Boundary estimation in the presence of measurement error with unknown variance

Alois Kneip

University of Bonn

Léopold Simar and Ingrid Van Keilegom

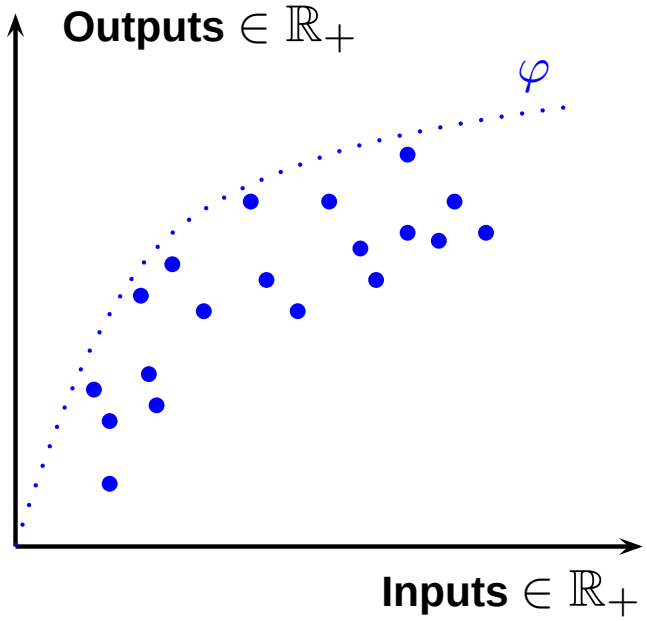
Université catholique de Louvain

Brussels, May 25, 2012

Outline of the talk

1. Introduction and motivation
2. Estimation method
3. Extension to covariates
4. Asymptotic results
5. Simulations
6. Data analysis
7. Conclusions

Introduction and motivation (1/3)



Goal : To estimate the boundary of the support, i.e. the (production) frontier φ

Some examples :

⋆ Family farms :

Input : Number of cows, hectars of land, labor, ...

Output : Liters of milk

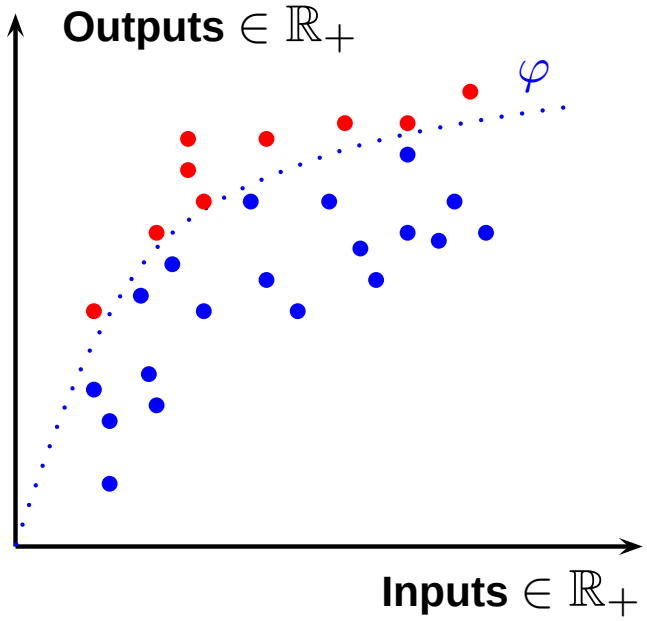
⋆ Productivity of universities :

Input : Human and financial capital

Output : Number of publications, of PhDs, ...

We restrict attention for the moment to the one-dimensional case (i.e. no inputs).

Introduction and motivation (1/3)



Goal : To estimate the boundary of the support, i.e. the (production) frontier φ

Some examples :

⋆ Family farms :

Input : Number of cows, hectars of land, labor, ...

Output : Liters of milk

⋆ Productivity of universities :

Input : Human and financial capital

Output : Number of publications, of PhDs, ...

We restrict attention for the moment to the one-dimensional case (i.e. no inputs).

Introduction and motivation (2/3)

Consider the model

$$Y = X \cdot Z,$$

where Y is observed, $Y \sim g$

X is the true unobserved variable of interest, $X \sim f$

Z is the noise, supposed to be independent of X.

Equivalently, we can also write $Y^\star = X^\star + Z^\star$, where $Y^\star = \log Y$, $X^\star = \log X$ and $Z^\star = \log Z$.

Assume

$$\log Z \sim N(0, \sigma^2),$$

where $\sigma^2 > 0$ is an unknown variance, i.e. Z is a log-normal.

Suppose Support$(X) = [0, \tau]$ with $f(\tau) > 0$ and τ unknown.

Aim of the paper : Estimation of τ (and of σ).

Introduction and motivation (3/3)

Data : $Y_1, \ldots, Y_n \sim Y$ i.i.d.

Literature :

- σ known : extensive literature, see e.g.
  - * Goldenshluger and Tsybakov (2004)
  - * Delaigle and Gijbels (2006)
  - * Meister (2006)
  - * Aarts, Groeneboom and Jongbloed (2007), among many others
- σ unknown : Hall and Simar (2002) :
  - * density of Z unknown but symmetric
  - * $\sigma = \sigma_n \to 0$

Related research : Estimation of f and of σ when f is smooth on $(0, \infty)$

- Butucea and Matias (2005), Butucea, Matias and Pouet (2008)
- Schwarz and Van Bellegem (2009)

Estimation method (1/6)

Note that the density of Z is given by (for $z > 0$)

$$\frac{1}{\sigma z}\phi\Big(\frac{\log z}{\sigma}\Big).$$

A subindex 0 will be added to indicate the true quantities (like $f_0, g_0, \tau_0, ...$).

It can be shown that for all $y > 0$:

$$g_0(y) \quad = \quad \frac{1}{\sigma_0 y}\int_0^1 h_0(t)\phi\left(\frac{1}{\sigma_0}\log\frac{y}{t\tau_0}\right)dt, \tag{1}$$

where $h_0(t) = \tau_0 f_0(t\tau_0)$ for $0 \leq t \leq 1$.

Theorem

There exists a unique $\sigma_0 > 0$, a unique $\tau_0 > 0$ and a unique density h_0 such that (1) holds true, i.e. such that the model is identifiable.

Estimation method (2/6)

Main idea :

Penalized profile likelihood maximization involving unknown functions.

Define

$$g_{h,\tau,\sigma} := \frac{1}{\sigma y} \int_0^1 h(t)\phi\left(\frac{1}{\sigma}\log\frac{y}{t\tau}\right) dt.$$

Obviously, $g_0 = g_{h_0,\tau_0,\sigma_0}$.

Two steps :

Step 1 : Estimation of h for fixed τ and σ

Step 2 : Estimation of τ_0 and σ_0

Estimation method (3/6)

Step 1 : Estimation of h for fixed τ and σ

Fix $\tau > 0$ and $\sigma > 0$ and let M be a natural number. Let

$$\Gamma = \left\{ \gamma = (\gamma_1, \ldots, \gamma_M) : \gamma_k > 0 \text{ for all } k \text{ and } \sum_{k=1}^{M} \gamma_k = M \right\},$$

and define

$$h_\gamma(t) = \gamma_1 I(t = 0) + \sum_{k=1}^{M} \gamma_k I(q_{k-1} < t \leq q_k)$$

for $0 \leq t \leq 1$, where $q_k = k/M$ ($k = 0, 1, \ldots, M$).

It is clear that h_γ is a density for all $\gamma \in \Gamma$. Then,

$$g_{h_\gamma,\tau,\sigma}(y) = \frac{1}{\sigma y} \sum_{k=1}^{M} \gamma_k \int_{q_{k-1}}^{q_k} \phi\left(\frac{1}{\sigma} \log \frac{y}{t\tau}\right) dt.$$

Estimation method (4/6)

The estimator of h is now defined as

$$\widehat{h}_{\tau,\sigma}(\cdot) = h_{\widehat{\gamma}_{\tau,\sigma}}(\cdot),$$

where

$$\widehat{\gamma}_{\tau,\sigma} = \text{argmax}_{\gamma\in\Gamma}\left\{n^{-1}\sum_{i=1}^{n}\log g_{h_\gamma,\tau,\sigma}(Y_i) - \lambda\text{pen}(g_{h_\gamma,\tau,\sigma})\right\},$$

where $\lambda \geq 0$ is a fixed value independent of n, and where

$$\text{pen}(g_{h_\gamma,\tau,\sigma}) = \max_{3\leq j\leq M}|\gamma_j - 2\gamma_{j-1} + \gamma_{j-2}|.$$

Estimation method (5/6)

Step 2 : Estimation of τ_0 and σ_0

Let

$$(\widehat{\tau}, \widehat{\sigma}) = \text{argmax}_{\tau>0,\sigma>0} \left\{ n^{-1} \sum_{i=1}^{n} \log g_{\widehat{h}_{\tau,\sigma},\tau,\sigma}(Y_i) - \lambda \text{pen}(g_{\widehat{h}_{\tau,\sigma},\tau,\sigma}) \right\}.$$

Moreover, $\widehat{h} := \widehat{h}_{\widehat{\tau},\widehat{\sigma}}$ estimates h_0, and $\widehat{g} := g_{\widehat{h},\widehat{\tau},\widehat{\sigma}}$ estimates g_0.

Note 1 :

- The estimation procedure can equivalently be carried out in one step, by maximizing jointly over γ, τ and σ.
- When $M = M_n \to \infty$, we get a sieve maximization procedure.

Estimation method (6/6)

Note 2 :

- λ can be taken equal to 0

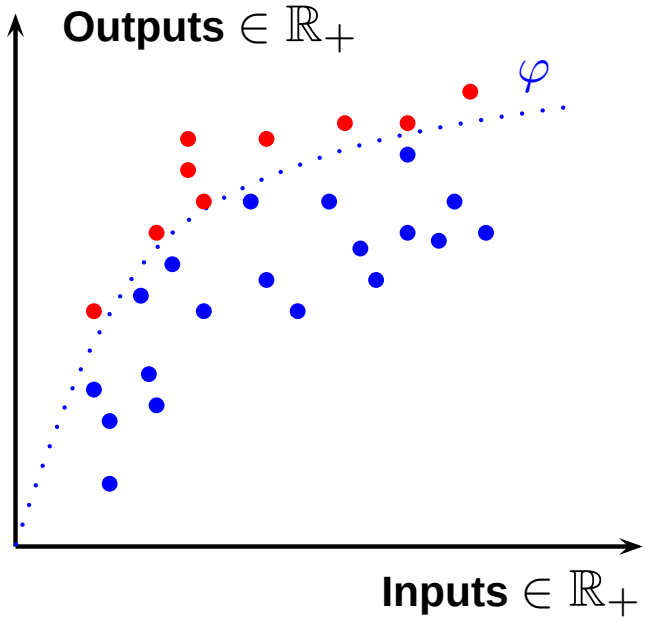
  $\Rightarrow$ Both penalized and non-penalized estimators are considered

  But : penalized estimator attains better rate of convergence.
- λ is chosen independent of n
- Choice of penalty function : motivated by the proof (see later)

Note 3 :

For the asymptotic results we will need to be a bit more precise regarding the space to which σ, τ and h belong (see below).

Extension to covariates (inputs) (1/4)



Goal : To estimate the boundary of the support, i.e. the (production) frontier φ

Some examples :

⋆ Family farms :

Input : Number of cows, hectars of land, labor, ...

Output : Liters of milk

⋆ Productivity of universities :

Input : Human and financial capital

Output : Number of publications, of PhDs, ...

We restrict attention for the moment to the one-dimensional case (i.e. no inputs).

Extension to covariates (2/4)

Consider the model

$$Y = \varphi(W)\exp(-U)\exp(V), \tag{2}$$

where $V \sim N(0, \sigma^2(W))$

$U > 0$ has a jump at the origin and its distribution (possibly) depends on W

U and V are independent given W

only W and Y are observed.

Equivalently, $\log Y = \log \varphi(W) - U + V$.

Note that

- If $\varphi \equiv \tau$ is constant, then the model can be written as $Y = X \cdot Z$, where $X = \tau \exp(-U)$ and $Z = \exp(V)$

  $\Rightarrow$ Model (2) extends our previous model to covariates.
- U represents the inefficiency, V represents the error.

Extension to covariates (3/4)

References :

- Fully parametric approach (φ, f_U and f_V parametric) : many papers; see Greene (2008) for a survey
- Semiparametric approach (φ nonparametric, f_U and f_V parametric) : see e.g. Fan et al (1996), Kumbhakar et al (2007)

Our goal :

φ and f_U nonparametric, f_V normal but with unknown variance.

But :

Dropping parametric assumptions on the distribution of U greatly complicates the problem and enforces to develop completely new methods.

Extension to covariates (4/4)

Let $(W_1, Y_1), \ldots, (W_n, Y_n) \sim (W, Y)$ i.i.d.

Fix w_0 in the support of W and define

$$(\widehat{\tau}(w_0), \widehat{\sigma}(w_0), \widehat{\gamma}(w_0))$$
$$= \text{argmax}_{\tau>0,\sigma>0,\gamma\in\Gamma} \left\{ n_b^{-1} \sum_{i:\|W_i - w_0\|_2 \le b} \log g_{h_\gamma,\tau,\sigma}(Y_i) - \lambda \, \text{pen}(g_{h_\gamma,\tau,\sigma}) \right\},$$

where b is a bandwidth, $n_b := \sum_{i=1}^n I\{\|W_i - w_0\|_2 \le b\}$, and

$$\text{pen}(g_{h_\gamma,\tau,\sigma}) = \max_{3\le j\le M} |\gamma_j - 2\gamma_{j-1} + \gamma_{j-2}|.$$

This ‘local constant’ estimator can be improved to a ‘local linear’ esimator (details omitted).

Asymptotic results (1/5)

Consider the case without covariates, and assume that

(A1) For some $0 < \sigma_l < \sigma_u < \infty$, $0 < \tau_l < \tau_u < \infty$, $0 < h_l < h_u < \infty$, and $0 < \delta < 1$ the estimators $(\widehat{g}, \widehat{\tau}, \widehat{\sigma})$ are determined by minimizing over all

$$(h_\gamma, \tau, \sigma) \in \mathcal{H}_n \times [\tau_l, \tau_u] \times [\sigma_l, \sigma_u],$$

where $\mathcal{H}_n \subset \mathcal{H}_{h_l, h_u, \delta}$, and

$\mathcal{H}_{h_l, h_u, \delta} = \{h | h$ is square integrable density with support $[0, 1]$ satisfying

$$\sup_t h(t) \leq h_u \text{ and } \inf_{1-\delta \leq t \leq 1} h(t) \geq h_l\}.$$

(A2) $h_0 \in \mathcal{H}_{h_l, h_u, \delta}$ and is twice continuously differentiable, $\tau_0 \in [\tau_l, \tau_u]$, and $\sigma_0 \in [\sigma_l, \sigma_u]$.

(A3) For some $0 < \beta < 1/5$, $M = M_n \sim n^\beta$ as n tends to ∞.

Asymptotic results (2/5)

(A4) For some $A > \sqrt{2}$, $P\Big(\log Y < -A(\log n)^{1/2}\sigma_0 \Big) = o(n^{-1})$.

Remark. Note that (A4) is satisfied when e.g. $h_0 \equiv 0$ on $[0, \epsilon]$ for some $\epsilon > 0$.

For two arbitrary densities g_1 and g_2, let

$$H^2(g_1, g_2) = \frac{1}{2} \int \left(\sqrt{g_1(y)} - \sqrt{g_2(y)} \right)^2 dy$$

be the Hellinger distance between g_1 and g_2.

Theorem 1. Assume (A1)-(A4). Then, if $\lambda \geq 0$,

$$H(\widehat{g}, g_0) = O_P(M_n^{-2}),$$

and if $\lambda > 0$,

$$\text{pen}(\widehat{g}) = O_P(M_n^{-2}).$$

Asymptotic results (3/5)

Theorem 2. Assume (A1)-(A4). Then,

a) If $\lambda = 0$ (i.e. without penalization),

$$\widehat{\sigma} - \sigma_0 = O_P\Big((\log n)^{-1}\Big),$$

$$\widehat{\tau} - \tau_0 = O_P\Big((\log n)^{-\frac{1}{2}}\Big).$$

b) If $\lambda > 0$ (i.e. with penalization),

$$\widehat{\sigma} - \sigma_0 = O_P\Big((\log n)^{-2}\Big),$$

$$\widehat{\tau} - \tau_0 = O_P\Big((\log n)^{-\frac{3}{2}}\Big),$$

$$\widehat{h}(1) - h_0(1) = O_P\Big((\log n)^{-1}\Big).$$

Asymptotic results (4/5)

Remark. Instead of using a histogram estimator for h_0, one could use suitable spline estimators to approximate h_0.

We have shown that if h_0 is m-times continuously differentiable for some $m > 2$, then

$$\widehat{\sigma} - \sigma_0 = O_P\left((\log n)^{-(1+\frac{m}{2})}\right),$$

$$\widehat{\tau} - \tau_0 = O_P\left((\log n)^{-\frac{m+1}{2}}\right),$$

$$\widehat{h}(1) - h_0(1) = O_P\left((\log n)^{-\frac{m}{2}}\right).$$

as long as $\widehat{g} = g_{\widehat{h},\widehat{\tau},\widehat{\sigma}}$ (obtained with splines or another approximation method) satisfies

$$H(\widehat{g}, g_0) = O_P(n^{-\kappa}) \quad \text{for some } \kappa > 0.$$

Asymptotic results (5/5)

Some remarks and thoughts

- Extension to covariates
- Choice of M and λ
- The proofs ...
- What changes if error is not normal ?

Simulations (1/11)

Recall that

$$Y = \tau \exp(-U) \exp(V), \text{ where } U > 0 \text{ and } V \sim \mathrm{N}(0, \sigma^2).$$

or equivalently $Y = X \cdot Z$, with $X = \tau \exp(-U)$ and $Z = \exp(V)$.

Suppose that $U \sim \mathrm{Exp}(\beta)$. Then, the density of X can be written as

$$f(x) = \frac{\beta}{\tau^\beta} x^{\beta-1} I(0 \leq x \leq \tau).$$

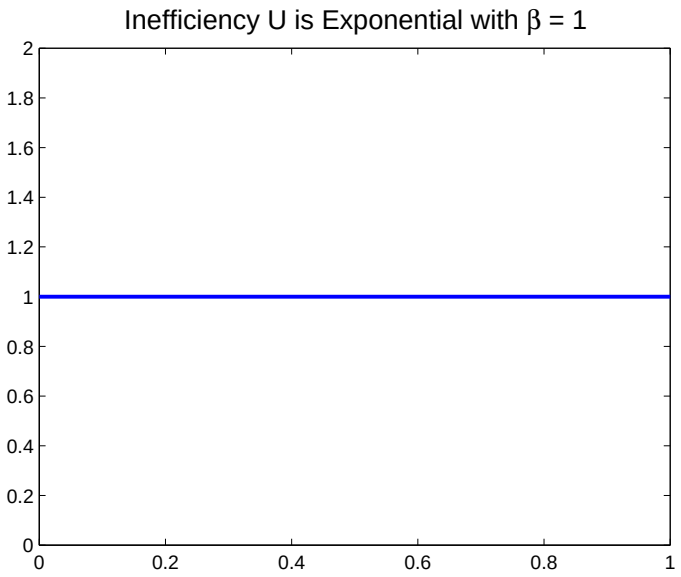
Let

- $\beta = 1$ and $\beta = 2$
- $\tau = 1$
- $\sigma = \sigma_V = \rho \sigma_U$ with $\rho = 0, 0.01, 0.05, 0.25, 0.50, 0.75$.
- $n = 100$

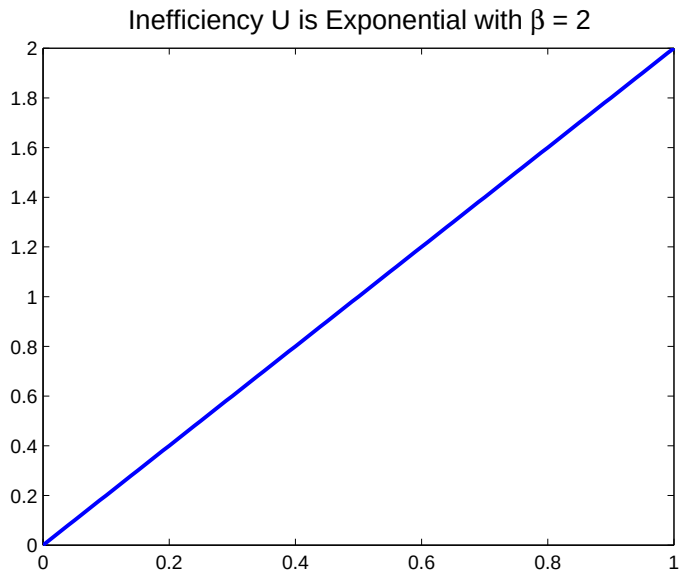
Simulations (2/11)

Density of X when $U \sim \mathrm{Exp}(\beta)$

$\beta = 1$



$\beta = 2$



Simulations (3/11)

Consider

- 500 replications of each experiment
- Choice of λ : minimization of

$$RMSE(\widehat{\tau}) + RMSE(\widehat{\sigma})$$

for $\log_{10} \lambda = -4, -3, -2, -1, 0, 1, 2, 3, 4$

- Choice of M : let

$$M = \max\left(3, c \times \text{round}(n^{1/5})\right)$$

(rule of thumb).

We fix $c = 2$. Very similar results were obtained with $c = 3$ (and even with $c = 1$ but here the number of bins was very small).

For $n = 100$ we have $M = 5$.

Simulations (4/11)

Case 1 : $\beta = 1$

ρ	$\log_{10}\lambda$		$\hat{\tau}$	$\hat{\sigma}$
0	-3	$RMSE$	0.0138	0.39e-04
		$BIAS$	-0.0098	0.13e-04
		STD	0.0098	0.37e-04
0.05	-2	$RMSE$	0.0370	0.0350
		$BIAS$	-0.0067	-0.0121
		STD	0.0365	0.0328
0.25	-1	$RMSE$	0.0988	0.0840
		$BIAS$	-0.0251	0.0182
		STD	0.0956	0.0821
0.75	1	$RMSE$	0.0872	0.1495
		$BIAS$	-0.0460	0.1153
		STD	0.0742	0.0952

Simulations (5/11)

Case 2 : $\beta = 2$

ρ	$\log_{10}\lambda$		$\hat{\tau}$	$\hat{\sigma}$
0	1	$RMSE$	0.0066	0.45e-03
		$BIAS$	-0.0042	0.42e-03
		STD	0.0050	0.17e-03
0.05	-2	$RMSE$	0.0178	0.0190
		$BIAS$	-0.0019	-0.0054
		STD	0.0177	0.0182
0.25	-1	$RMSE$	0.0352	0.0332
		$BIAS$	0.0020	0.0049
		STD	0.0351	0.0329
0.75	-1	$RMSE$	0.0750	0.0544
		$BIAS$	0.0250	-0.0090
		STD	0.0708	0.0537

Simulations (6/11)

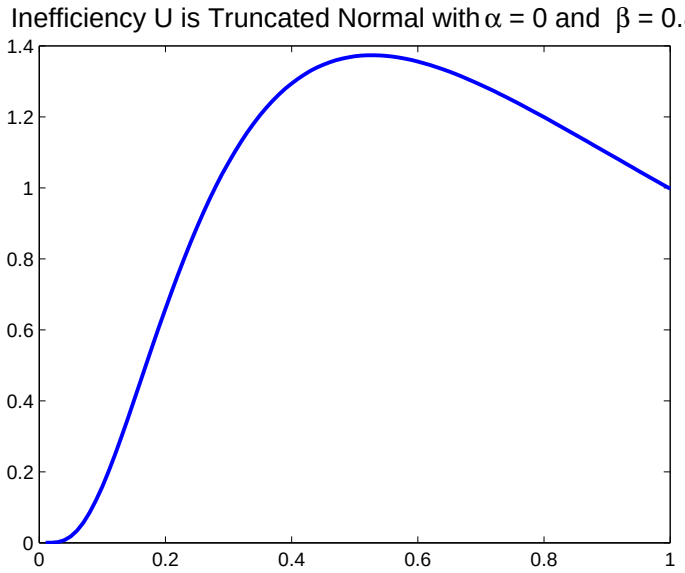
Results of the simulations

- When ρ increases the performance deteriorates.
- Selection of λ : seems not to be crucial. In practice, we suggest to use a bootstrap procedure to estimate the $RMSE$ of the estimators as a tool for selecting λ.
- Performance for $\beta = 2$ is better than for $\beta = 1$.
- Rule of thumb for selecting M (number of bins) seems to work well.
- Simulations not shown here, demonstrate that
  - Performance improves when n increases.
  - Method also works when jump size is rather small (truncated normal).

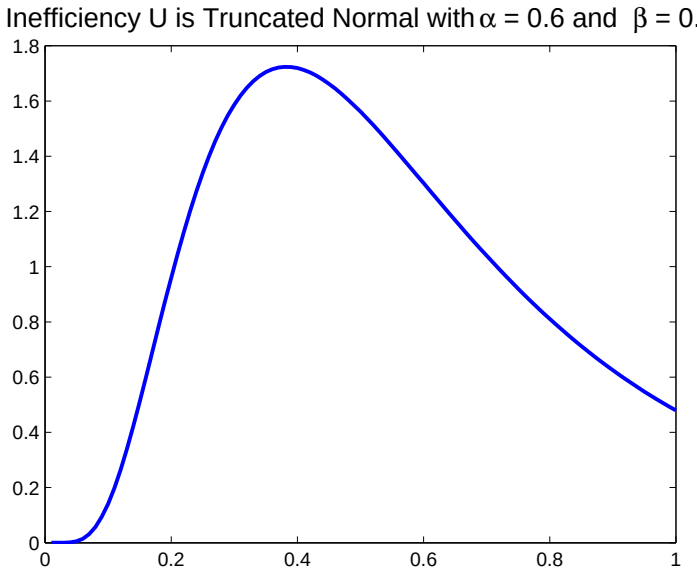
Simulations (7/11)

Density of X when $U \sim N^{+}(\alpha, \beta^2)$

$\alpha = 0, \beta = 0.8$



$\alpha = 0.6, \beta = 0.6$



Simulations (8/11)

Extension to covariates :

Recall the model :

$$Y = \varphi(W)\exp(-U)\exp(V),$$

and consider

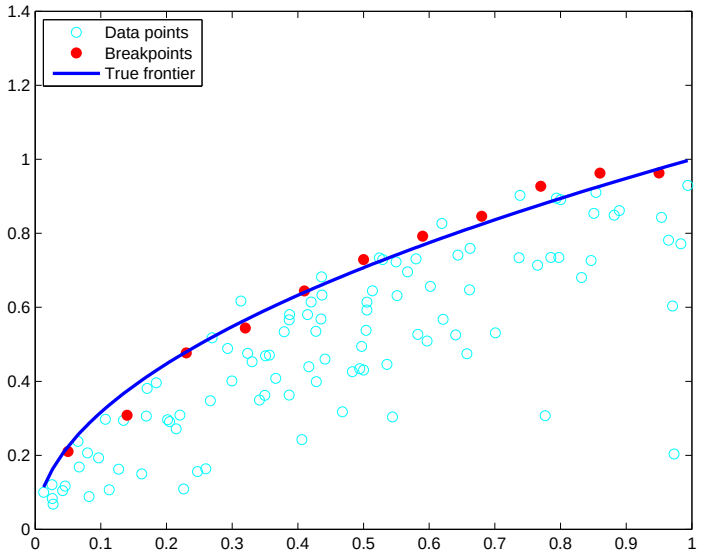
- $W \sim Un[0,1]$, $U \sim$ Exp(3), $V \sim N(0,(.667)^2)$
- $\varphi(w) = w^{1/2}$
- W, U and V are independent
- $w =$ 0.25, 0.50 and 0.75
- $h = 1.06\min(s, r/1.349)n^{-1/5}$ (normal reference rule)
- λ computed from $B = 200$ loops
- 500 Monte-Carlo loops

Then, $\rho = \sigma_V/\sigma_U = 0.20$.

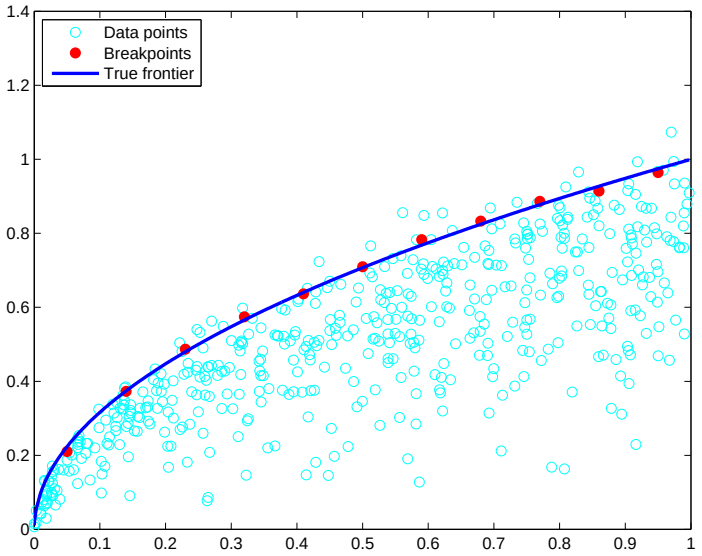
Simulations (9/11)

Frontier estimation for one sample

$n = 100$



$n = 500$



Simulations (10/11)

n				$w = 0.25$	$w = 0.50$	$w = 0.75$
100	$\tau(w)$	new	$BIAS$	.0034	.0046	-.0030
			MSE	.0004	.0008	.0014
		HS	$BIAS$	.0048	.0017	-.0003
			MSE	.0014	.0025	.0028
	$\sigma(w)$	new	$BIAS$	-.0071	-.0031	.0009
			MSE	.0003	.0003	.0003
500	$\tau(w)$	new	$BIAS$	.0083	.0105	.0064
			MSE	.0002	.0004	.0004
		HS	$BIAS$	.0009	.0041	.0041
			MSE	.0005	.0007	.0011
	$\sigma(w)$	new	$BIAS$	-.0053	-.0033	-.0005
			MSE	.0001	.0001	.0001

HS = Hall-Simar (2002)

Simulations (11/11)

Results

- Choice of λ does not seem to be crucial.
- For $\tau(w)$: MSE much better than in HS :
  - For $n = 100$: MSE is 30% of the MSE in HS (50% when $w = 0.75$).
  - For $n = 500$: MSE is less than 50% of the MSE in HS for all cases.
- For $\sigma(w)$: MSE good, but no comparison with HS is possible.

Data analysis (1/3)

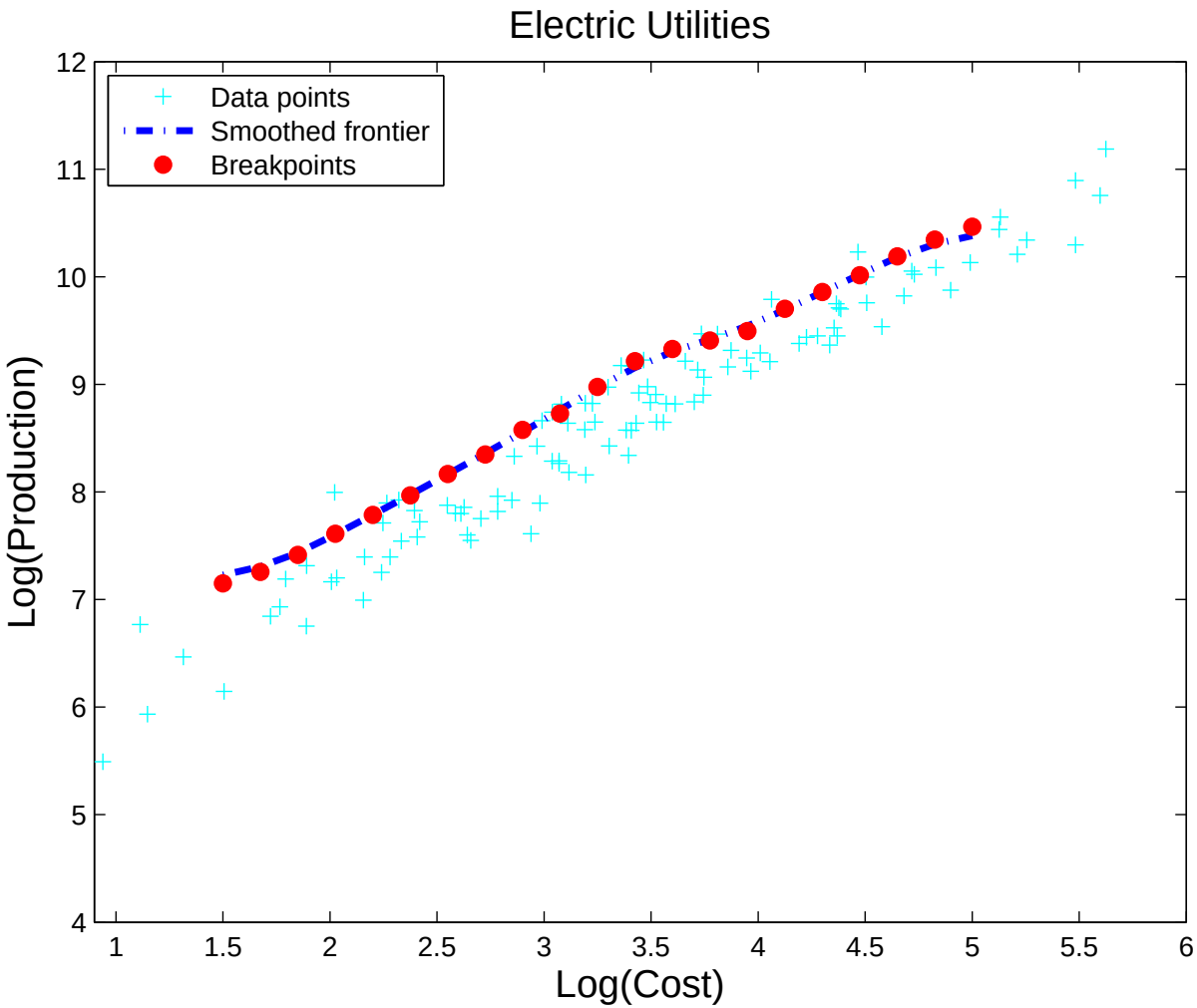
Data from Christensen and Greene (1976) concerning 123 American electricity utility companies.

For each company, we observe

- $Y =$ production output
- $W =$ total cost involved in production

Frontier function = curve of most efficient companies

Data analysis (2/3)



Data analysis (3/3)

values of w	$n_b(w)$	$\widehat{\tau}(w)$	$\widehat{\sigma}(w)$	$\widehat{E}(U\|W=w)$
1.50	10	7.15	0.0342	0.0769
1.85	17	7.41	0.0404	0.0607
2.20	26	7.79	0.0357	0.0463
2.55	28	8.17	0.0257	0.0504
2.90	37	8.58	0.0193	0.0504
3.25	43	8.98	0.0246	0.0464
3.60	38	9.33	0.0226	0.0404
3.95	32	9.49	0.0232	0.0255
4.30	24	9.86	0.0227	0.0278
4.65	21	10.19	0.0201	0.0294
5.00	15	10.47	0.0123	0.0278

Conclusions

- We considered the model $Y = X \cdot Z$, where Y is observed, X is the variable of interest with support $[0, \tau]$ and Z is the noise.
- We supposed that $f_0(\tau_0) > 0$ and that Z is independent of X and is log-normal with unknown variance σ_0^2.
- We showed that the model is identifiable.
- We proposed estimators for τ_0 and σ_0 and proved their consistency and rate of convergence.
- We extended the method to covariates (inputs).
- We showed that the estimators work well for small n, both with and without covariates.
- We applied the method to data on efficiency of electricity companies.