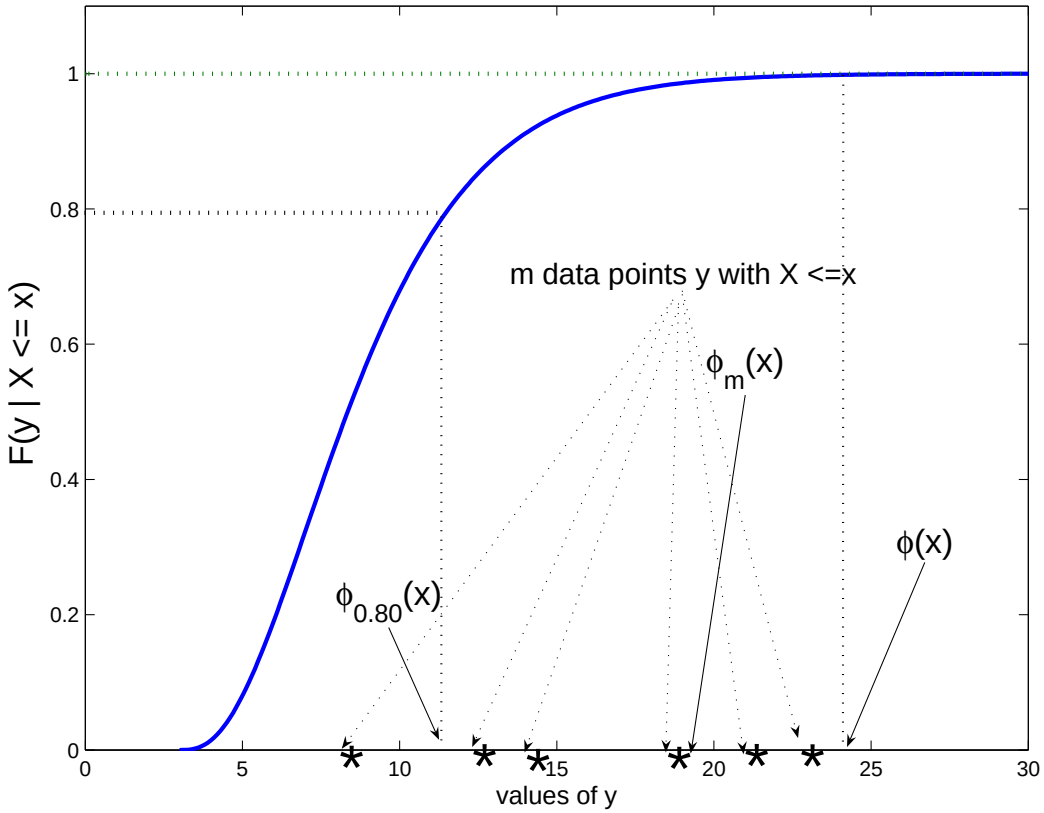

Figure 3: *Illustration of full and partial frontiers:* $m = 6$ *and* $\alpha = 0.80$

Robust Benchmark Frontier -5-

Nonparametric estimators of partial order frontier

- **Plug-in principle**

$$\hat{\varphi}_{m,n}(x) = \int_0^\infty (1 - [\widehat{F}_{n,Y|X}(y|x)]^m)\, dy$$

$$\hat{\varphi}_{\alpha,n}(x) = \inf\{y \in \mathbb{R}_+ | \widehat{F}_{n,Y|X}(y|x) \geq \alpha\}$$

- **Properties**
  - **$\sqrt{n}$-consistency and asymptotic normality:**

$$\sqrt{n}(\hat{\varphi}_{m,n}(x) - \varphi_m(x)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_m^2(x)) \text{ and } \sqrt{n}(\hat{\varphi}_{\alpha,n}(x) - \varphi_\alpha(x)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_\alpha^2(x))$$

  - **Convergence to FDH estimator:**

$$\text{as } m \to \infty,\ \hat{\varphi}_{m,n}(x) \to \hat{\varphi}_{FDH,n}(x) \quad \text{and as} \quad \alpha \to 1,\ \hat{\varphi}_{\alpha,n}(x) \to \hat{\varphi}_{FDH,n}(x)$$

- **Detection of Outliers** (Simar, JPA, 2003)

Robust Benchmark Frontier -6-

Robust estimator of "full frontier" $\varphi(x)$

- When $m \to \infty$ or $\alpha \to 1$, the partial frontiers and their nonparametric estimator converge to full frontier and to the FDH frontier respectively.

  Theorem 1. *If* $m = m(n)$ *is such that* $m(n) = O(n \log(n))$ *when* $n \to \infty$ *and if* $\alpha = \alpha(n)$ *is such that* $n^{(p+2)/(p+1)}(1 - \alpha(n)) \to 0$ *as* $n \to \infty$, *then*

$$n^{1/(p+1)}(\hat{\varphi}_{m(n),n}(x) - \varphi(x)) \xrightarrow{\mathcal{L}} Weibull(\cdot)$$
$$n^{1/(p+1)}(\hat{\varphi}_{\alpha(n),n}(x) - \varphi(x)) \xrightarrow{\mathcal{L}} Weibull(\cdot)$$

  (see CFS and ADT for details)

- Same asymptotic properties that the FDH frontier, but, **for finite** n, $\hat{\varphi}_{m(n),n}(x)$ and $\hat{\varphi}_{\alpha(n),n}(x)$ provide estimators of $\varphi(x)$ that **will not envelop all the data points** and so, are **more robust to extreme and outliers**.

- In practice, m and α are chosen as tuning parameters that tune the **percentage of points left out** the obtained partial frontier estimate.

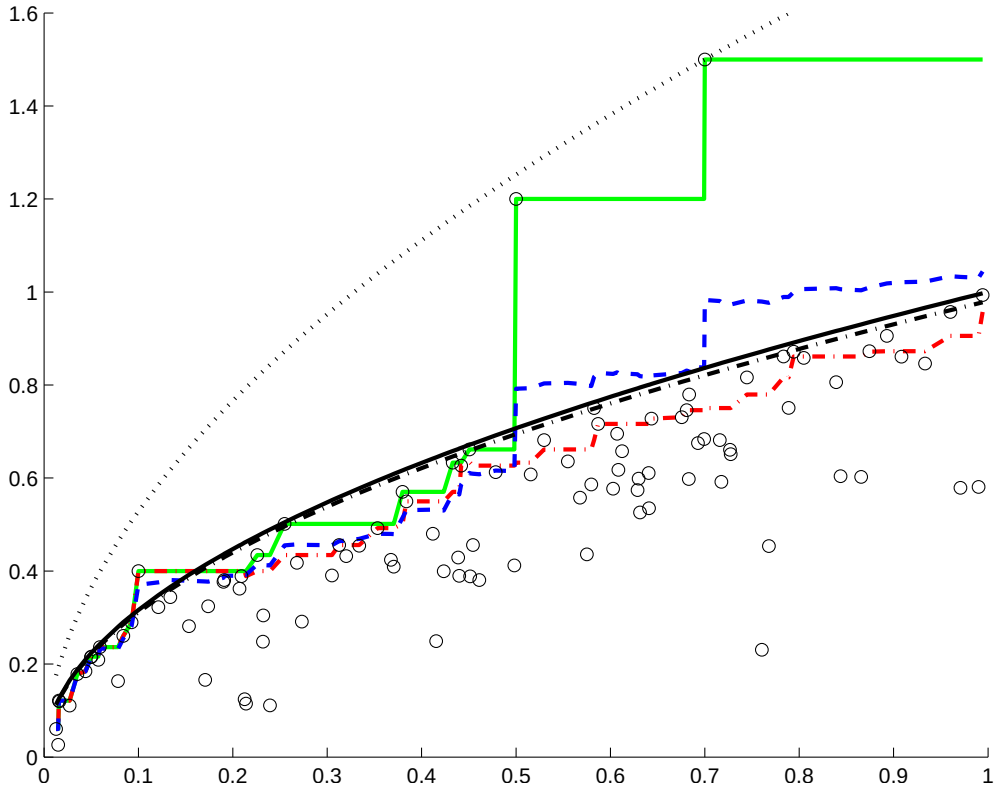

Figure 4: *Example 1. In solid black line, the true frontier* $y = x^{0.5}$. *In cyan solid, the FDH frontier estimate, in blue dashed the estimated order-m frontier and in dash-dot red the estimate of the order-*α *frontier. In black dotted, the shifted OLS estimate and in dash-dot black, the parametric stochastic fit,* $m = 20$ *and* $\alpha = 0.95$.

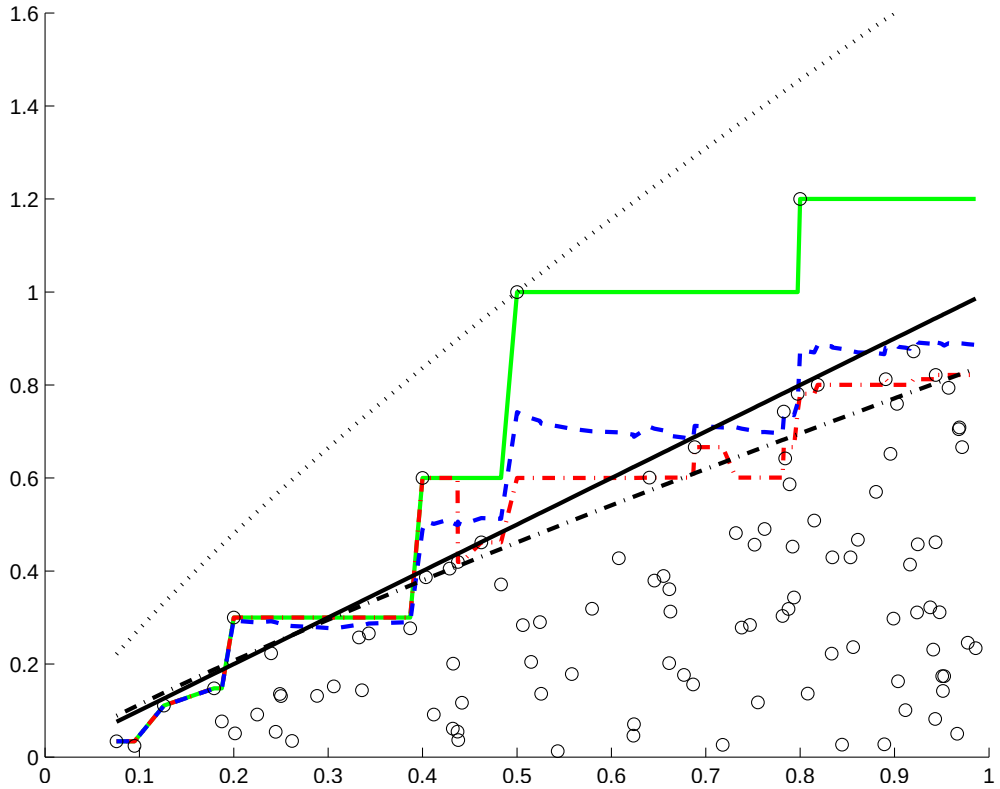

Figure 5: *Example 2. In solid black line, the true frontier* $y = x$. *In cyan solid, the FDH frontier estimate, in blue dashed the estimated order-m frontier and in dash-dot red the estimate of the order-*α *frontier. In black dotted, the shifted OLS estimate and in dash-dot black, the parametric stochastic fit,* $m = 20$ *and* $\alpha = 0.95$.

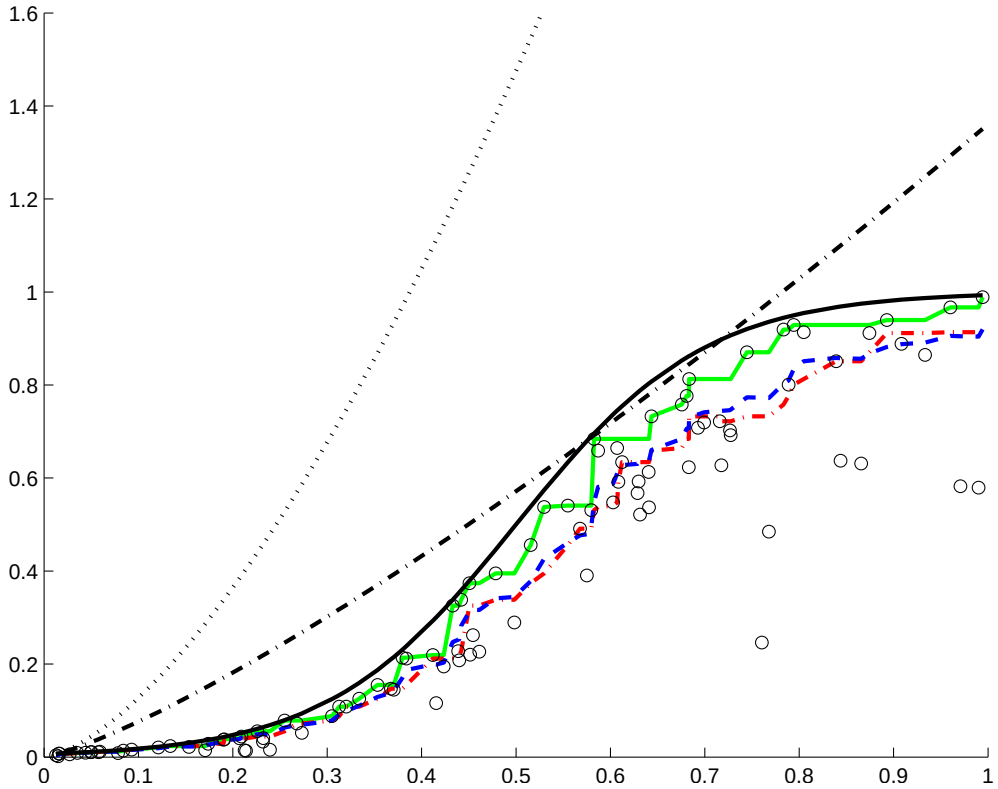

Figure 6: *Example 3. In solid black line, the true logit frontier. In cyan solid, the FDH frontier estimate, in blue dashed the estimated order-m frontier and in dash-dot red the estimate of the order-α frontier. In black dotted, the shifted OLS estimate and in dash-dot black, the parametric stochastic fit,* $m = 20$ *and* $\alpha = 0.95$.

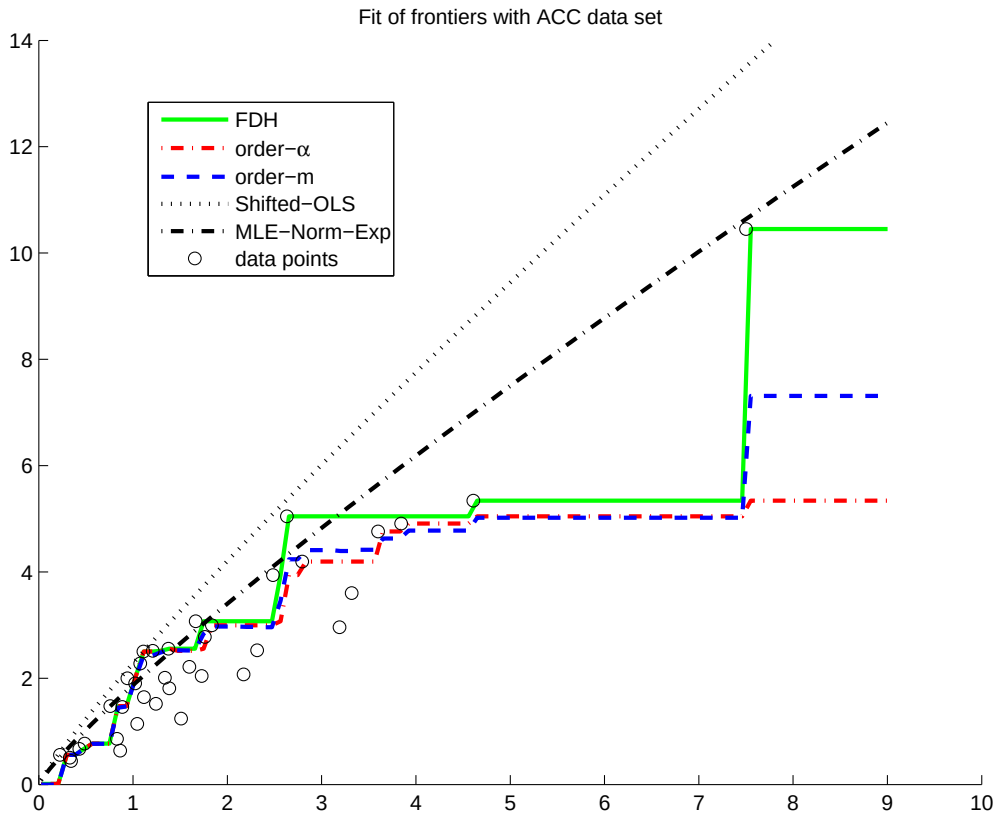

Figure 7: *ACC data: Illustration of full and partial frontiers:* $m = 20$ *and* $\alpha = 0.95$.

Robust Benchmark Frontier: Multivariate versions -7-

- **Farell-Debreu Efficiency** (input orientation)

$$\theta(x,y) = \inf\{\theta \mid F_{X|Y}(\theta x|y) > 0\}$$

If X univariate, $\phi(y) = \theta(x,y)\,x$ is an **minimal** input frontier (benchmark)

- **Order-α quantile frontier**: an other benchmark

$$\theta_\alpha(x,y) = \inf\{\theta \mid F_{X|Y}(\theta x|y) > 1-\alpha\}.$$

If X univariate, $\phi_\alpha(y) = \theta_\alpha(x,y)\,x$ is an **order-α** input frontier (new benchmark)

- **Order-m frontier**: still an other benchmark

$$\theta_m(x,y) \;=\; \int_0^\infty (1 - F_{X|Y}(ux \mid y))^m du$$

If X univariate, $\phi_m(y) = \theta_m(x,y)\,x = E(\min(X_1,\ldots,X_m|Y \geq y))$ is an **order-m** input frontier (new benchmark)

Environmental Factors Z

- **Very easy and natural**
  - No **separability** conditions (Simar and Wilson, 2006 JE: 2-stage story)
  - No **prior** information of the role of Z (favorable or not to the process)
- Replace $H_{XY}(x,y) = \text{Prob}(X \leq x, Y \geq y)$ by
  $H_{XY|Z}(x,y|Z=z) = \text{Prob}(X \leq x, Y \geq y|Z=z)$
- Nonparametric estimator: kernel smoothing on Z

$$\widehat{H}_{XY,n|Z}(x,y|Z=z) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y) K((Z_i - z)/h)}{\sum_{i=1}^{n} K((Z_i - z)/h)}$$

- **All the relevant information is there** and **the theory is done**!
- **Effect of** Z (favorable, neutral or detrimental) by analyzing $\widehat{\theta}(x,y|z)/\widehat{\theta}(x,y)$ as a function of z.

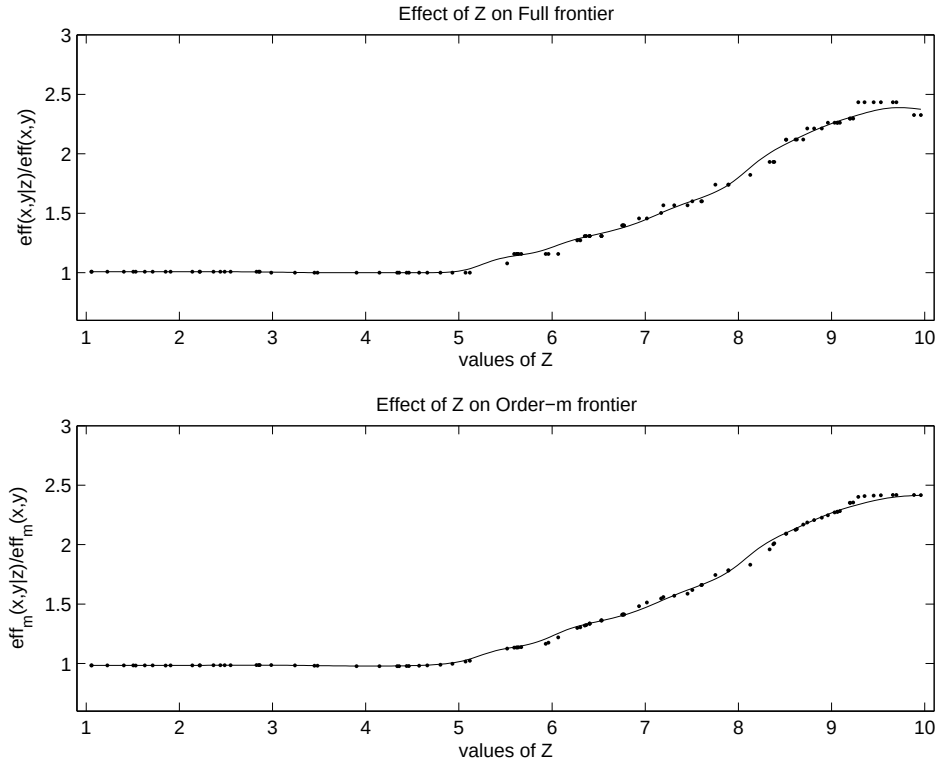

Figure 8: *"Unfavorable" effect of* Z *on production efficiency, only after* $Z > 5$

Conclusions

- **Nonparametric frontiers are very flexible**
- **Statistical inference is available**
- **Robust versions are very useful and easy to compute**
- **Environmental factors are easy to introduce**

Main References

- Daraio, C. and L. Simar (2006), *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications*, forthcoming Springer, New-York, September 2006.
- Cazals, C. Florens, J.P. and L. Simar (2002), Nonparametric Frontier Estimation: a Robust Approach , in *Journal of Econometrics*, 106, 1–25.
- Daouia, A. and L. Simar (2004), Nonparametric efficiency analysis: a multivariate conditional quantile approach, Discussion paper 0419, Institut de Statistique, UCL, forthcoming *Journal of Econometrics*.
- 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.
- Daraio, C. and L. Simar (2006), Conditional nonparametric frontier models for convex and non convex technologies: a unifying approach, Discussion paper 0502, Institut de Statistique, UCL, forthcoming *Journal of Productivity Analysis*.