Nonparametric Frontier Estimation: recent developments and new challenges

Toulouse - October 2011

Léopold Simar
Institut de Statistique
Université Catholique de Louvain, Belgium

Contents

- **Frontier Models and Efficiency Measures**
  - Economy of Production and Farell-Debreu efficiency scores
- **Statistical Paradigm**
  - Different models and Different approaches
- **Nonparametric approaches**
  - FDH and DEA estimators and Statistical inference
- **Challenges**: **Drawbacks of FDH/DEA and Solutions**
  - Robustness to outliers: Partial-order frontier (order-m and order-α quantile)
  - Economic interpretation of the frontier: Parametric approximations
  - Hetrogeneity: introducing Environmental Factors
  - Introducing noise: Stochastic Nonparametric Frontiers

I. Frontier Models and Efficiency Measures

The Frontier Model -1-

- **Economic Theory** Koopmans (1951), Debreu (1951): "Activity Analysis"
  - $x \in \mathbb{R}_+^p$ vector of **inputs**
  - $y \in \mathbb{R}_+^q$ vector of **outputs**
  - **Production set** Ψ of physically attainable points (x, y):

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

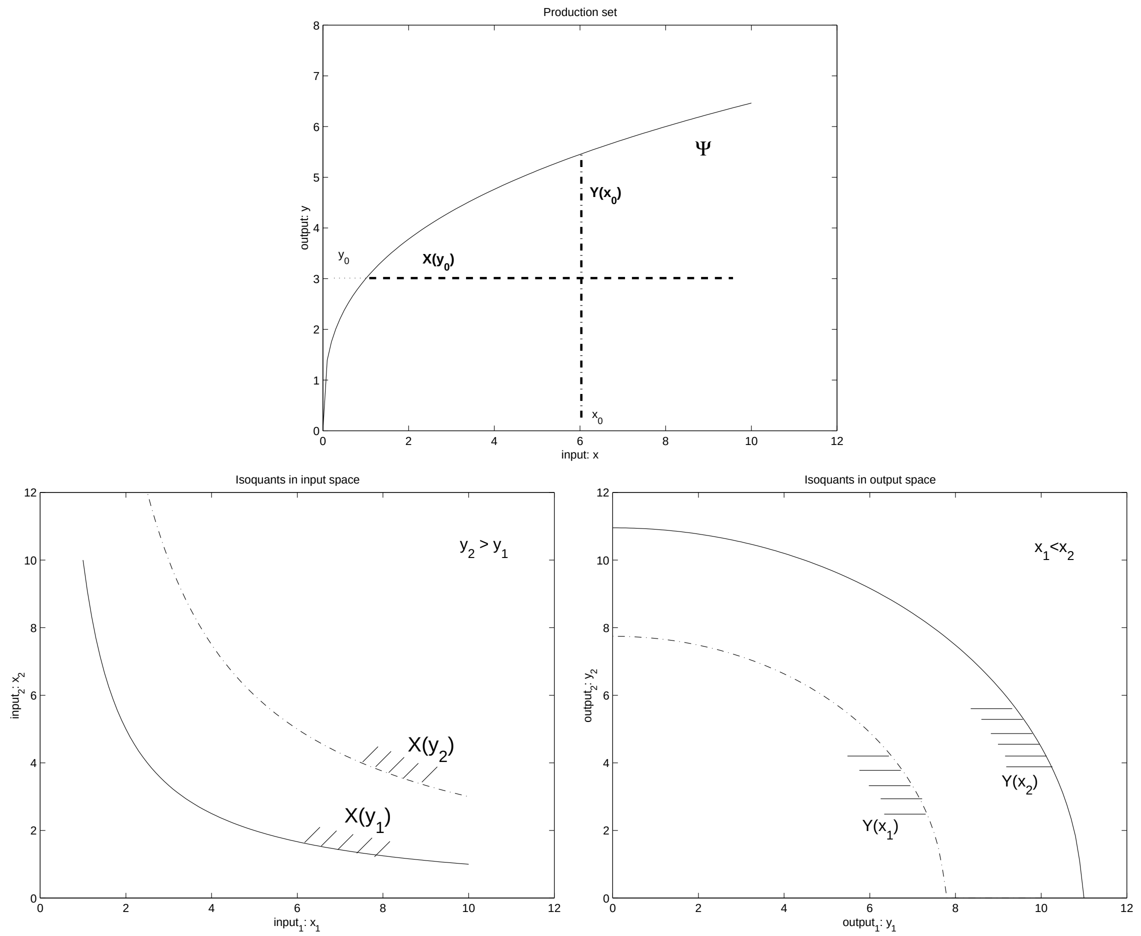
- **The input (output) correspondence sets**
  - Ψ can be described by its sections:

$$\forall\, \boldsymbol{y} \in \boldsymbol{\Psi},\quad \boldsymbol{X}(\boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{R}_+^{\boldsymbol{p}} \mid (\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{\Psi}\}$$

$$\forall\, \boldsymbol{x} \in \boldsymbol{\Psi},\quad \boldsymbol{Y}(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}_+^{\boldsymbol{q}} \mid (\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{\Psi}\}.$$

  - We have

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{\Psi}\,,\ \boldsymbol{x} \in \boldsymbol{X}(\boldsymbol{y}) \Leftrightarrow \boldsymbol{y} \in \boldsymbol{Y}(\boldsymbol{x}).$$



- **Top panel**: **Production set Ψ** for $p = q = 1$.
- **Bottom Panels**: **Correspondence sets $X(y)$ and $Y(x)$** for $p = 2$ and $q = 2$

The Frontier Model -2-

- **Usual Assumptions (a.o.):** (Shephard, 1970)
  - Free Disposability of inputs and outputs

$$\forall (x,y) \in \Psi, \text{ then if } x' \geq x, y' \leq y, \ (x',y') \in \Psi$$

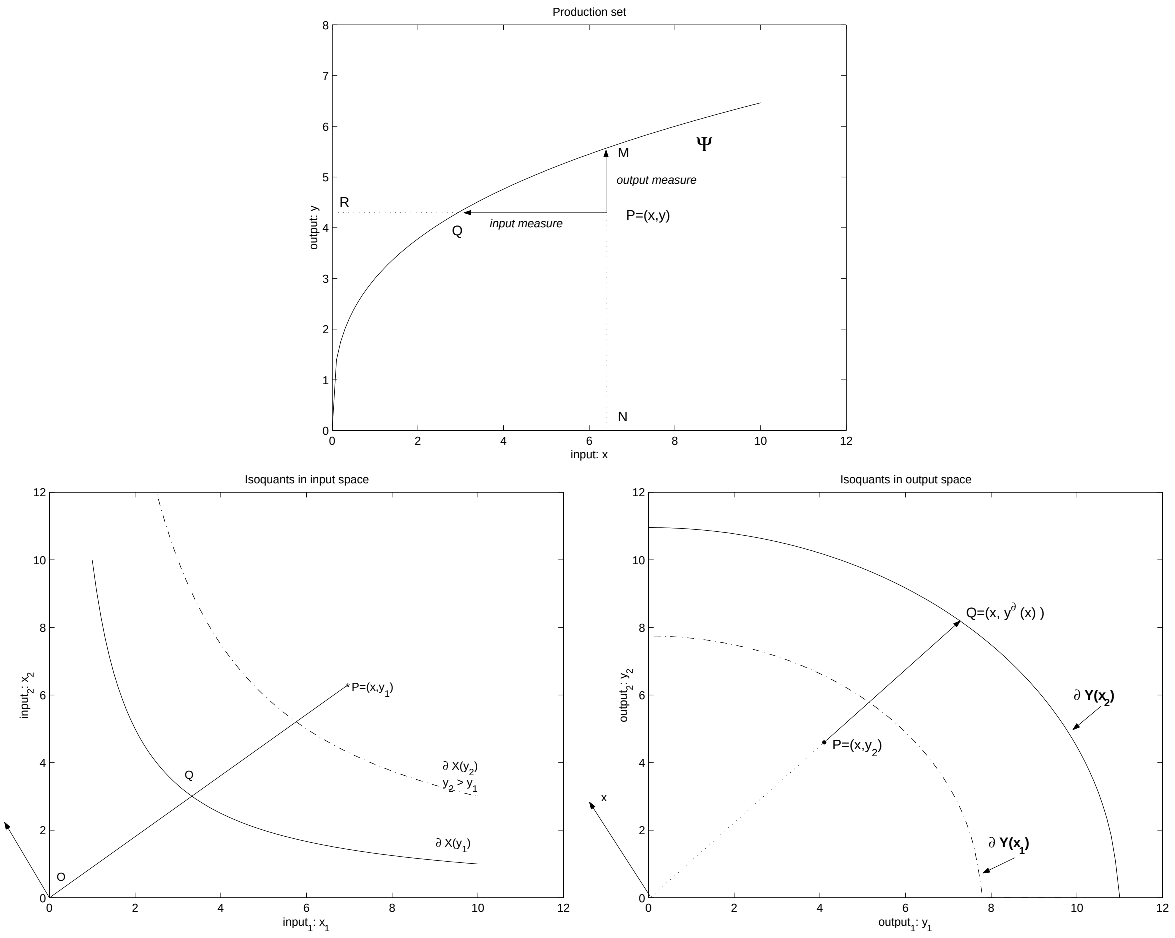
  - Convexity: if $(x_1, y_1), (x_2, y_2) \in \Psi$, then for all $\alpha \in [0,1]$ we have:

$$(x,y) = \alpha(x_1, y_1) + (1-\alpha)(x_2, y_2) \in \Psi$$

  - No Free Lunches: $(x,y) \notin \Psi$ if $x = 0$ and $y \geq 0, y \neq 0$.

- **Farrell-Debreu Efficiency scores**

  radial measures of distance to the boundary of Ψ
  - **Input oriented**: $\boldsymbol{\theta(x,y) = \inf\{\theta \mid (\theta x, y) \in \Psi\} \leq 1}$
  - **Output oriented**: $\boldsymbol{\lambda(x,y) = \sup\{\lambda \mid (x, \lambda y) \in \Psi\} \geq 1}$



- **Top panel**: $\theta_P = |RQ|/|RP| \leq 1$ and $\lambda_P = |NM|/|NP| \geq 1$.
- **Bottom panels**: $\theta_P = |OQ|/|OP| \leq 1$ and $\lambda_P = |OQ|/|OP| \geq 1$

The Frontier Model -3-

- **Extensions**
  - **Hyperbolic Distances**: adjusts simultaneously input and output levels (Färe et al., 1985, Färe and Grosskopf, 2004).

$$\gamma(x, y|\Psi) = \sup\{\gamma > 0|(\gamma^{-1}x, \gamma y) \in \Psi\}.$$

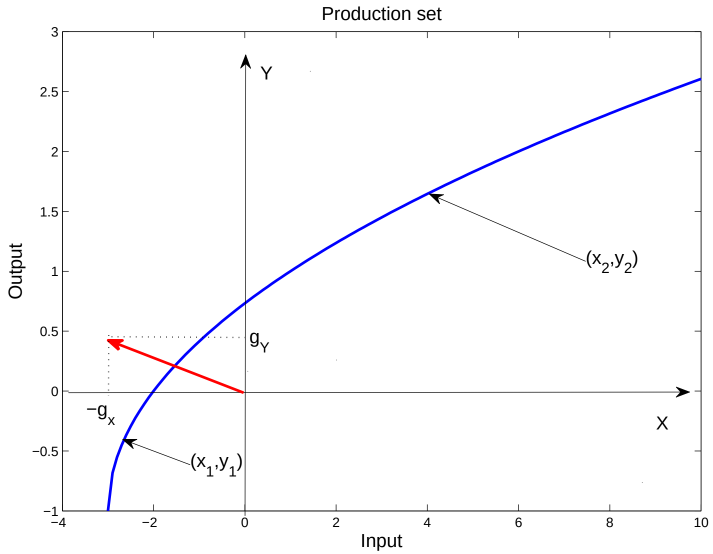
  - **Directional Distances**: Projection of (x, y) onto the technology frontier in a direction $d = (-d_x, d_y)$. (Chambers et al., 1998, Färe and Grosskopf, 2000).

$$\delta(x, y|d_x, d_y, \Psi) = \sup\{\delta|(x - \delta d_x, y + \delta d_y) \in \Psi\}.$$

    * **Additive**: allow negative values of x and/or y.
    * **Special cases**:
      · If $d = (-x, 0)$ with $x > 0$: $\delta(x, y|d_x, d_y, \Psi) = 1 - \theta(x, y|\Psi)^{-1}$
      · If $d = (0, y)$ with $y > 0$: $\delta(x, y|d_x, d_y, \Psi) = \lambda(x, y|\Psi)^{-1} - 1$

The Frontier Model -4-

- Under free disposability, characterization of the technology
  - $\delta(x, y|d_x, d_y, \Psi) \geq 0$ if and only if $(x, y) \in \Psi$
  - $\delta(x, y|d_x, d_y, \Psi) = 0$ if (x, y) is on the frontier.



- **Presentation today and below**: Radial cases, but can be extended (Wilson, 2011, Simar and Vanhems, 2010, Simar, Vanhems and Wilson, 2011)

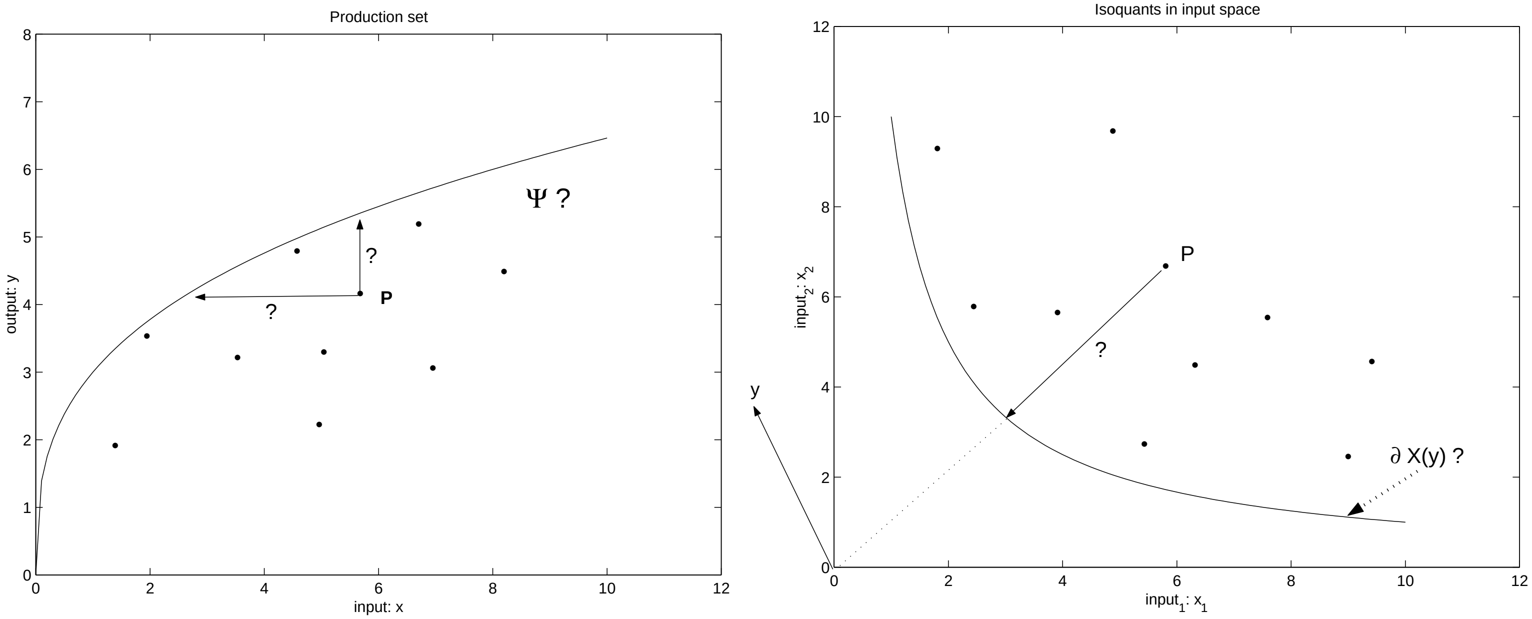
II. The Statistical Paradigm

The Statistical Paradigm

- In practice, Ψ is **unknown**

$$\Rightarrow \theta(x,y) \textbf{ and/or } \lambda(x,y) \textbf{ are also unknown}.$$

- Estimation based on a **sample** $\mathcal{X} = \{(x_i, y_i),\ i = 1, \ldots, n\}$



The Statistical Paradigm -2-

- **Different Approaches**
  - **Deterministic** Frontiers: $\text{Prob}\{(x_i, y_i) \in \Psi\} = 1$, pour tout $i = 1, \ldots, n$.
    - $*$ No noise on the data, no random shocks ...
    - $*$ Distance to frontier is pure inefficiency.
    - $*$ Drawback: **sensitivity** to outliers (superefficient units or errors)
  - **Stochastic** Frontiers
    - $*$ Random noise: some observations may $\notin \Psi$.
    - $*$ Distance to frontier has 2 components (noise and inefficiency)
    - $*$ Drawback: **identification** problems
- **Different Models:** for frontier function and for the law of (X, Y), $F(x, y)$
  - Parametric Models: very restrictive, standard methods (MLE, OLS,...)

    e.g. SFA $\quad Y_i = \beta' X_i + V_i - U_i,$ where $V_i \sim N(0, \sigma_V^2), U_i \sim N^+(0, \sigma_U^2),$ indep.
  - Nonparametric Models: very flexible but more difficult and more challenging.

Choosing a Model: A Summary

Models	**Parametric $\mathcal{P}$**	**Nonparametric $\mathcal{NP}$**
Deterministic $\mathcal{D}$	Analytical models for frontier and for $F(x, y)$	No specific model for frontier and for $F(x, y)$
Stochastic $\mathcal{S}$	Analytical models for frontier for $F(x, y)$ including noise	No specific model for frontier and for $F(x, y)$ including noise (Some structure on noise)

Remarks:

– $\mathcal{D} \subseteq \mathcal{S}$ and $\mathcal{P} \subseteq \mathcal{NP}$

– **Horizontal and Vertical** comparisons are legitimate and may be useful.

– **Semiparametric Models**: combine $\mathcal{P}$ and $\mathcal{NP}$ (see below)

Choosing a Model: Inference

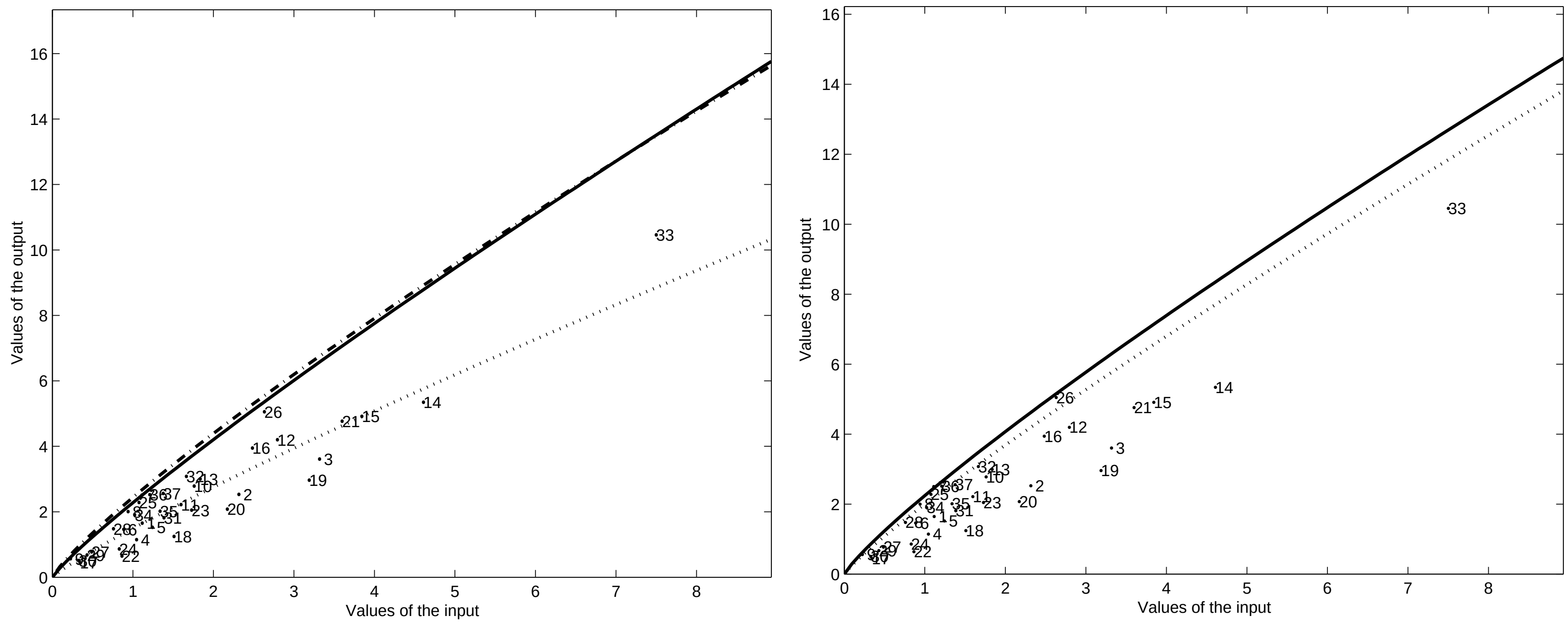
Inference is:	**Parametric $\mathcal{P}$**	**Nonparametric $\mathcal{NP}$**
Deterministic $\mathcal{D}$	**Very Easy** COLS, MOLS, MLE (restrictive) **Two-stages**: $\mathcal{P}$ fit of $\mathcal{NP}$ Bootstrap for efficiency scores	**Easy** FDH: $\widehat{F}_n(x,y) \Rightarrow F(x,y)$ DEA: convexify FDH Bootstrap
Stochastic $\mathcal{S}$	**Easy** MOLS, MLE (restricted models) Identification problems (noise vs ineffcency) Sensitivity: Bagging	**Complicated** Identification problems (deconvolution problem) Localizing $\mathcal{P}$ and SFDH/SDEA Semi-(non)parametric models

Bootstrap is needed everywhere!

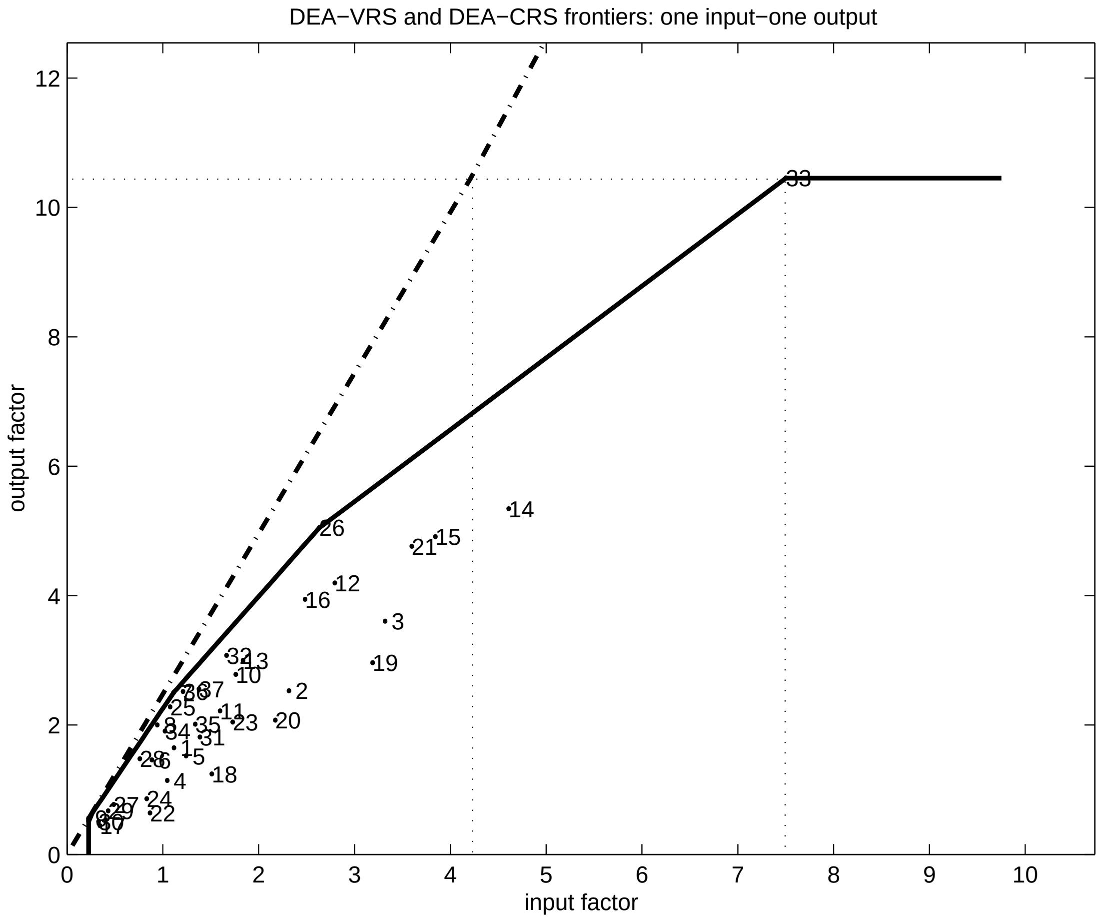
One Example

Efficiency Analysis of Air Controlers (Mouchart and Simar, 2002).
Data are available on the activity of 37 european air controler units in 2000,

- **four outputs**:
  - total flight hours controlled,
  - number of air movements controlled,
  - number of sectors controlled and
  - sum of sector hours worked.
- **two inputs**:
  - the number of air controllers in EFT and
  - the total number of hours worked by air controlers.
- For the example: **aggregated** in one output and one input



The frontier estimates for the Air controlers data: left various deterministic estimates (Solid: Shifted OLS, dash-dotted: MLE-Gamma) , right stochastic (N+HN) (solid for MLE).



DEA estimators VRS (solid) and CRS (dash-dot) for Air Controlers data.

The Statistical Paradigm -3-

- **Statistical Inference**
  - Estimation individual inefficiencies ("rankings")
  - Confidence intervals for these measures
  - Specification tests
    - Aggregation of inputs and/or outputs
    - Relevance of the chosen variables
  - Hypothesis testing on the shape of the efficient frontier ("technology")
    - Convexity
    - Returns to scale (increasing/decreasing/constant)
  - Evolution over time
    - Panel data
    - Gain or loss of productivity?
    - Technical progress or gain of efficiency?

The Literature

- **Parametric deterministic or stochastic frontier models**: hundreds of papers in Econometric literature (*Journal of Econometrics*,...)

  Easier but are the parametric assumptions reasonable ones?

- **Nonparametric deterministic frontier models**: thousands of papers in hundreds of different journals (Management sciences, OR, Econometrics)

  Very popular (flexibility) but some drawbacks (see below).

- **Nonparametric stochastic frontier models**: very recent, very few applications (theoretical econometric literature)

  Flexible but so far, hard to use: "work in progress"...

- **Applications:** Banks, Transports (Air, Railways,...), Public Services, Municipalities, Post, School, Education, Research, University, Insurance, Hospitals, Finance, Mutual funds, Industry, Electric plants, Food industry, Agronomy, Macroeconomic, Economy of development, Regional economy,... (*Journal of Productivity Analysis*)

III. Nonparametric Approaches

Nonparametric Estimators: FDH -1-

- **Envelopment Estimators**: estimate $\boldsymbol{\Psi}$ by $\widehat{\boldsymbol{\Psi}}$ which **"envelops"** at best the cloud of n data points $\mathcal{X}$.

- **Free Disposal Hull: FDH** Deprins, Simar, Tulkens (1984)

$$\widehat{\Psi}_{FDH}(\mathcal{X}) = \left\{(x,y) \in \mathbb{R}_+^{p+q} | y \leq y_i,\ x \geq x_i, \quad (x_i, y_i) \in \mathcal{X}\right\}$$

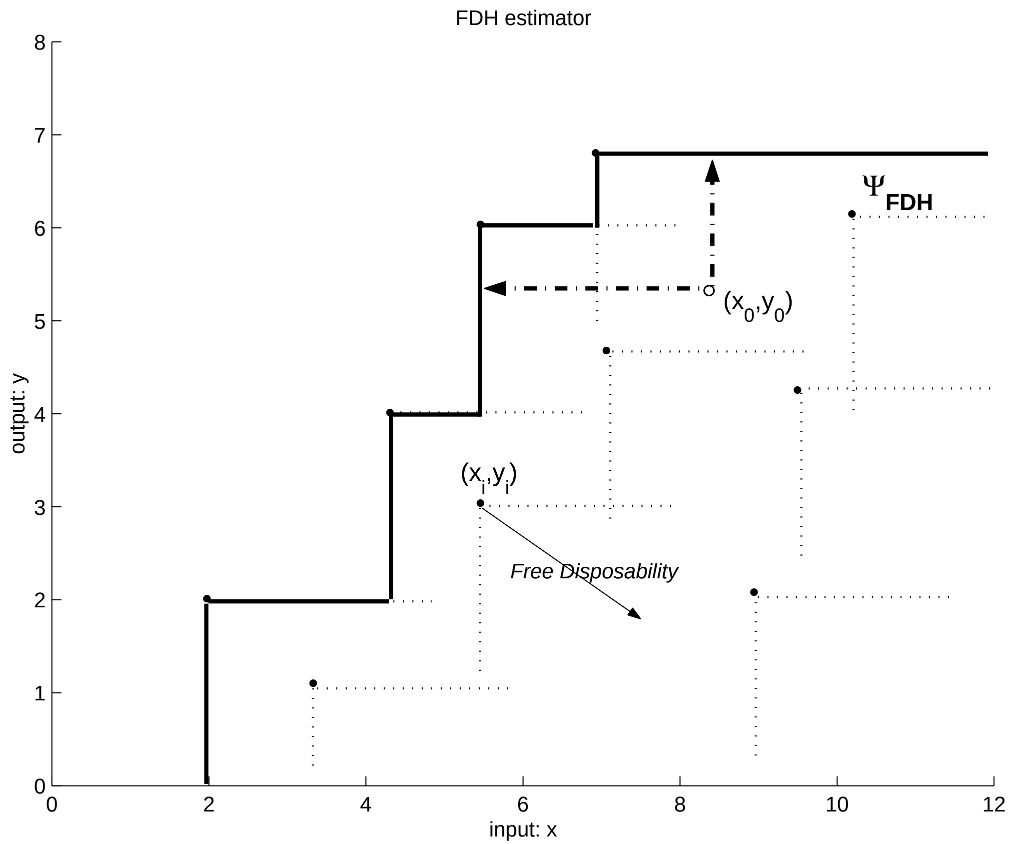
- **FDH efficiency scores**

$$\begin{aligned} \hat{\theta}(x_0, y_0) &= \inf\{\theta \mid (\theta x_0, y_0) \in \widehat{\Psi}_{FDH}(\mathcal{X})\} \\ \hat{\lambda}(x_0, y_0) &= \sup\{\lambda \mid (x_0, \lambda y_0) \in \widehat{\Psi}_{FDH}(\mathcal{X})\}. \end{aligned}$$

- **Practical computations**: fast and easy (sorting algorithms)
  - The set **dominating** points: $D_0 = \{i \mid (x_i, y_i) \in \mathcal{X},\ x_i \leq x_0,\ y_i \geq y_0\}$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \min_{i \in D_0} \max_{j=1,\ldots,p} \left(\frac{\boldsymbol{x}_i^j}{\boldsymbol{x}_0^j}\right); \qquad \hat{\boldsymbol{\lambda}}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \max_{i \in D_0} \min_{j=1,\ldots,q} \left(\frac{\boldsymbol{y}_i^j}{\boldsymbol{y}_0^j}\right)$$

Nonparametric Estimators: FDH -2-



FDH estimator $\widehat{\Psi}_{FDH}$ of the production set Ψ: the $\bullet$ are the observations.

Nonparametric Estimators: DEA -1-

- **Data Envelopment Analysis: DEA** If Ψ is convex:
  - Take the **convex hull** of $\widehat{\Psi}_{FDH}$ (Farrell, 1957, Charnes, Cooper and Rhodes, 1978)

$$\widehat{\Psi}_{DEA} = \quad \{(x,y) \in \mathbb{R}^{p+q} | y \leq \sum_{i=1}^{n} \gamma_i y_i ; x \geq \sum_{i=1}^{n} \gamma_i x_i \text{ for } (\gamma_1, \ldots, \gamma_n)$$

$$\text{such that } \sum_{i=1}^{n} \gamma_i = 1 ; \gamma_i \geq 0, i = 1, \ldots, n\}.$$

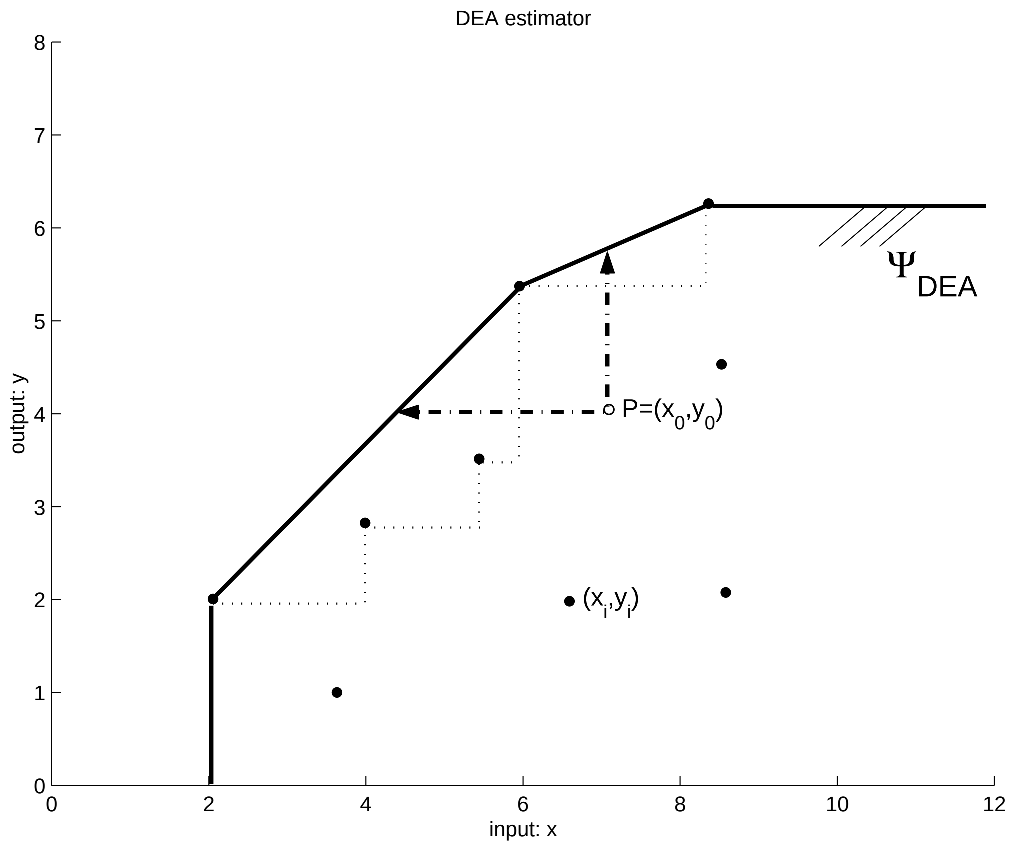
- **Estimation of efficiency score**

$$\begin{aligned} \hat{\theta}(x,y) &= \inf\{\theta \mid (\theta x, y) \in \widehat{\Psi}_{DEA}(\mathcal{X})\} \\ \hat{\lambda}(x,y) &= \sup\{\lambda \mid (x, \lambda y) \in \widehat{\Psi}_{DEA}(\mathcal{X})\} \end{aligned}$$

- **Computation** through **linear programs**.

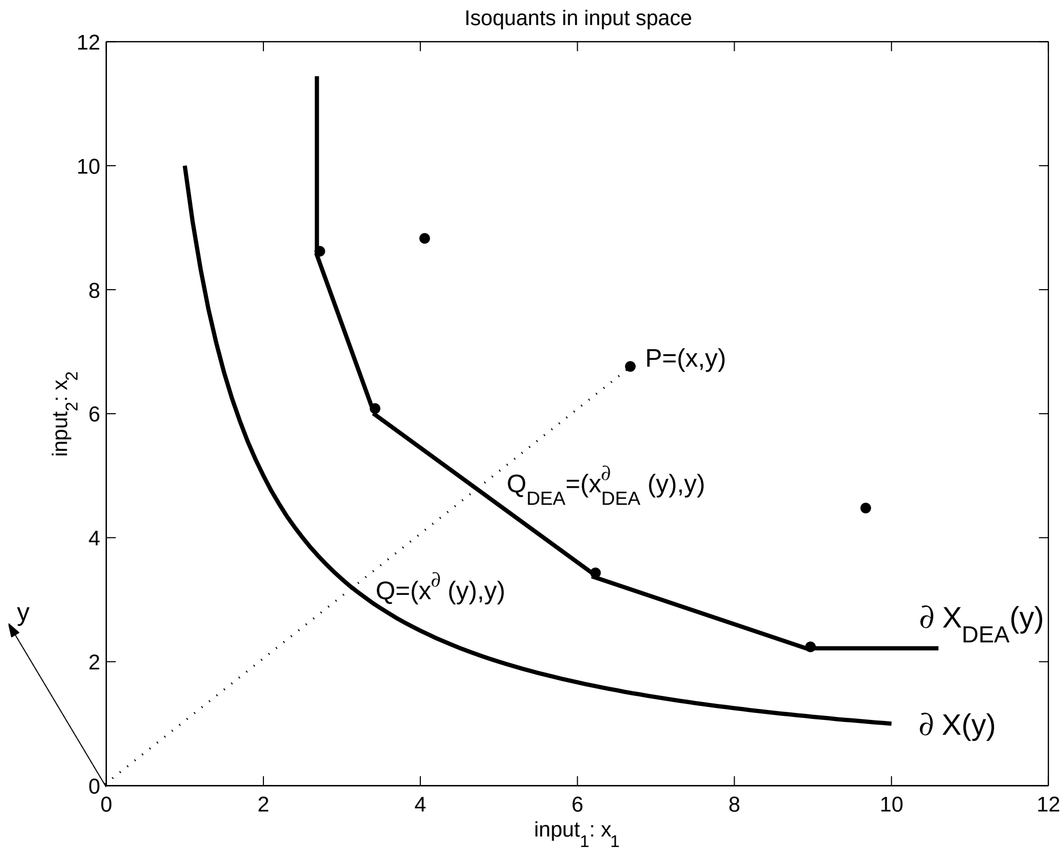
  Available free software: **FEAR** (Wilson, 2008)

Nonparametric Estimators: DEA -2-



DEA estimator $\widehat{\Psi}_{DEA}$ of the production set Ψ: the $\bullet$ are the observations.

Nonparametric Estimators: DEA -3-



Properties of DEA estimators: Relations between $\hat{\theta}_{DEA}(x, y)$ and $\theta(x, y)$?

Statistical Inference: State of the Art -1-

Properties: recent survey, Simar and Wilson (2008)

- **Consistency and rate of convergence**:

$$\left(\hat{\theta}(x,y) - \theta(x,y)\right) = O_p(n^{-\tau}), \text{ as } n \to \infty?$$

  - **FDH**: Korostelev, Simar and Tsybakov (1995a) and Park, Simar and Weiner (2000). Rate is $n^{-1/(p+q)}$.
    Recent Extensions: Daouia, Florens and Simar (2010)
  - **DEA**: Korostelev, Simar and Tsybakov (1995b) and Kneip, Park and Simar (1998). Rate is $n^{-2/(p+q+1)}$. Park, Jeong and Simar (2010) (CRS case), rate is $n^{-2/(p+q)}$.

- **Nice!** but not very useful for the practitionners.
- **Curse of dimensionality**: bad rates if $p + q \uparrow$.

Statistical Inference: State of the Art -2-

Is Inference possible ?

- **Asymptotic sampling distribution**:

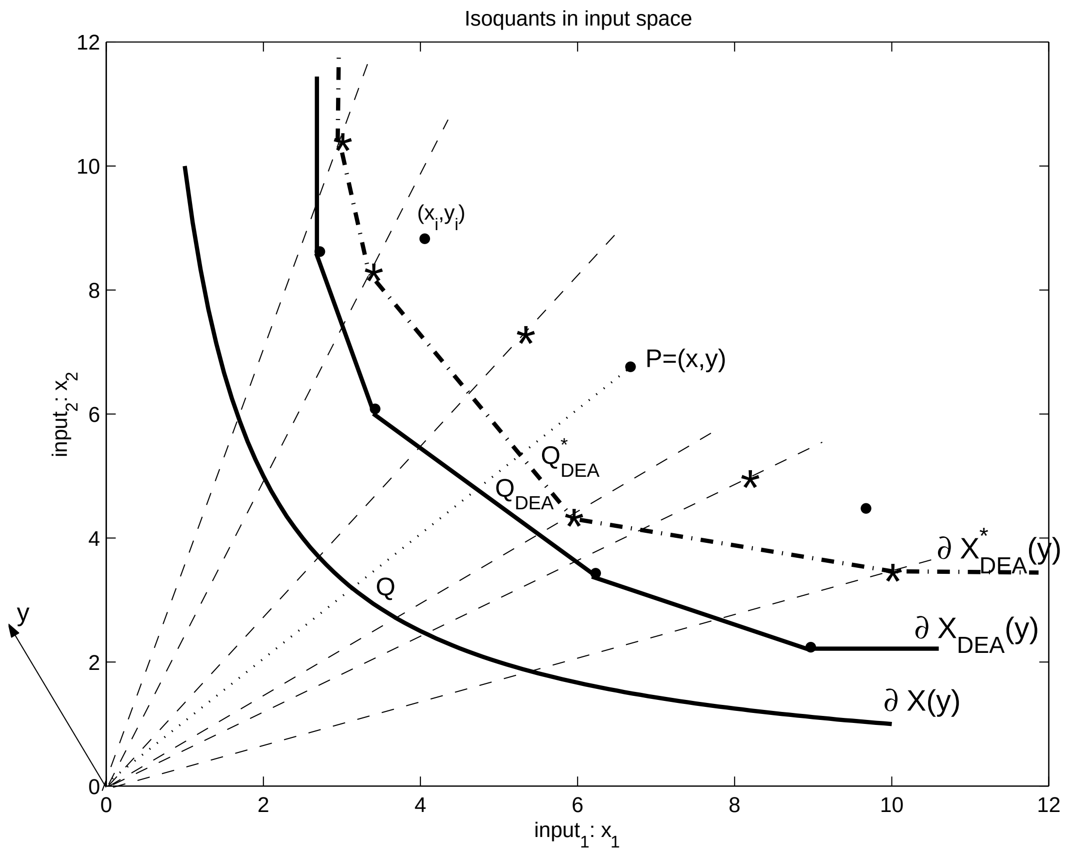
$$n^{\tau}\left(\hat{\theta}(x,y)-\theta(x,y)\right)\sim Q(\eta),\ \text{as } n\to\infty?$$

  - **FDH**: Park, Simar and Weiner (2000), Badin, Simar (2009), Daouia, Florens and Simar (2010); $Q(\eta)$ is a **Weibull distribution** with unknown parameters to be estimated: not easy to handle and need large sample sizes if $p+q$ increases.
  - **DEA**: Gijbels, Mammen, Park and Simar (1999), Kneip, Simar and Wilson (2008), Park, Jeong, Simar (2010); $Q(\eta)$ is a **Regular distribution** depending on unknown parameters but no closed forms available (untractable for practical purposes) when p or $q>1$.
- **No hope ?** Yes: the bootstrap.

The Bootstrap -1-

Basic Idea

- **The "Real World"**: The Data Generating Process $\mathcal{P}$
  (x_i, y_i) in $\mathcal{X}$ are realizations of iid random variables (X, Y) with probability density function $f(x, y)$ with support Ψ, and $\text{Prob}\big((X, Y) \in \Psi\big) = 1$.
  - $\widehat{\Psi}(\mathcal{X})$ is an estimator of Ψ (FDH or DEA)
  - $\hat{\theta}(x, y) = \inf\{\theta \mid (\theta x, y) \in \widehat{\Psi}(\mathcal{X})\}$ is an estimator of $\theta(x, y)$
- **The "Bootstrap World"**: Consider a DGP $\widehat{\mathcal{P}}$, a consistent estimator of $\mathcal{P}$.
  We can use $\widehat{\Psi}(\mathcal{X})$ (FDH or DEA) and **some appropriate** $\hat{f}(x, y)$ with support $\widehat{\Psi}(\mathcal{X})$, and $\text{Prob}\big((X, Y) \in \widehat{\Psi}(\mathcal{X})\big) = 1$.
- **Bootstrap Analogy**:
  Define a new data set $\mathcal{X}^* = \{(x_i^*, y_i^*),\ i = 1, \ldots, n\}$ drawn from $\widehat{\mathcal{P}}$.
  - $\widehat{\Psi}(\mathcal{X}^*)$ is an estimator of $\widehat{\Psi}(\mathcal{X})$: here, $\widehat{\Psi}(\mathcal{X}^*)$ is the FDH or DEA set computed with $\mathcal{X}^*$ as reference data set.
  - $\hat{\theta}^*(x, y) = \inf\{\theta \mid (\theta x, y) \in \widehat{\Psi}(\mathcal{X}^*)\}$ is an estimator of $\hat{\theta}(x, y)$



The Bootstrap idea:

the • are the original observations (x_i, y_i) generated by the **unknown** $\mathcal{P}$, and the * are the pseudo-observations (x_i^*, y_i^*) generated by the **known** $\widehat{\mathcal{P}}$.

The Bootstrap -2-

- **The Key Relation** : If the Bootstrap is **consistent**, for large n,

$$(\hat{\theta}^*(x,y) - \hat{\theta}(x,y)) \mid \widehat{\mathcal{P}} \quad \approx \quad (\hat{\theta}(x,y) - \theta(x,y)) \mid \mathcal{P}.$$

  - The right part is **unknown** and/or difficult to handle
  - The left part can be approximated by **Monte-Carlo** simulation methods

- **Inference is now available**
  - Bias correction and Standard errors of $\hat{\theta}(x,y)$ are available
  - Confidence intervals for $\theta(x,y)$ can be builded

- **How to generate $\mathcal{X}^*$?** Naive bootstrap **looks easy**: n random drawns of (x_i^*, y_i^*) from $\mathcal{X}$.

- **But naive bootstrap is inconsistent** Simar and Wilson (1998, 1999a, 1999b)
  - The efficient facet, which determines in the original sample $\mathcal{X}$ the value of $\hat{\theta}$, appears **too often**, and with a **fixed** probability, in $\mathcal{X}^*$ and this fixed probability **does not vanish** even when $n \to \infty$.

The Bootstrap -3-

Two Solutions: see Simar and Wilson (1998, 2000, 2011a), Jeong and Simar (2006), Kneip, Simar and Wilson (2008)

- **Subsampling**: draw from $\widehat{\mathcal{P}}$ pseudo-samples of size $m = n^{\kappa}$ where $\kappa < 1$.
  - How to chose m in practice: Simar and Wilson (2011a).
- **Smoothing**: Use smoothed density estimate $\hat{f}(x, y)$ and smooth the boundary of $\widehat{\Psi}$ when defining $\widehat{\mathcal{P}}$: not easy to implement due to the double smoothing.
  - Simplification: homogeneous bootstrap, Simar and Wilson (1998), similar to homoskedastic assumption in regression. But restrictive...
  - Consistent efficient algorithm in the heterogeneous case: Kneip, Simar and Wilson (2011).

Testing issues: Returns to scale, Simar and Wilson (2002), Comparison of groups of firms, Simar and Zelenyuk (2006, 2007), Testing significancy of variables and/or aggregation of variables, Simar and Wilson (2001), and work in progress (convexity,...).

Extensions available: **Hyperbolic distances**, Wilson (2011), **Directional distances**, Simar and Vanhems (2010), Simar, Vanhems and Wilson (2011).

An Example: Program Follow Through (PFT)

- Charnes, Cooper, Rhodes (1981): analysis of an experimental education program administered in US schools: data for 49 schools that implemented PFT, and 21 schools that did not, for a total of 70 observations. 5 inputs and 3 outputs
  - x_1: Education level of the mother (percentage of high school graduates among the mothers),
  - x_2: Highest occupation of a family member (according a pre-arranged rating scale),
  - x_3: Parental visit to school index (number of visits to the school)
  - x_4: Parent counseling index (time spent with child on school related topics)
  - x_5: Number of teachers of the school.

  There are three outputs (results to standard tests):
  - y_1: Total Reading Score (MAT: Metropolitan Achievement Test),
  - y_2: Total Mathematics Score (MAT) and
  - y_3: Coopersmith Self-Esteem Inventory (measure of self-esteem).
- We look for **output efficiency** of the Schools $\lambda(x, y)$ using DEA estimators.

Units	$\hat{\lambda}(x, y)$	Units	$\hat{\lambda}(x, y)$
1	1.0323	50	1.0436
2	1.1093	51	1.0871
3	1.0684	52	1.0000
4	1.1074	53	1.1465
5	1.0000	54	1.0000
⋮	⋮	⋮	⋮

- **Questions**:
  - What is the real value of $\lambda(x, y)$ (bias correction, confidence intervals)?
  - Comparaison of the 2 groups of school:
    - Mean of Group A (49 PFT schools): $\bar{\hat{\lambda}}_A = 1.0589$
    - Mean of Group B (21 Non-PFT schools): $\bar{\hat{\lambda}}_B = 1.0384$ (more efficient?)
  - Is it **significant**?

• The Bootstrap

Units	Eff. Scores	Eff. Bias-Corrected	Bias	Std	Lower Bound	Upper Bound
1	1.0323	1.0671	-0.0348	0.0246	1.0343	1.1268
2	1.1093	1.1387	-0.0294	0.0162	1.1111	1.1702
3	1.0684	1.0979	-0.0295	0.0186	1.0703	1.1396
4	1.1074	1.1264	-0.0190	0.0098	1.1094	1.1463
5	1.0000	1.0530	-0.0530	0.0444	1.0020	1.1651
50	1.0436	1.0725	-0.0289	0.0221	1.0450	1.1239
51	1.0871	1.1102	-0.0231	0.0125	1.0895	1.1373
52	1.0000	1.0558	-0.0558	0.0435	1.0021	1.1542
53	1.1465	1.1718	-0.0253	0.0121	1.1485	1.1954
54	1.0000	1.0520	-0.0520	0.0418	1.0019	1.1484

- After **bias correction** the mean are:
  - Group A (PFT): 1.0940
  - Group B (Non-PFT): 1.0740
- **Formal Test:** $H_0 : E[\lambda(X,Y)|A] = E[\lambda(X,Y)|B]$ vs $H_0 : E[\lambda(X,Y)|A] > E[\lambda(X,Y)|B]$
  - p-value of $H_0 = 0.5590$: $\Rightarrow$ **We do not reject** H_0.

An Other Example: Role of Innovation on Exports -1-

Schubert and Simar (2011) analyze the relations between exports and innovation in the sector of **"Mechanical Engineering"** in Germany (CIS survey, 2007)

- The economic literature is unclear and divided on the role of innovation
- Empirical studies, so far, used parametric models with restrictive assumptions and analyze mean behavior of firms (regression)
- We want to analyze the role of innovation on efficient production plans, not on averages
  - Data: 215 firms
  - 3 inputs: X_1 Expenses for personnel, X_2 Expenses for equipment and materials, X_3 Expenses for innovation
  - 2 outputs: Y_1 Domestic turnover et Y_2 Exports

An Other Example: Role of Innovation on Exports -2-

- NB: here individual efficiency scores are not of interest
- **Results**: tests using bootstrap on different models
  - The inputs X_1 and X_2 can be aggregated, without changing the structure of the efficient frontier, but not X_3.
  - The outputs Y_1 and Y_2 can be aggregated
- **Empirical Conclusions**:
  - The expenditures in innovation (even measured in monetary terms) are really different than other routine expenses: they influence the shape of the frontier, it is really an input, and not a by-product (as claimed by some economic theory).
  - For efficient firms, there is no clear empirical evidence of a link between the expenses in innovation and the exports (as claimed by some economic theory).

IV. Challenges

Challenges: Drawbacks of DEA/FDH and Solutions

- **Sensitivity to extreme/outliers**: **robust** methods and/or detection of outliers
  - **Order-m frontiers**: Cazals, Florens and Simar (2002), Simar (2003), Daouia, Florens and Simar (2009).
  - **Order-α quantile frontiers**: Aragon, Daouia and Thomas (2005), Daouia and Simar (2005, 2007), Daouia, Florens and Simar (2009, 2010).
- **Lack of Economic interpretation**: **Semiparametric Model**, parametric approximations of nonparametric frontiers, Simar (1992), Florens and Simar (2005), Daouia, Florens and Simar (2008)
- **Heterogeneity:** How to explain inefficiency by environmental/external factors ?
  - **Two-stage methods**, Simar and Wilson (2007, 2011b).
  - **Conditional measures of efficiency**, Cazals, Florens and Simar (2002), Daraio and Simar (2005, 2006, 2007a, 2007,b), Jeong, Park and Simar (2010), Badin, Daraio and Simar (2010, 2011).
- **No noise is allowed**: deterministic frontiers $\text{Prob}\big((X,Y)\in\Psi\big)=1$: **Nonparametric Stochastic Frontiers?**: Simar (2007), Kumbhakar, Park, Simar and Tsionas (2008), Simar and Zelenyuk, (2011), Kneip, Simar and Van Keilegom (2010), flexible semiparametric models.

IV.1 Sensitivity to Outliers

Robust Frontier -1

Probabilitic Formulation of DGP

- **The DGP**: $H(x,y) = \text{Prob}(X \leq x, Y \geq y)$, **$\Psi$ is the support of $H(x,y)$**
- **Farrell-Debreu Efficiency score** (case of input orientation)

$$H(x,y) = \text{Prob}(X \leq x \,|\, Y \geq y)\ \text{Prob}(Y \geq y) = F_{X|Y}(x|y)\, S_Y(y)$$
$$\theta(x_0,y_0) = \inf\{\theta|(\theta x_0, y_0) \in \Psi\} = \inf\{\theta|F_{X|Y}(\theta x_0|y_0) > 0\}$$

- **Nonparametric Estimator**: Plug-in the empirical version of $H(x,y)$

$$\widehat{H}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y), \text{ then } \widehat{F}_{X|Y,n}(x|y) = \frac{\widehat{H}_n(x,y)}{\widehat{H}_n(\infty,y)}$$

- **The FDH estimators**: Cazals, Florens and Simar (2002)
  - $\widehat{\Psi}_{FDH}$ is the support of $\widehat{H}_n(x,y)$
  - Estimation (input) efficiency score: $\widehat{\theta}(x_0,y_0) = \inf\{\theta \,|\, \widehat{F}_{X|Y,n}(\theta x_0|y_0) > 0\}$

Robust Frontier -2-

Partial order frontiers. Economic interpretation (case of univariate output)

Another benchmark frontier less extreme than the "full frontier".

- **Order-m:** Cazals, Florens, Simar (2002)
  - 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:** Aragon, Daouia, Thomas (2005), Daouia and Simar (2007)
  - 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 Frontier -2-

Partial order frontiers: Mathematical definition for univariate output

- **Full Frontier Benchmark:** $\varphi(x) = \inf\{y | F_{Y|X}(y|x) \geq 1\}$ and
- **Less Extreme Benchmarks:**
  - **Order-m frontier:**

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

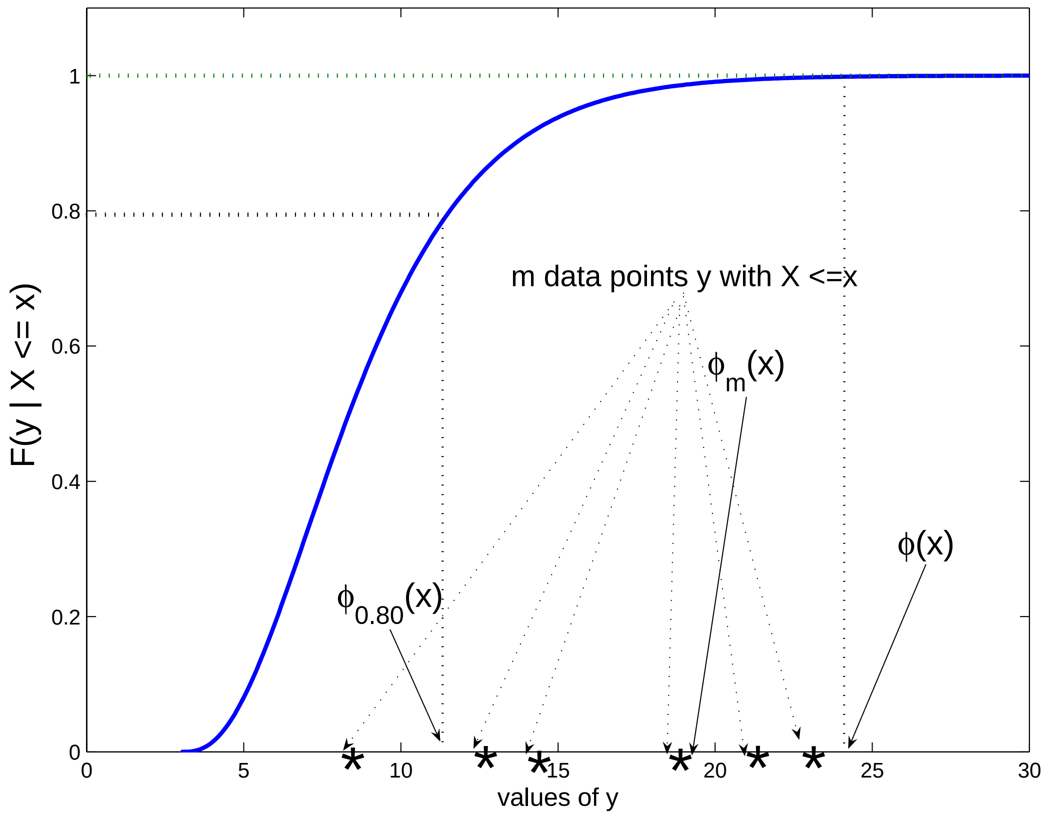
  - **Order-α quantile frontier:**

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

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 Frontier -3-



Picture of $F_{Y|X}(y|x) = \mathrm{Prob}(Y \le y|X \le x)$

Illustration of **full and partial frontiers**: one output with $m = 6$ and $\alpha = 0.80$

Robust Frontier -4-

Nonparametric estimators of partial order frontier

- **Plug-in principle**

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

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

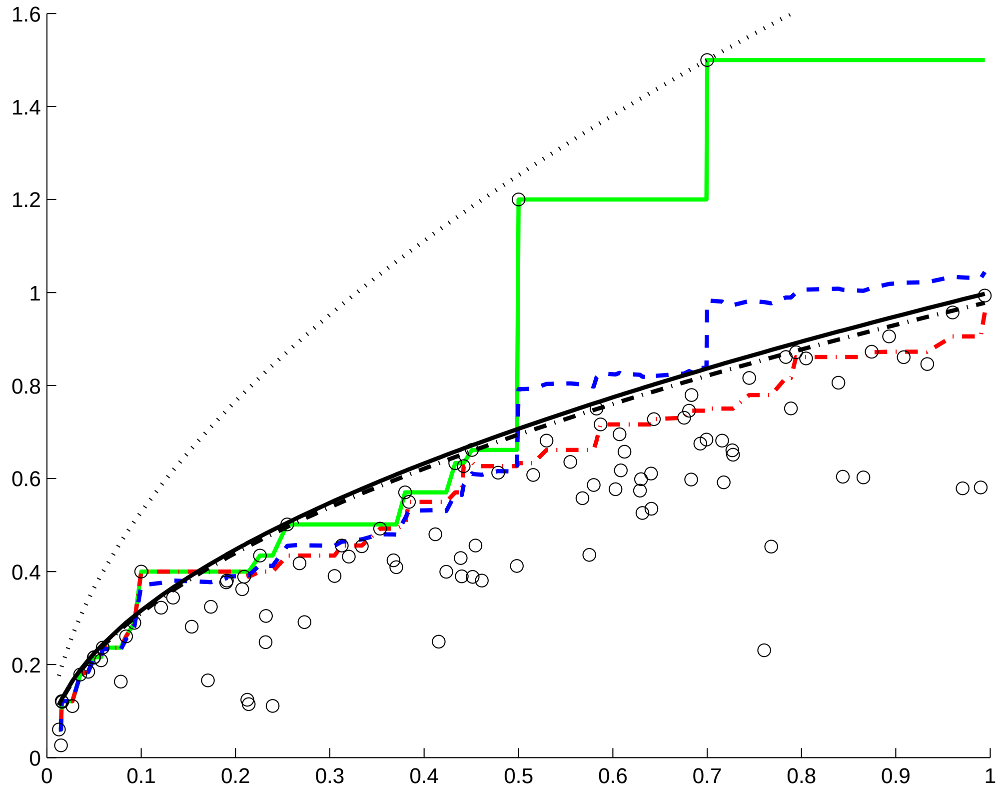
- **Properties**
  - **$\sqrt{n}$-consistency and asymptotic normality:**

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

  - **Convergence to FDH estimator:**

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

- **Choice of m and α:** tune the percentage of points left out estimated partial frontier, see Simar (2003), Daraio, Simar (2005, 2007a).



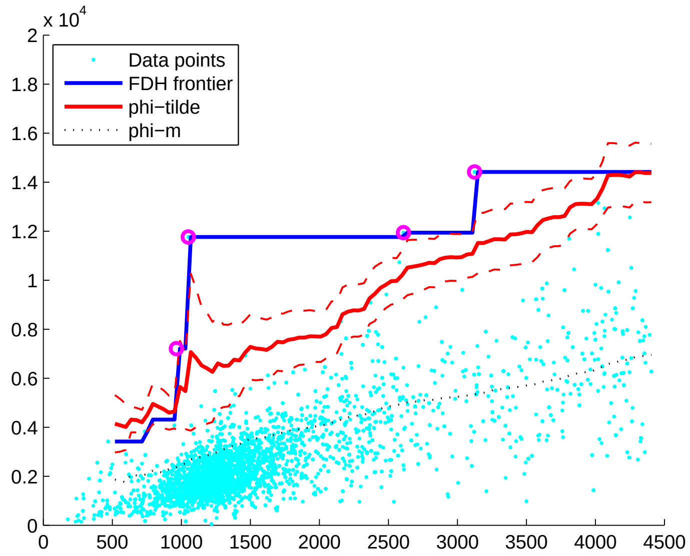
In solid black line, the **true** frontier $y = x^{0.5}$. In green solid, the **FDH** frontier estimate, in blue dashed the estimated **order-m** frontier and in dash-dot red the estimate of the **order-α** frontier. In black dotted, the shifted OLS estimate and in dash-dot black, the parametric stochastic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

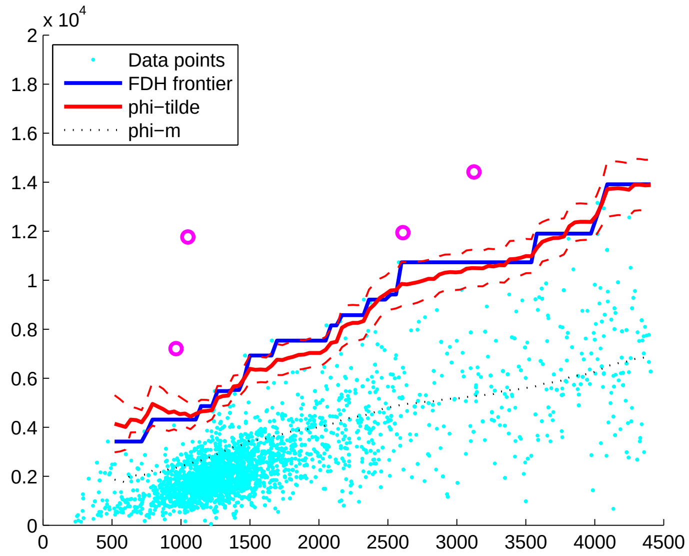
Robust Frontier -5-

Robust Nonparametric Estimator of Full-Frontier $\varphi(x)$, Daouia, Florens, Simar (2009, 2010)

- If $m = m(n)$ (and $\alpha = \alpha(n)$) converges to ∞ (to 1) when $n \to \infty$, but at a slow rate, we obtain an estimator (after bias correction) that converges to the full frontier with a Normal limiting distribution
  - Easy to build confidence intervals for $\varphi(x)$ using Normal Tables.
- **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**.

Robust Frontier -5-





Post Offices in France (from Daouia, Florens, Simar, 2009).

Left panel: estimation with the 4 extreme points.

Right panel: estimation without these 4 points

IV.2 Lack of Economic Interpretation

Parametric Approximation of Deterministic Frontiers -1-

- **Parametric models:** easy economic interpretation of the model (returns to scale, elasticities, elasticities of substitution, . . .)
- **Standard parametric approches: some drawbacks**
  - strong restrictive assumptions on the stochastic part of the models
  - sensitive to extreme/outliers
  - most are "regression-based" models and capture the shape of the cloud of points near its center (not at the efficient boundary)
- **Two stage semiparametric approaches:** Simar (1992), Florens, Simar (2005), Daouia, Florens, Simar (2008)
  - First estimate the efficient frontier using **nonparametric** or **robust nonparametric methods**;
  - Then fit, by standard OLS, the approriate **parametric model** on the obtained nonparametric frontier

Parametric Approximation of Deterministic Frontiers -2-

- **More sensible estimator** of the parametric frontier model and **allows for some noise** by tunning the robustness parameter.

- **Asymptotic theory** of the resulting estimators (for fix m and fix α):

  If FDH is used as 1st step: $\hat{\theta}_n \xrightarrow{p} \theta_0$

  If order-m is used is used as 1st step: $\sqrt{n}(\hat{\theta}_n^m - \theta_0^m) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, V_m)$

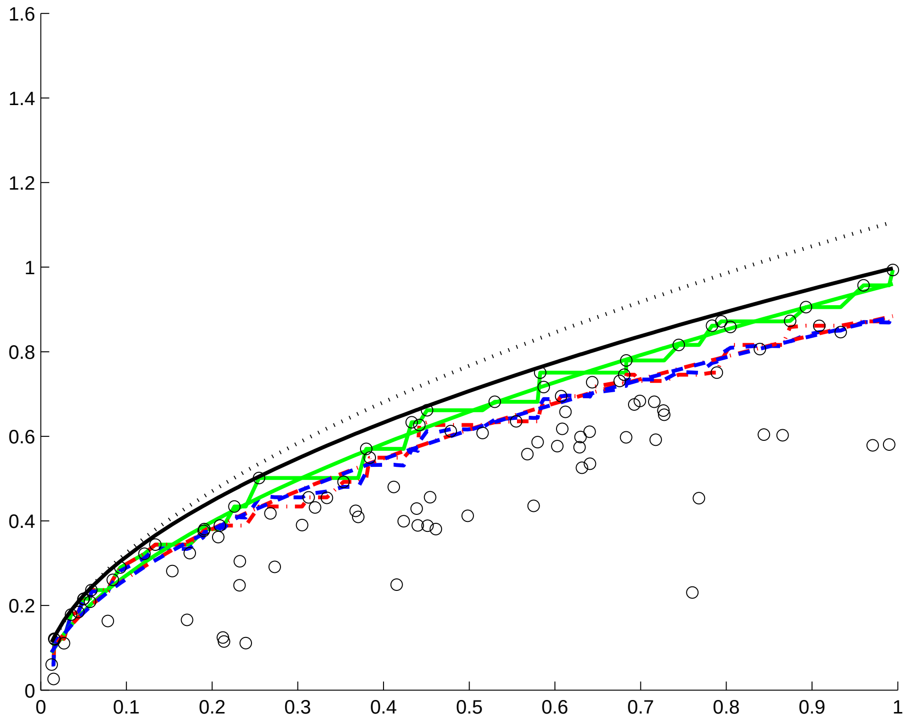
  If order-α is used is used as 1st step: $\sqrt{n}(\hat{\theta}_n^\alpha - \theta_0^\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, V_\alpha)$

  where θ_0, (θ_0^m, θ_0^α), are the **pseudo-true values** of the parameters of the best approximation of the corresponding frontier $\varphi(x)$, ($\varphi_m(x)$, $\varphi_\alpha(x)$).

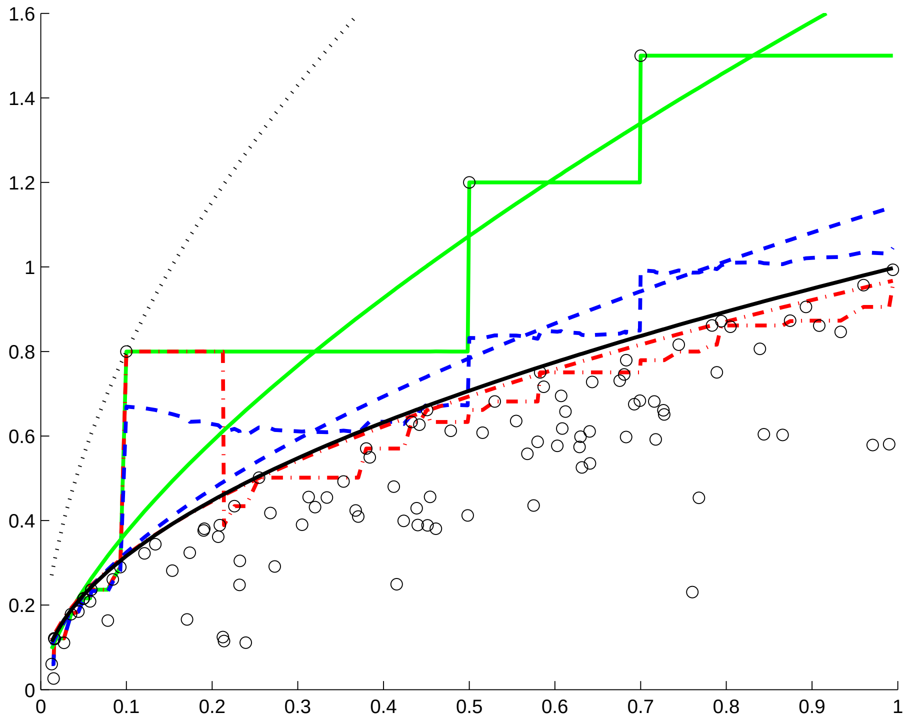
- If $m(n) \to \infty$ and $\alpha(n) \to 1$ as $n \to \infty$ at appropriate rates:

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0\,; \qquad \hat{\theta}_n^{m(n)} \xrightarrow{a.s.} \theta_0\,; \qquad \hat{\theta}_n^{\alpha(n)} \xrightarrow{a.s.} \theta_0$$

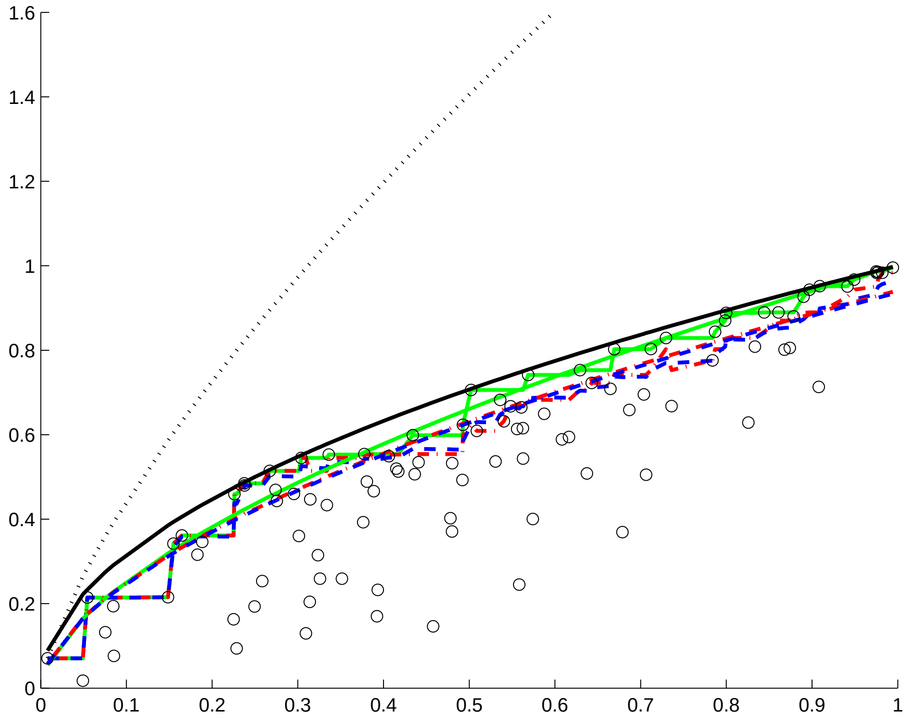
- **Multivariate case:** multi-input/muti-output, see Daraio and Simar (2007a)



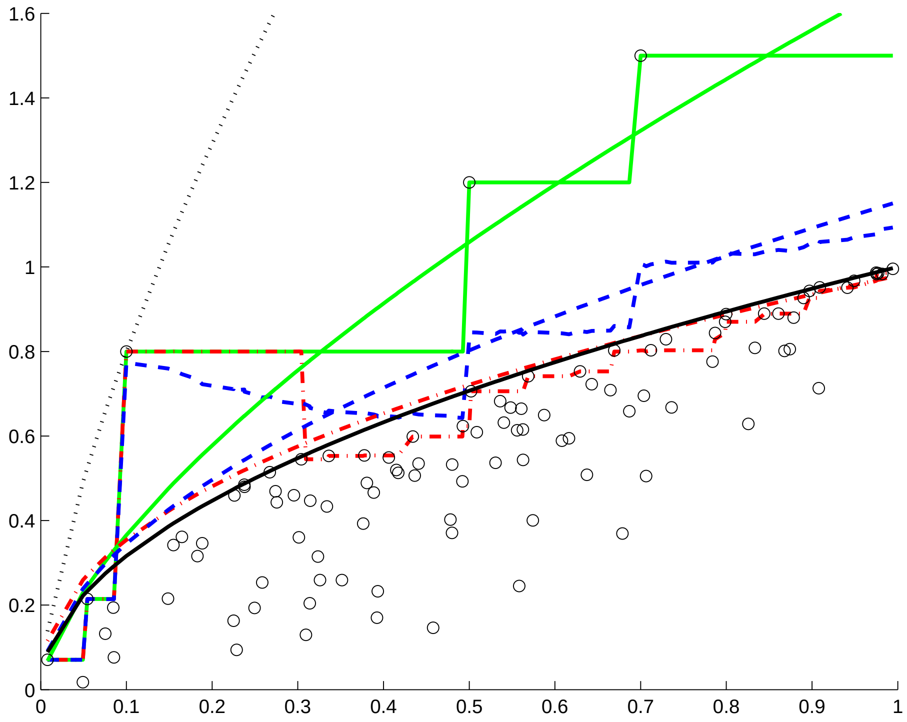
In solid black line, the true frontier $y = x^{0.5}$ **homoscedastic inefficiency**. In cyan solid, the FDH frontier, in blue dashed the order-m frontier and in dash-dot red the order-α frontier. Here, $m = 20$ and $\alpha = .9622$. In black dotted, the shifted OLS estimate.



Same with 3 outliers included.



Same with **heteroscedastic inefficiency**. In cyan solid, the FDH frontier estimate, in blue dashed the order-m frontier and in dash-dot red the order-α frontier. Here, $m = 20$ and $\alpha = .9622$. In black dotted, the shifted OLS estimate.



Same with 3 outliers included.

IV.3 Heterogeneity

Introducing Environmental Factors -1-

- **Motivation**
  - The analysis of productive efficiency should have two components:
    1. Estimation of a production frontier (best-practice) which serve as a benchmark against which **efficiency** of a producer can be measured;
    2. Incorporation into the analysis of **exogenous variables** (Z) which are neither inputs, nor outputs, and so are **not under the control of the producer**, but which may influence the process.
  - How to explain inefficiencies of firms by these factors?
  - How to introduce heterogeneity in the production process?

Introducing Environmental Factors -2-

- **One-stage approaches** Banker and Morey (1986)
  - Z is like an input(favorable) or like an output (defavorable) $\Rightarrow$ Adapt FDH/DEA
  - Free disposability ? Convexity ? RTS assumption ?
  - Which direction for Z?
  - What if the effect of Z changes?
    (say, favorable if $Z \leq z_0$ and then defavorable or neutral for $Z > z_0$)
- **Two-stage approaches** Simar and Wilson (2007, 2011)
  - DEA efficiency scores are regressed on Z (in an **appropriate** way)
  - **Implicit Separability Condition**:
    - Z does not influence Ψ
    - Z only affects the probability of being more or less efficient
    - The second stage regression is nonstandard (correlation among efficiency scores, bias,...): **inference by bootstrap.**

Traditional 2-stage approaches

- **First stage** get efficiency estimates $\widehat{\lambda}(X_i, Y_i)$ (or $\widehat{\theta}(X_i, Y_i)$, $\widehat{\gamma}(X_i, Y_i)$,...) with respect to $\widehat{\Psi}$ (by DEA or FDH, ...)
- **Second stage** regression of $\widehat{\lambda}(X_i, Y_i)$ on Z.
  - Parametric models (truncated regression, logistic, etc,...)
  - Nonparametric models (truncated, etc,...)
- **Problems: $\Psi^z = \{(x, y)|Z = z,$ x can produce y$\}$** Simar and Wilson (2007, 2011b):
  - If $\Psi^z \neq \Psi$, what is the **Economic meaning** of $\lambda(x, y)$ (and so, of $\widehat{\lambda}(X_i, Y_i)$), for a unit facing environmental conditions z?
  - Separability issue: condition for giving economic meaning to $\widehat{\Psi}$ and $\widehat{\lambda}(x, y)$.

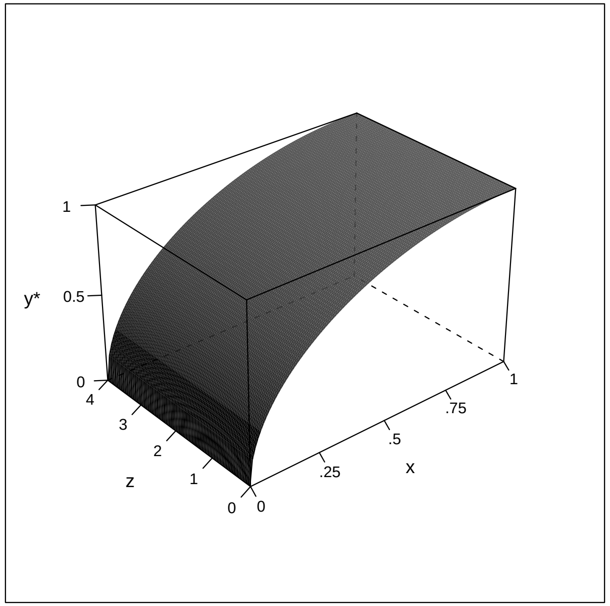
"Separability" condition: $\Psi^z = \Psi$, for all $z \in \mathcal{Z}$.

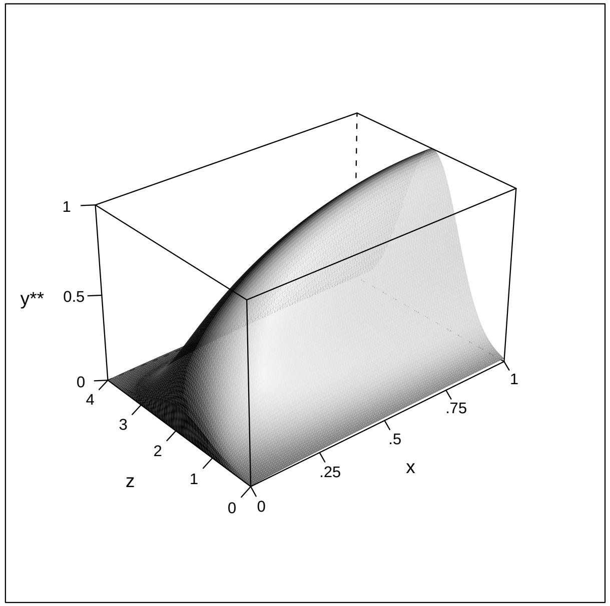
  - Even if separability holds, **Inference** in second stage is nonstandard (bootstrap).

"Separability" Condition

$$g(X) = [1 - (X - 1)^2]^{1/2}$$

$$Y^* = g(X)e^{-(Z-2)^2 U} \qquad Y^{**} = g(X)e^{-(Z-2)^2}e^{-U}$$





Left Panel: **Separable**, Right Panel: **Not Separable**

Conditional Efficiency -1-

- **Conditional Measures** Cazals, Florens, Simar (2002), Daraio Simar (2005, 2007a, 2007b), Jeong, Park, Simar (2010)
  - **The DGP** (A Model for the Production process) is now characterized by
    - $F(x, y|z) = \text{Prob}(X \leq x, Y \leq y|Z = z)$ or
    - $H(x, y|z) = \text{Prob}(X \leq x, Y \geq y|Z = z)$
    - The attainable set is Ψ^z: the support of $F(x, y|z)$
  - **Natural and very easy:** A firm combines inputs $X \in \mathbb{R}_+^p$ and outputs $Y \in \mathbb{R}_+^q$ facing the environmental conditions $Z \in \mathbb{R}^r$
    - No **separability** conditions
    - No **prior** information of the role of Z (favorable or not to the process)
  - Note that the **separability condition** of 2-stages methods relies on:

$$\Psi \equiv \Psi^z \quad \text{for all} \quad z.$$

Conditional Efficiency -2-

- **Conditional efficiency score**
  - Same idea as the unconditional measure:

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

where
$S_{Y|X,Z}(y|x,z) = H_{XY|Z}(x,y|z)/H_{XY|Z}(x,0|z) = \text{Prob}(Y \geq y \mid X \leq x, Z = z)$.

- **Nonparametric estimator**: kernel smoothing on Z (here continuous)

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

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

Conditional Efficiency -3-

- **Conditional FDH efficiency estimator**: Kernels with compact support,

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

- **Conditional FDH attainable set**:

$$\widehat{\Psi}^Z_{FDH} = \{(x,y) \in \mathbb{R}^{p+q}_+ \mid x \ge x_i, y \le y_i \text{ for } i \text{ s.t. } ||Z_i - z|| \le h\}$$

- **DEA versions**: Convexify the FDH attainable set, see Daraio, Simar (2007b)

$$\widehat{\Psi}^Z_{DEA} = \{(x,y) \in \mathbb{R}^{p+q}_+ \mid x \ge \sum_{\{i|||Z_i - z|| \le h\}} \gamma_i x_i, \quad y \le \sum_{\{i|||Z_i - z|| \le h\}} \gamma_i y_i$$
$$\text{for } \gamma_i \text{ s.t. } \sum_{\{i|||Z_i - z|| \le h\}} \gamma_i = 1\},$$
$$\widehat{\lambda}_{DEA}(x,y|z) = \sup\{\lambda \mid (x, \lambda y) \in \widehat{\Psi}^Z_{DEA}\}.$$

Conditional Efficiency -4-

- **Properties**
  - Optimal bandwidth selection by data-driven methods, Badin, Daraio, Simar (2010)
  - Asymptotic properties: similar to FDH/DEA with n replaced by nh^r, Jeong, Park, Simar (2010)
  - Allow to detect the direction of the "influence" of Z on efficiency, see Dario, Simar (2005, 2007a)
  - Inference (confidence intervals) by bootstrap
  - Robust versions (using order-m and order-α) are also available
  - Z can be continuous, categorical or discrete

Conditional Efficiency -5-

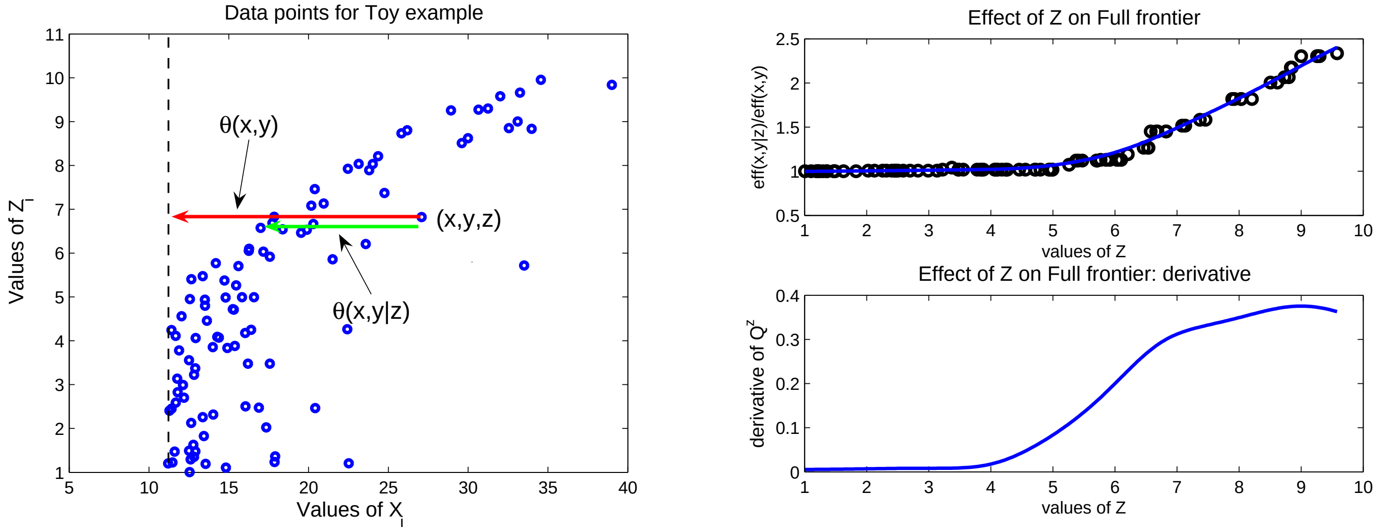
- **Usefulness**
  - Define a **"pure measure of technical efficiency"**, Badin, Daraio, Simar (2011)
    - Eliminate most of the influence of Z on $\widehat{\lambda}(x, y|z)$ by using a flexible location-scale nonparametric model: $\widehat{\lambda}(x, y|z) = \mu(z) + \sigma(z)\varepsilon$, where $\mu(z)$ and $\sigma(z)$ are unspecified functions
    - $\hat{\epsilon}_i$ allows to rank firms facing different operating conditions.
- **N.B.: An other approach**: Florens, Simar, Van Keilegom (2011).
  - First eliminate influence of Z on inputs X and outputs Y by using two flexible location-scale nonparametric models
  - The residuals are **"pure inputs and outputs"** $\tilde{X}_i$ and $\tilde{Y}_i$
  - Search for the frontier in these new units, to define "pure measure of technical efficiency"
  - Full frontier and order-m frontiers

Conditional Efficiency, Example -1-

- **A Toy example**:
  - No output ($Y_i \equiv 1$) and one input (input orientation)
  - Z has no effect on X when $Z \leq 5$ and then a defavorable effect on X when $Z > 5$.
  - The input are generated according

$$X_i = 5^{1.5}\, 1\!\!I(Z_i <= 5) + Z_i^{1.5}\, 1\!\!I(Z_i > 5) + U_i,$$

    where $Z_i \sim \mathrm{U}(1, 10)$, $U_i \sim \mathrm{Expo}(\mu = 3)$ and $n = 100$.



Effect of Z on the ratios $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$.

Conditional Efficiency, Examples -2a-

- **2 inputs/ 2 outputs : output orientation**
  - The efficient frontier is described by: $y^{(2)} = 1.0845(x^{(1)})^{0.3}(x^{(2)})^{0.4} - y^{(1)}$.
  - $X_i^{(j)} \sim U(1,2)$ and $\tilde{Y}_i^{(j)} \sim U(0.2,5)$ for $j = 1,2$.
  - The output efficient random points on the frontier are

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

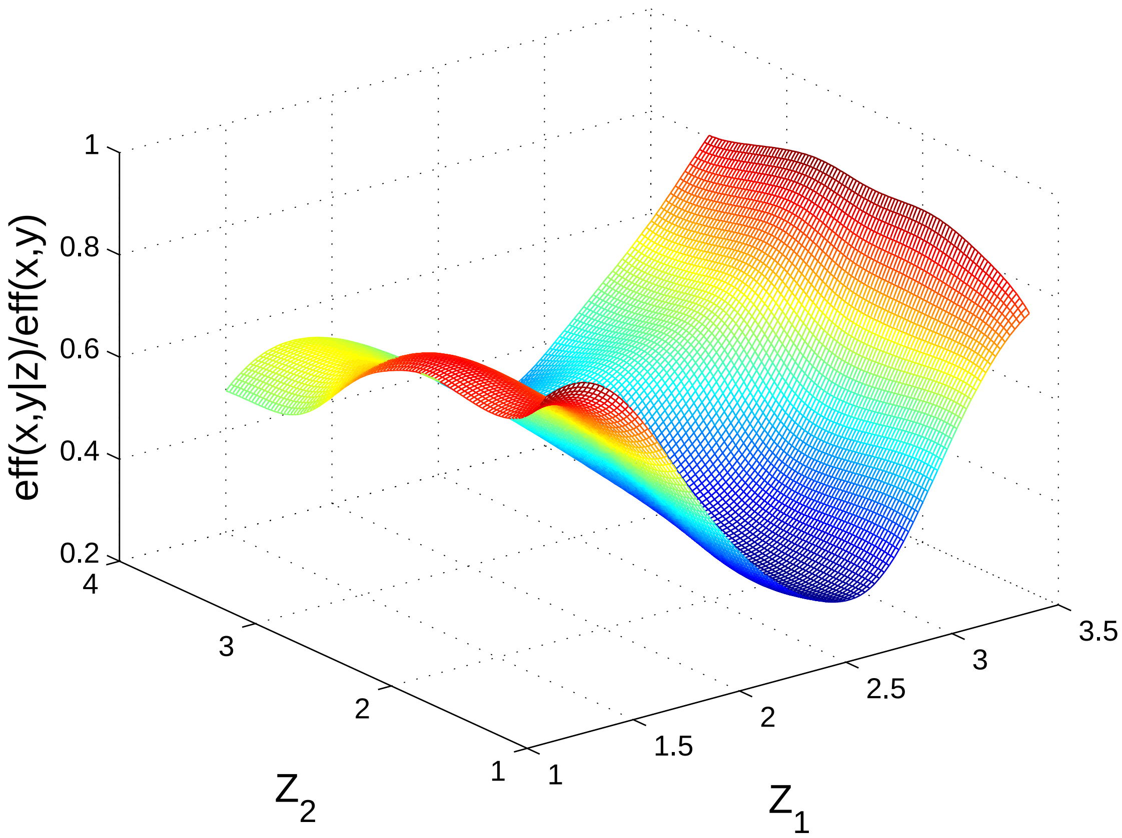
    where $S_i = \tilde{Y}_i^{(2)}/\tilde{Y}_i^{(1)}$ represent the generated random rays in the output space.
  - The efficiencies are simulated according to $\exp(-U_i)$
  - The observed output are defined by $Y_i = Y_{i,eff} * \exp(-U_i)$ where $U_i \sim Exp(\mu_U = 1/2)$.
  - $n = 100$.

Conditional Efficiency, Examples -2b-

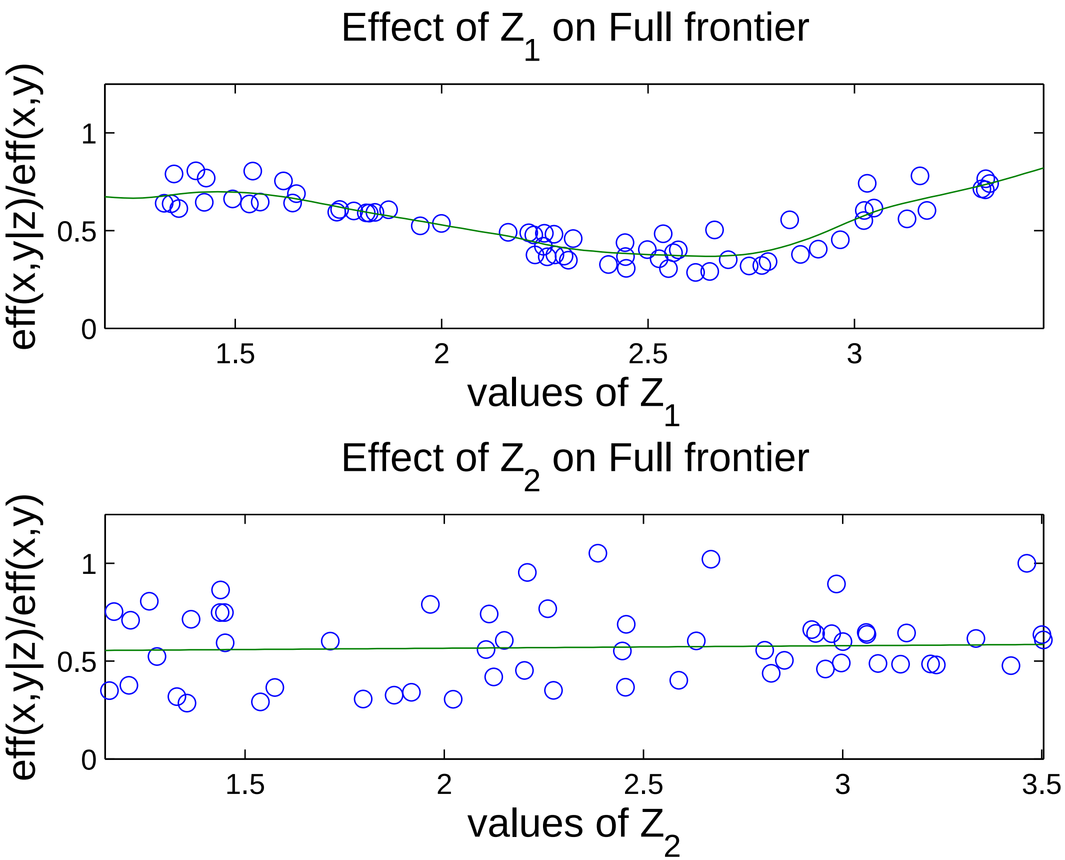
- **Environmental factors Z bivariate**
  - We generate two independent uniform variables $Z_j \sim U(1,4)$ to build the bivariate variable $Z = (Z_1, Z_2)$.
  - The influence of Z on the production process is described by:

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

    - Z_1 pushes the efficient frontier above when far from 2.5, in both directions, with a **cubic effect**,
    - Z_2 has **no effect** on the frontier or on the distribution of inefficiencies: Z_2 is irrelevant.
    - Note that there is no interaction between Z_1 and Z_2 (independent) and no interaction between X and Z.
  - Remember: only $n = 100$ observations, with $p = q = r = 2$!



Smoothed nonparametric surface regression of $\widehat{\lambda}_n(x,y|z)/\widehat{\lambda}_n(x,y)$ on Z_1 and Z_2.



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).

IV.4 Introducing Noise

Nonparametric Stochastic Frontiers -1-

- **Basic Idea:** localize (using kernels) an anchorage parametric model, Kumbhakar, Park, Simar, Tsionas (2007)

$$Y_i = r(X_i) + v_i - u_i$$

  - $u|X = x \sim |\mathcal{N}(0, \sigma_u^2(x))|$ and $v|X = x \sim \mathcal{N}(0, \sigma_v^2(x))$ and u and v being independent conditionally on X.
  - $r(x), \sigma_u^2(x)$ and $\sigma_v^2(x)$ are **unknown functional parameters**
  - Estimation by **Local Maximum Likelihhood** method: $r(x), \sigma_u^2(x)$ and $\sigma_v^2(x)$ are approximated by local polynomials (linear or quadratic).
  - Asymptotic properties are available
  - Bandwidths selection by **LS cross-validation**

Nonparametric Stochastic Frontiers -2-

- **Multivariate extension**: Simar (2007), Simar, Zelenyuk (2011)
  - Use (partial-)polar coordinates: $(x, y) \Leftrightarrow (\omega, \eta, x)$, where $\omega \in \mathbb{R}_+$ is the modulus and $\eta \in [0, \pi/2]^{q-1}$ is the amplitude (angle) of the vector y.
  - The joint density $f_{X,Y}(x, y)$ induces a density on (ω, η, x):

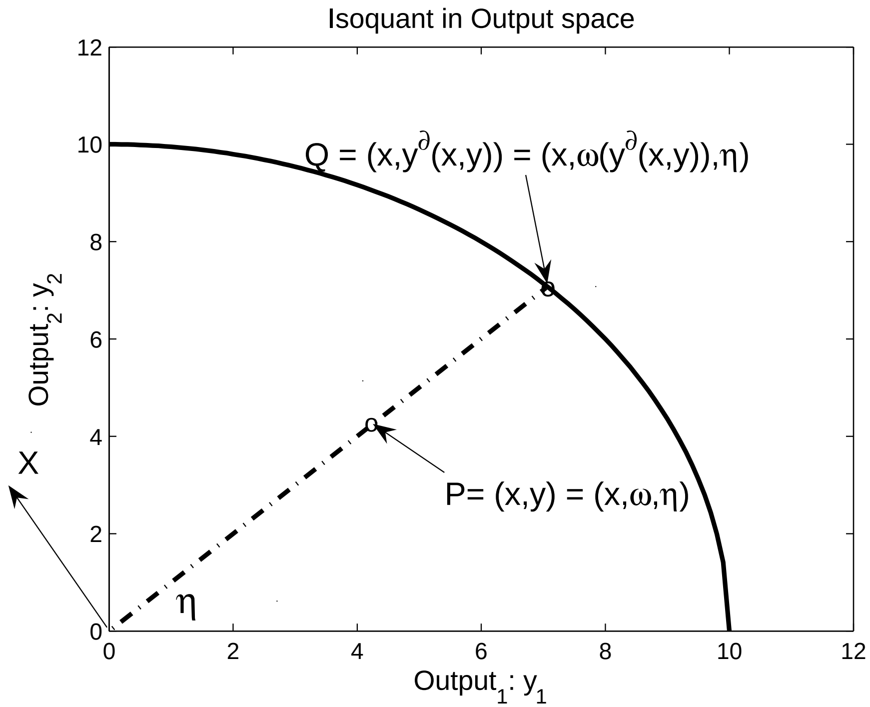
$$f_{\omega,\eta,X}(\omega, \eta, x) = f_\omega(\omega \mid \eta, x)\, f_{\eta,X}(\eta, x)$$

  - For a given (x, y) the frontier point $y^\partial(x, y) = \lambda(x, y)\, y$ has a modulus:

$$\omega(y^\partial(x, y)) = \sup\{\omega \in \mathbb{R}^+ \mid f_\omega(\omega \mid \eta, x) > 0\}$$

- **Back to a univariate frontier problem!**
  - Given (η, x) find $\omega(y^\partial(x, y))$.

Nonparametric Stochastic Frontiers -3-



Polar coordinates in the output space for a particular section $Y(x)$. Output efficiency of $P = (x, y)$ is $\lambda(x, y) = |OQ|/|OP| = \omega(y^\partial(x, y))/\omega(x, y) \geq 1$.

Nonparametric Stochastic Frontiers -4-

- **The Model**:
  - The observations are made on noisy data in the output radial-direction
  - The data $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ have polar coordinates (ω_i, η_i, X_i)

$$\omega_i = \omega(y^\partial(X_i, Y_i))\, e^{-u_i}\, e^{v_i},$$

  where $u_i > 0$ is inefficiency and v_i is noise $(E(v_i|X_i, Y_i) = 0)$.

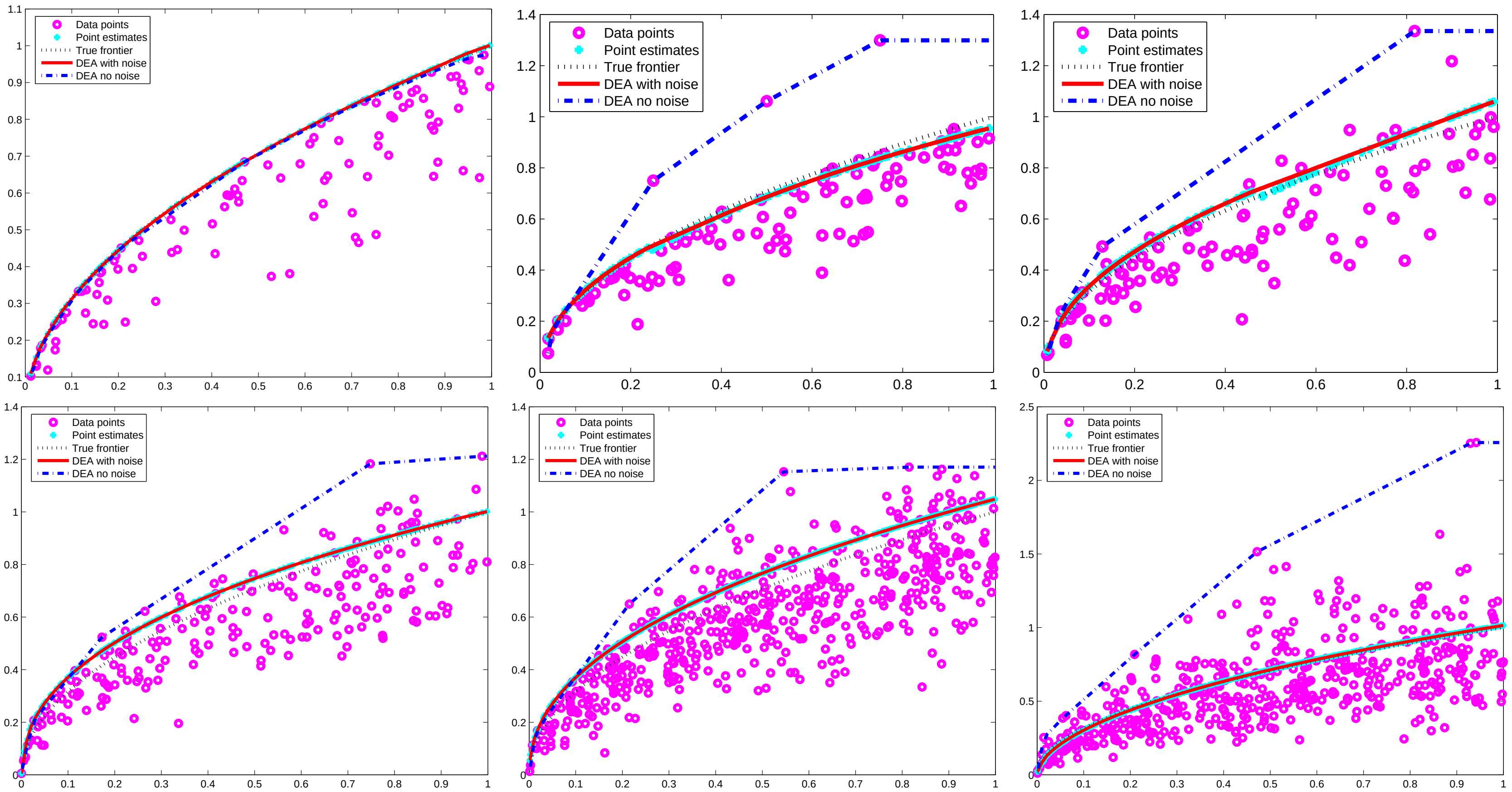
  - $\omega(y^\partial(X_i, Y_i))$ is only a function of (η_i, X_i).
- **In the log-scale**, the model could be written as

$$\log \omega_i = r(\eta_i, X_i) - u_i + v_i,$$

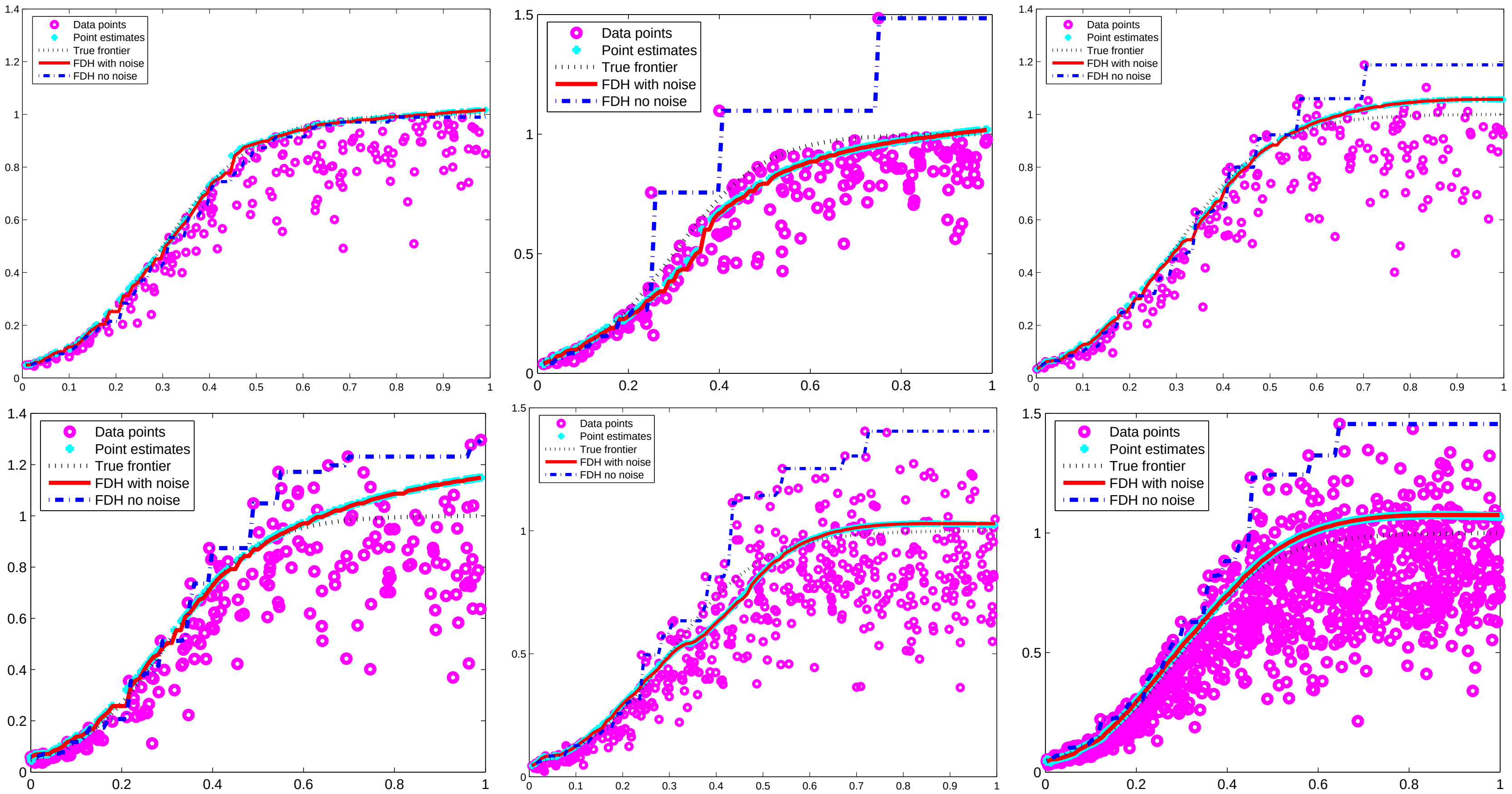
with $u_i > 0$ and $E(v_i|\eta_i, X_i) = 0$.

Nonparametric Stochastic Frontiers -5-

- **Stochastic Versions of DEA/FDH :** Two-stage procedure
  - [1] "Whitening the noise": Compute the consistent estimator of the frontier levels $\hat{r}(\eta_i, X_i)$ for each data points
    * This gives points (X_i, Y_i^*) where $Y_i^* = \exp(\hat{r}(\eta_i, X_i))Y_i/\omega_i$
  - [2] Run a DEA (or FDH) program with reference set (X_i, Y_i^*).
- **Summary:**
  - Very encouraging results
  - Computationally demanding (cross-validation for bandwidth selection)
  - Below, some bivariate examples (see multivariate examples in Simar and Zelenyuk, 2011)



(a) $n = 100, \rho_{\text{nts}} = 0$, (b) $n = 103, \rho_{\text{nts}} = 0 + 3$ outliers, (c) $n = 100, \rho_{\text{nts}} = 1$, (d) $n = 200, \rho_{\text{nts}} = 1$, (e) $n = 500, \rho_{\text{nts}} = 1$, (f) $n = 500, \rho_{\text{nts}} = 2$.



(a) $n = 200, \rho_{\text{nts}} = 0$, (b) $n = 203, \rho_{\text{nts}} = 0 + 3$ outliers, (c) $n = 200, \rho_{\text{nts}} = 0.5$, (d) $n = 200, \rho_{\text{nts}} = 1$, (e) $n = 500, \rho_{\text{nts}} = 1$, (f) $n = 1000, \rho_{\text{nts}} = 1$.

Conclusions

- **Nonparametric models $\mathcal{NP}$ are Econometric Models**
  - Flexible and can be "robustified", Inference is available (bootstrap)
  - Noise can be introduced
  - Environmental factors (heterogeneity) can be introduced
- **$\mathcal{P}$ and $\mathcal{NP}$ are complimentary models**
  - $\mathcal{NP}$ models can be used to check (test) $\mathcal{P}$ models (not the contrary).
  - Parametric approximations of $\mathcal{NP}$ models can be useful for economic analysis.
  - Semiparametric models should be developed.
- **Other challenges**
  - Panel Data: introduce dynamic behavior of units
  - Theory for functions of DEA/FDH scores: Kneip, Simar and Wilson (2011)
    * Useful for justifying testing procedures
    * RTS, Convexity, Simar and Wilson (2011a), Separabilty, Daraio, Simar and Wilson (2010),...

References

- **Basic References**

  - Fried, H., Lovell, K. and S. Schmidt (eds) (2008), *The Measurement of Productive Efficiency*, 2nd Edition, Oxford University Press.
  - Kumbhakar, S.C. and C.A.K. Lovell (2000), *Stochastic Frontier Analysis*, Cambridge University Press.
  - Daraio, C. and L. Simar (2007a), *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications*, Springer, New-York.
  - Simar, L. and P.W. Wilson (2008), Statistical Inference in Nonparametric Frontier Models: recent Developments and Perspectives, in *The Measurement of Productive Efficiency*, 2nd Edition, Harold Fried, C.A.Knox Lovell and Shelton Schmidt, editors, Oxford University Press, 2008.

- **References list**

- Aigner, D.J., Lovell, C.A.K. and P. Schmidt (1977), Formulation and estimation of stochastic frontier models, *Journal of Econometrics*, 6, 21-37.
- Aragon, Y. and Daouia, A. and Thomas-Agnan, C. (2005), Nonparametric Frontier Estimation: A Conditional Quantile-based Approach, *Econometric Theory*, 21, 358–389.
- Badin, M., Daraio, C. and L. Simar (2010), Optimal Bandwidth Selection for Conditional Efficiency Measures: a Data-driven Approach, *European Journal of Operational Research*, 201, 633–640
- Badin, L., Daraio, C. and L. Simar (2011), How to measure the impact of environmental factors in a nonparametric production model? Discussion paper 2011/19, Institut de Statistique, UCL.
- Badin, L. and L. Simar (2009), A bias corrected nonparametric envelopment estimator of frontiers, *Econometric Theory*, 25, 5, 1289–1318.
- Banker, R.D. and R.C. Morey (1986), Efficiency analysis for exogenously fixed inputs and outputs, *Operations Research*, 34(4), 513–521.
- Cazals, C. Florens, J.P. and L. Simar (2002), Nonparametric Frontier Estimation: a Robust Approach , in *Journal of Econometrics*, 106, 1–25.
- Chambers, R. G., Y. Chung, and R. Färe (1998), Profit, directional distance functions, and nerlovian efficiency, *Journal of Optimization Theory and Applications*, 98, 351–364.

- Charnes, A., Cooper W.W. and E. Rhodes (1978), Measuring the inefficiency of decision making units, *European Journal of Operational Research* 2 (6), 429-444.
- Daouia, A., J.P. Florens and L. Simar (2008), Functional Convergence of Quantile-type Frontiers with Application to Parametric Approximations, *Journal of Statistical Planning and Inference*, 138, 708–725.
- Daouia, A., Florens, J.P. and L. Simar (2009), Regularization of Non-parametric Frontier Estimators. Discussion paper 0922, Institut de Statistique, UCL.
- Daouia, A., Florens, J.P. and L. Simar (2010), Frontier estimation and extreme values theory. *Bernoulli*, 16(4), 1039–1063.
- Daouia, A. and L. Simar (2005), Robust Nonparametric Estimators of Monotone Boundaries, *Journal of Multivariate Analysis*, 96, 311–331.
- Daouia, A. and L. Simar (2007), Nonparametric efficiency analysis: a multivariate conditional quantile approach, *Journal of Econometrics*, 140, 375–400.
- 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), A Robust Nonparametric Approach to Evaluate and Explain the Performance of Mutual Funds, *European Journal of Operational Research*, vol 175, 1, 516–542.

- Daraio, C. and L. Simar (2007a), *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications*, Springer, New-York.
- Daraio, C. and L. Simar (2007), Conditional nonparametric frontier models for convex and non convex technologies: a unifying approach, *Journal of Productivity Analysis*, vol 28, 13–32.
- Daraio, C., Simar, L. and P.W. Wilson (2010), Testing whether Two-Stage Estimation is Meaningful in Non-Parametric Models of Production, Discussion paper 1031, Institut de Statistique, UCL.
- Debreu, G. (1951), The coefficient of ressource utilization, *Econometrica*, 19:3, 273-292.
- Deprins, D., Simar, L. and H. Tulkens (1984), Measuring labor inefficiency in post offices. In *The Performance of Public Enterprises: Concepts and measurements*. M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243–267.
- Farrell, M.J. (1957), The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, 120, 253–281.
- Färe, R. and S. Grosskopf (2000), Theory and application of directional distance functions, *Journal of Productivity Analysis* 13, 93–103.
- Färe, R., and S. Grosskopf (2004), *New Directions: Efficiency and Productivity*, Boston: Kluwer Academic Publishers.
- Färe, R., S. Grosskopf, and C. A. K. Lovell (1985), *The Measurement of Efficiency of Production*, Boston: Kluwer-Nijhoff Publishing.

- Florens, J.P. and L. Simar, (2005), Parametric Approximations of Nonparametric Frontier, *Journal of Econometrics*, vol 124, 1, 91–116
- Fried, H., Lovell, K. and S. Schmidt (eds) (2008),*The Measurement of Productive Efficiency*, 2nd Edition, Oxford University Press.
- Gijbels, I., E. Mammen, B.U. Park and L. Simar (1999), On Estimation of Monotone and Concave Frontier Functions, *Journal of the American Statistical Association*, vol 94, 445, 220-228.
- Hall, P. and L. Simar (2002), Estimating a Changepoint, Boundary or Frontier in the Presence of Observation Error, *Journal of the American Statistical Association*, 97, 523–534.
- Jeong, S.O. , B. U. Park and L. Simar (2010), Nonparametric conditional efficiency measures: asymptotic properties. *Annals of Operations Research*, 173, 105–122.
- Jeong, S.O. and L. Simar (2006), Linearly interpolated FDH efficiency score for nonconvex frontiers, *Journal of Multivariate Analysis*, 97, 2141–2161.
- Kneip, A., Park, B.U. and Simar, L. (1998). : A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14, 783–793.
- Kneip, A., Simar, L. and I. Van Keilegom (2010), Boundary Estimation in the Presence of Measurement Errors. Discussion paper 1046, Institut de Statistique, UCL.
- 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.

- Kneip, A., Simar, L. and P.W. Wilson (2011), A Computational Efficient, Consistent Bootstrap for Inference with Non-parametric DEA Estimators. Discussion paper 0903, Institut de Statistique, UCL, in press *Computational Economics.*
- Kneip, A., Simar, L. and P.W. Wilson (2011), Central Limit Theorems for DEA efficiency scores: when bias can kill the variance. Discussion paper 2011/**, Institut de Statistique, UCL.
- Koopmans, T.C.(1951), An analysis of production as an efficient combination of activities, in Koopmans, T.C. (ed) *Activity Analysis of Production and Allocation*, Cowles Commision for Research in Economics, Monograph 13, John-Wiley, New-York.
- Korostelev, A., Simar, L. and A. Tsybakov (1995a), Efficient Estimation of Monotone Boundaries, *Annals of Statistics*, 23(2), 476–489.
- Korostelev, A., Simar, L. and A. Tsybakov (1995b), On Estimation of Monotone and Convex Boundaries, *Publications des Instituts de Statistique des Universités de Paris*, 1, 3–18.
- Kumbhakar, S.C. and C.A.K. Lovell (2000), *Stochastic Frontier Analysis*, Cambridge University Press.
- Kumbhakar, S.C. , Park, B.U., Simar, L. and E.G. Tsionas (2007), Nonparametric stochastic frontiers: a local likelihood approach, *Journal of Econometrics*, 137, 1, 1–27.
- Mouchart, M. and L. Simar (2002), Efficiency analysis of Air Controlers: first insights, Consulting report 0202, Institut de Statistique, Université Catholique de Louvain, Belgium.

- Park, B.U., Jeong, S.-O. and L. Simar (2010), Asymptotic Distribution of Conical-Hull Estimators of Directional Edges. *Annals of Statistics*, Vol 38, 6, 1320–1340.
- Park, B. Simar, L. and Ch. Weiner (2000), The FDH Estimator for Productivity Efficiency Scores: Asymptotic Properties, *Econometric Theory*, Vol 16, 855–877.
- Park, B., L. Simar and V. Zelenyuk (2008), Local Likelihood Estimation of Truncated Regression and its Partial Derivatives: Theory and Application, *Journal of Econometrics*, 146, 185–198.
- Ritter, C. and L. Simar (1997), Pitfalls of normal-gamma stochastic frontier models, *Journal of Productivity Analysis*, 8, 167–182.
- Schubert, T. and L. Simar (2011), Innovation and export activities in the German mechanical engineering sector: an application of testing restrictions in production analysis. *Journal of Productivity Analysis*, 36, 55–69.
- Shephard, R.W. (1970). *Theory of Cost and Production Function.* Princeton University Press, Princeton, New-Jersey.
- Simar, L. (1992), Estimating efficiencies from frontier models with panel data: a comparison of parametric, non-parametric and semi-parametric methods with bootstrapping, *Journal of Productivity Analysis*, 3, 167-203.

- Simar, L. (2003), Detecting Outliers in Frontiers Models: a Simple Approach, *Journal of Productivity Analysis*, 20, 391–424.
- Simar, L. (2007), How to Improve the Performances of DEA/FDH Estimators in the Presence of Noise, *Journal of Productivity Analysis*, vol 28, 183–201.
- Simar, L. and A. Vanhems (2010), Probabilist Characterization of Directional Distances and their Robust versions. Discussion paper 1040, Institut de Statistique, UCL, to appear *Journal of Econometrics*.
- Simar, L., Vanhems, A. and P.W. Wilson (2011), Statistical inference with DEA estimators of directional distances. Discussion paper 2011/**, Institut de Statistique, UCL.
- Simar, L. and P. Wilson (1998), Sensitivity of efficiency scores : How to bootstrap in Nonparametric frontier models, *Management Sciences*, 44, 1, 49–61.
- Simar, L. and P. Wilson (1999a), Some problems with the Ferrier/Hirschberg Bootstrap Idea, *Journal of Productivity Analysis*, 11, 67–80.
- Simar, L. and P. Wilson (1999b), Of Course we can bootstrap DEA scores ! But does it mean anything ? Logic trumps wishful thinking, *Journal of Productivity Analysis*, 11, 93–97.
- Simar L. and P. Wilson (2000), A General Methodology for Bootstrapping in Nonparametric Frontier Models, *Journal of Applied Statistics*, Vol 27, 6, 779–802.

- Simar L. and P. Wilson (2001), Testing Restrictions in Nonparametric Efficiency Models, *Communications in Statistics, simulation and computation*, 30 (1), 159–184.
- Simar L. and P. Wilson (2002), Nonparametric Test of Return to Scale, *European Journal of Operational Research*, 139, 115–132.
- 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.
- Simar, L. and P.W. Wilson (2008), Statistical Inference in Nonparametric Frontier Models: recent Developments and Perspectives, in *The Measurement of Productive Efficiency*, 2nd Edition, Harold Fried, C.A.Knox Lovell and Shelton Schmidt, editors, Oxford University Press, 2008.
- Simar, L. and P.W. Wilson (2010), Inference From Cross-Sectional Stochastic Frontier Models. *Econometric Review*, 29, 1, 62–98.
- Simar, L. and P.W. Wilson (2011a), Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis*, 36, 33–53.
- Simar, L. and P.W. Wilson (2011b), Two-Stage DEA: *Caveat Emptor.* Discussion paper 1041, Institut de Statistique, UCL, in press *Journal of Productivity Analysis.*

- Simar, L. and V. Zelenyuk (2006), On Testing Equality of Two Distribution Functions of Efficiency Score Estimated via DEA, *Econometric Review*, 25(4), 497-522.
- Simar, L. and V. Zelenyuk (2007), Statistical Inference for Aggregates of Farrell-type Efficiencies, *Journal of Applied Econometrics*, vol 22, 7, 1367–1394.
- Simar, L. and V. Zelenyuk (2011), Stochastic FDH/DEA estimators for frontier analysis. *Journal of Productivity Analysis*, 36, 1–20..
- Wilson, P.W. (2008), FEAR 1.0: A Software Package for Frontier Efficiency Analysis with R, *Socio-Economic Planning Sciences* 42, 247–254.
- Wilson, P.W. (2011), Asymptotic properties of some non-parametric hyperbolic efficiency estimators, in I. van Keilegom and P. W. Wilson, eds., *Exploring Research Frontiers in Contemporary Statistics and Econometrics*, Berlin: Springer-Verlag.