Boundary estimation in the presence of measurement error with unknown distribution

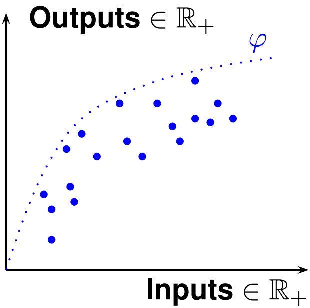
Jean-Pierre Florens Alois Kneip Léopold Simar
Ingrid Van Keilegom

Université catholique de Louvain

August 17-18, 2016

1. Introduction
2. Case 1 : unknown variance
   - Joint work with Alois Kneip and Léopold Simar
   - Published in J. Econometrics (2015)
3. Simulations
4. Case 2 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - Inspired by Delaigle and Hall, 2016
5. Case 3 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - New idea

Consider first the simple case of deterministic frontier models.



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

Some examples :

$\star$ Family farms :

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

Output : Liters of milk

$\star$ Productivity of universities :

Input : Human and financial capital

Output : Number of publications, PhDs, ...

Typically,

- ◇ Input = labor, energy, capital, . . .
- ◇ Output = amount of goods produced

Other areas of application :

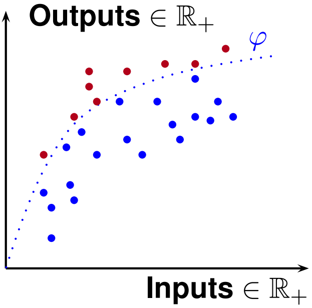
Industry, hospitals, transportation, schools, banks, public services, . . .

Nonparametric estimators of the frontier :

- ◇ DEA (Data Envelopment Analysis) : Farrell (1957)
- ◇ FDH (Free Disposal Hull) : Deprins, Simar, Tulkens (1984)

Output oriented versus input oriented frontiers

We now add some noise to the outputs, i.e. we consider stochastic frontier models.



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

Consider the model

$$Y = X \cdot Z \quad \text{or equivalently} \quad Y^* = X^* + Z^*,$$

where $Y^* = \log Y$, $X^* = \log X$, $Z^* = \log Z$

Y is observed

X is the true unobserved variable of interest

Z is the noise, supposed to be independent of X

the distribution (or variance) of Z is unknown

We suppose that X lives on $(0, \tau]$ (or X^* lives on $(-\infty, \log \tau]$).

Goal : Estimation of τ

Two cases :

- ◇ $Z^* \sim N(0, \sigma^2)$ with σ unknown
- ◇ density of Z^* is symmetric around 0

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 when f is smooth on $(0, \infty)$

- ◇ Butucea and Matias (2005)
- ◇ Meister (2006, 2007)
- ◇ Butucea, Matias and Pouet (2008)
- ◇ Schwarz and Van Bellegem (2010)
- ◇ Delaigle and Hall (2016)

1. Introduction
2. Case 1 : unknown variance
   - Joint work with Alois Kneip and Léopold Simar
   - Published in J. Econometrics (2015)
3. Simulations
4. Case 2 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - Inspired by Delaigle and Hall, 2016
5. Case 3 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - New idea

We suppose for case 1 that

$$\log Z \sim N(0, \sigma^2) \text{ with } \sigma \text{ unknown.}$$

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

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

Let $Y \sim g$ and $X \sim f$. 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 $y > 0$:

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

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

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.

Remark. The proof follows from Schwarz and Van Bellegem (2010), who prove the identifiability for any P_X belonging to

$$\{P \in \mathcal{P} | \exists A \in \mathcal{B}(R) : |A| > 0 \text{ and } P(A) = 0\},$$

where $\mathcal{B}(R)$ = set of Borel sets in R
$\mathcal{P}$ = set of all probability distributions on R
$|A|$ = Lebesgue measure of A.

Other error densities that allow to identify the model :

- Cauchy
- stable, ...

(see Schwarz and Van Bellegem, 2010).

We use penalized profile likelihood maximization to estimate τ :

Define

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

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

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

for some $M < \infty$, 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 $q_k = k/M$ $(k = 0, 1, \ldots, M)$. Then,

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

Let

$$(\widehat{\tau}, \widehat{\sigma}, \widehat{\gamma}) = \text{argmax}_{\tau>0,\sigma>0,\gamma\in\Gamma}\Big\{n^{-1}\sum_{i=1}^{n}\log g_{h_\gamma,\tau,\sigma}(Y_i) - \lambda \text{pen}(g_{h_\gamma,\tau,\sigma})\Big\},$$

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

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

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

Note :

- 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

Asymptotic results

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

$$\mathcal{H}_{h_l, h_u, \delta} = \{h | h \text{ is square integrable density with support } [0,1] \text{ satisfying } \sup_t h(t) \leq h_u \text{ and } \inf_{1-\delta \leq t \leq 1} h(t) \geq h_l\}.$$

(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 ∞.

(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) holds if e.g. $h_0 \equiv 0$ on $[0, \epsilon]$ for some $\epsilon > 0$.

For two arbitrary densities g_1 and g_2, let

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

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 $\lambda > 0$,

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

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),

$$\widehat{\sigma} - \sigma_0 = O_P\Big((\log n)^{-2}\Big),$$
$$\widehat{\tau} - \tau_0 = O_P\Big((\log n)^{-\frac{3}{2}}\Big),$$
$$\widehat{h}(1) - h_0(1) = O_P\Big((\log n)^{-1}\Big).$$

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

$$\begin{aligned}
\widehat{\sigma} - \sigma_0 &= O_P\left((\log n)^{-(1+\frac{m}{2})}\right), \\
\widehat{\tau} - \tau_0 &= O_P\left((\log n)^{-\frac{m+1}{2}}\right), \\
\widehat{h}(1) - h_0(1) &= O_P\left((\log n)^{-\frac{m}{2}}\right).
\end{aligned}$$

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

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

Extension to covariates (inputs)

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

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.

References :

- ◇ Fully parametric approach (φ, f_U and f_V parametric) : many papers; see Greene (2008) for a survey
- ◇ Semiparametric approach (φ nonpar., f_U and f_V param.) : 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.

Suppose that $\dim(W) = d$.

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

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

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

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

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

Recall that $Y = X \cdot Z$, or equivalently,

$$\begin{aligned} \log Y &= \log X + \log Z \\ &= \log \tau - U + \log Z, \end{aligned}$$

where $U > 0$ and $\log Z \sim N(0, \sigma^2)$.

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

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

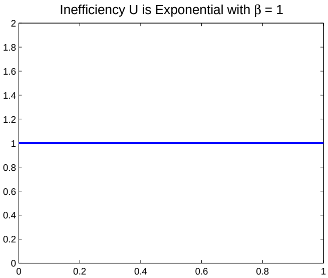
Let

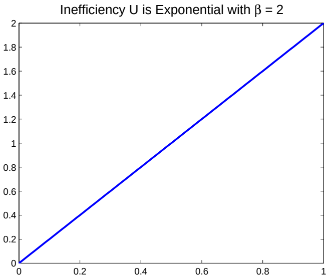
- $\beta = 1$ and $\beta = 2$
- $\tau = 1$
- $\sigma = \sigma_{\log Z} = \rho\sigma_U$ with $\rho = 0, 0.05, 0.25, 0.75$.
- $n = 100$

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

$\beta = 1$

$\beta = 2$





Consider

- $\diamond$ 500 replications of each experiment
- $\diamond$ Choice of λ : minimization of

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

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

- $\diamond$ 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 $M = 5$.

Case 1 : $\beta = 1$

ρ	$\log_{10} \lambda$		$\widehat{\tau}$	$\widehat{\sigma}$
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

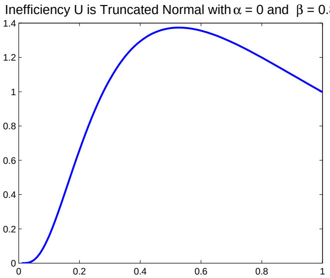
Case 2 : $\beta = 2$

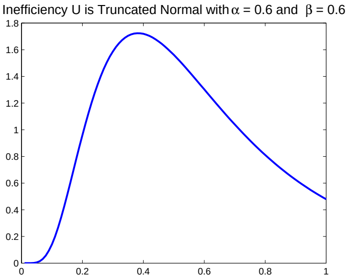
ρ	$\log_{10} \lambda$		$\widehat{\tau}$	$\widehat{\sigma}$
0	1	*RMSE*	0.0066	0.45e-03
		BIAS	-0.0042	0.42e-03
		STD	0.0050	0.17e-03
0.05	-2	*RMSE*	0.0178	0.0190
		BIAS	-0.0019	-0.0054
		STD	0.0177	0.0182
0.25	-1	*RMSE*	0.0352	0.0332
		BIAS	0.0020	0.0049
		STD	0.0351	0.0329
0.75	-1	*RMSE*	0.0750	0.0544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

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

$\alpha = 0, \beta = 0.8$

$\alpha = 0.6, \beta = 0.6$





Case 1 : $\alpha = 0, \beta = 0.8$

ρ	$\log_{10}\lambda$		$\widehat{\tau}$	$\widehat{\sigma}$
0	-4	*RMSE*	0.0127	0.29e-04
		BIAS	-0.0090	0.14e-04
		STD	0.0090	0.26e-04
0.05	-2	*RMSE*	0.0441	0.0385
		BIAS	-0.0227	0.0083
		STD	0.0378	0.0376
0.25	-2	*RMSE*	0.0999	0.0672
		BIAS	-0.0436	0.0126
		STD	0.0900	0.0661
0.75	1	*RMSE*	0.0777	0.0716
		BIAS	-0.0529	0.0555
		STD	0.0570	0.0452

Case 2 : $\alpha = 0.6, \beta = 0.6$

ρ	$\log_{10} \lambda$		$\widehat{\tau}$	$\widehat{\sigma}$
0	-2	*RMSE*	0.0255	0.13e-03
		BIAS	-0.0169	0.50e-04
		STD	0.0191	0.12e-03
0.05	-4	*RMSE*	0.1038	0.0776
		BIAS	-0.0672	0.0428
		STD	0.0792	0.0649
0.25	-3	*RMSE*	0.1483	0.0894
		BIAS	-0.0959	0.0301
		STD	0.1132	0.0843
0.75	2	*RMSE*	0.1812	0.0876
		BIAS	-0.1226	0.0711
		STD	0.1336	0.0512

Robustness to Gaussian assumption :

We now compare our method with the one by Hall and Simar (2002, JASA), who assumed that

- $\diamond$ density of $\log Z$ unknown but symmetric
- $\diamond$ $\sigma = \sigma_n \to 0$

Consider (as before) the case where $U \sim N^+(0, 0.8^2)$.

Consider the same model settings as before, except that

$$\log Z \sim C_1 t_4 \quad \text{or} \quad \log Z \sim C_2 \text{ Laplace},$$

where the scaling factors C_1 and C_2 are chosen to obtain the same noise to signal ratios as in the preceding simulation.

Case 1 : $\log Z \sim C_1 t_4$

ρ		$\widehat{\tau}$	$\widehat{\tau}_{HS}$	$\widehat{\sigma}$
0	*RMSE*	0.0129	0.0427	0.38e-03
	BIAS	-0.0088	-0.0087	0.30e-03
	STD	0.0095	0.0418	0.24e-03
0.05	*RMSE*	0.0415	0.0431	0.0363
	BIAS	-0.0210	-0.0085	0.0070
	STD	0.0359	0.0423	0.0356
0.25	*RMSE*	0.0963	0.0600	0.0788
	BIAS	-0.0492	-0.0075	0.0155
	STD	0.0829	0.0596	0.0774
0.75	*RMSE*	0.0767	3.0828	0.1027
	BIAS	-0.0550	0.6044	0.0552
	STD	0.0535	3.0260	0.0867

Case 2 : $\log Z \sim C_2$ Laplace

ρ		$\widehat{\tau}$	$\widehat{\tau}_{HS}$	$\widehat{\sigma}$
0	*RMSE*	0.0128	0.0394	0.12e-03
	BIAS	-0.0086	-0.0062	0.48e-04
	STD	0.0095	0.0389	0.11e-03
0.05	*RMSE*	0.0434	0.0431	0.0361
	BIAS	-0.0235	-0.0085	0.0073
	STD	0.0365	0.0423	0.0354
0.25	*RMSE*	0.0959	0.0598	0.0703
	BIAS	-0.0481	-0.0045	0.0142
	STD	0.0830	0.0597	0.0689
0.75	*RMSE*	0.0777	0.4797	0.0818
	BIAS	-0.0564	0.1189	0.0542
	STD	0.0535	0.4652	0.0614

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 $f(\tau) > 0$ and that Z is independent of X and is log-normal with unknown variance σ^2.
- We showed that the model is identifiable.
- We proposed a profile likelihood estimator for τ and σ and proved their consistency and rate of convergence.
- We showed that the estimators work well for small n.

1. Introduction
2. Case 1 : unknown variance
   - Joint work with Alois Kneip and Léopold Simar
   - Published in J. Econometrics (2015)
3. Simulations
4. Case 2 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - Inspired by Delaigle and Hall, 2016
5. Case 3 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - New idea

Consider the model

$$Y = X + \varepsilon = \tau + Z + \varepsilon, \text{ where } Z \geq 0, X \perp\!\!\!\perp \varepsilon,$$

and we assume now that

the density of ε is symmetric around 0, but otherwise unknown.

Note that X lives on $[\tau, \infty)$.

As in Delaigle and Hall (2016) we assume that X is non-decomposable, i.e. it is not possible to write X as

$$X = X_1 + X_2,$$

with $X_1 \perp\!\!\!\perp X_2$
$X_1 \geq \tau_1$ for some τ_1
X_2 is symmetric around 0.

This assumption is necessary to make the model identifiable.

Let

$$\psi_Y(t) = E\{\exp(itY)\}$$

be the characteristic function of Y. We propose to estimate τ by minimizing the following distance between two estimators of $\psi_Y(t)$:

$$\int w(t)\left|\widehat{\psi}_{NP}(t) - \widehat{\psi}_{\tau}(t)\right|^2 dt,$$

where $\widehat{\psi}_{NP}(t) = n^{-1}\sum_{j=1}^{n}\exp(itY_j) =$ nonparametric estimator

$\widehat{\psi}_{\tau}(t) =$ certain estimator depending on τ

$w(t) =$ certain weight function

Note that

$$\psi_Y(t) = |\psi_Y(t)| P_Y(t),$$

where $|\psi_Y(t)|$ is the modulus of Y
$P_Y(t) = \psi_Y(t)/|\psi_Y(t)|$ is the phase function of Y.

An estimator of $|\psi_Y(t)|$ is given by $|\widehat{\psi}_{NP}(t)|$.

For $P_Y(t)$, note that (since X and ε are independent)

$$P_Y(t) = P_X(t) P_\varepsilon(t) = P_X(t)$$

since ε is symmetric around 0 and hence $\psi_\varepsilon(t)$ is real.

Moreover,

$$\begin{aligned} P_X(t) &= \exp(it\tau) P_Z(t) \\ &= \exp(it\tau) \frac{E\{\exp(itZ)\}}{|E\{\exp(itZ)\}|}. \end{aligned}$$

$E\{\exp(itZ)\}$ can be estimated by

$$\sum_{k=1}^{m} p_k \exp(itz_k),$$

where $z_1, \ldots, z_m =$ fixed grid of points in the support of Z
$p_1, \ldots, p_m =$ parameters satisfying $p_k \geq 0$ and $\sum_{k=1}^{m} p_k = 1$
$m = 5\, n^{1/2}$

(see also Delaigle and Hall, 2016). Hence,

$$\widehat{\psi}_{\tau,p}(t) = |\widehat{\psi}_{NP}(t)| \exp(it\tau) \frac{\sum_{k=1}^{m} p_k \exp(itz_k)}{|\sum_{k=1}^{m} p_k \exp(itz_k)|}.$$

We now define

$$(\widehat{\tau}, \widehat{p}_1, \ldots, \widehat{p}_m) = \text{argmin}_{\tau, p_1, \ldots, p_m} \int w(t) \left| \widehat{\psi}_{NP}(t) - \widehat{\psi}_{\tau,p}(t) \right|^2 dt,$$

under the constraint of maximizing τ.

Asymptotic properties and simulations for these estimators : work in progress...

1. Introduction
2. Case 1 : unknown variance
   - Joint work with Alois Kneip and Léopold Simar
   - Published in J. Econometrics (2015)
3. Simulations
4. Case 2 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - Inspired by Delaigle and Hall, 2016
5. Case 3 : unknown distribution
   - Joint work with Jean-Pierre Florens and Léopold Simar
   - Work in progress
   - New idea

Write

$$P_Y(t) = \exp(i\theta_Y(t)).$$

Instead of comparing two estimators of $\psi_Y(t)$, we now compare two estimators of $\theta_Y(t)$. Note that

$$P_Y(t) = \cos(\theta_Y(t)) + i\sin(\theta_Y(t)),$$

and hence

$$\theta_Y(t) = \arctan\frac{Im(P_Y(t))}{Re(P_Y(t))},$$

which can be estimated in a nonparametric way by

$$\widehat{\theta}_{NP}(t) = \arctan\frac{Im(\widehat{P}_{NP}(t))}{Re(\widehat{P}_{NP}(t))}.$$

with $\widehat{P}_{NP}(t) = \frac{n^{-1}\sum_{j=1}^{n}\exp(itY_j)}{|n^{-1}\sum_{j=1}^{n}\exp(itY_j)|}$.

Next, note that

$$\theta_Y(t) = \theta_X(t) = t\tau + \theta_Z(t),$$

since $P_Y(t) = P_X(t)$ and since $X = \tau + Z$. it can be shown that

$$\theta_Z(t) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1} t^{2k-1} \kappa_{2k-1}}{(2k-1)!},$$

where κ_k is the k-th cumulant of the distribution of Z (see Delaigle and Hall, 2016).

Next, since $Z \geq 0$, we can write

$$Z = -\log \widetilde{Z},$$

where the support of $\widetilde{Z}$ is $(0, 1]$.

We know that

$$F_Z(z) = P(Z \le z) = P(\widetilde{Z} \ge e^{-z}) = 1 - F_{\widetilde{Z}}(e^{-z}),$$

and hence

$$f_Z(z) = f_{\widetilde{Z}}(e^{-z})e^{-z}.$$

We approximate $f_{\widetilde{Z}}(z)$ by a histogram estimator :

$$f_{\widetilde{Z}}(z; \alpha) = \sum_{j=1}^{M_n} \alpha_j I(q_{j-1} < z \le q_j),$$

where $q_j = j/M_n$ ($j = 0, \ldots, M_n \to \infty$) and $\sum_{j=1}^{M_n} \alpha_j = M_n$.

Note that Z should have positive mass near 0
$\Rightarrow$ $\widetilde{Z}$ should have positive mass near 1
$\Rightarrow$ We impose that $\alpha_{M_n} > \epsilon$ for some small $\epsilon > 0$

Now, for any $k \geq 1$,

$$\mu_{Zk} = \int_0^\infty z^k f_Z(z) dz = \int_0^\infty z^k f_{\widetilde{Z}}(e^{-z}) e^{-z} dz$$

which can be approximated by

$$\int_0^\infty z^k f_{\widetilde{Z}}(e^{-z}; \alpha) e^{-z} dz = \sum_{j=1}^{M_n} \alpha_j \int_{-\log q_j}^{-\log q_{j-1}} z^k e^{-z} dz,$$

which is a known function of $\alpha_1, \ldots, \alpha_{M_n}$. Hence, we also have that

$$\begin{aligned} \kappa_k &= \text{known function of } \mu_{Z1}, \ldots, \mu_{Zk} \\ &= \text{known function of } \alpha_1, \ldots, \alpha_{M_n} \end{aligned}$$

We conclude that $\theta_X(t)$ can be approximated by

$$\widehat{\theta}_{\tau,\alpha}(t) = t\tau + \sum_{k=1}^{K_n} h_k(\alpha_1, \ldots, \alpha_{M_n}) t^{2k-1},$$

where $K_n \to \infty$ and $h_k(\alpha_1, \ldots, \alpha_{M_n})$ is a known function of $\alpha_1, \ldots, \alpha_{M_n}$.

Finally, we define the following estimators of τ and $\alpha_1, \ldots, \alpha_{M_n}$:

$$
\begin{aligned}
&(\widehat{\tau}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_{M_n}) \\
&= \operatorname{argmin}_{\substack{\alpha_1, \ldots, \alpha_{M_n} > 0 \\ \alpha_{M_n} > \epsilon, \tau \in \mathbb{R}}} \left\{ \int w(t) \Big| \tan \widehat{\theta}_{NP}(t) - \tan \widehat{\theta}_{\tau,\alpha}(t) \Big|^2 dt \right. \\
&\qquad\qquad\qquad\qquad \left. + \lambda \max_{3 \le j \le M_n} |\alpha_j - 2\alpha_{j-1} + \alpha_{j-2}| \right\},
\end{aligned}
$$

under the constraint of maximizing τ, where $\lambda \geq 0$ is a smoothness penalty parameter. Or alternatively, we could also define

$$
\begin{aligned}
&(\widehat{\tau}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_{M_n})_{alt} \\
&= \operatorname{argmin}_{\substack{\alpha_1, \ldots, \alpha_{M_n} > 0 \\ \alpha_{M_n} > \epsilon, \tau \in \mathbb{R}}} \left\{ \int w(t) \Big| \exp(i\widehat{\theta}_{NP}(t)) - \exp(i\widehat{\theta}_{\tau,\alpha}(t)) \Big|^2 dt \right. \\
&\qquad\qquad\qquad\qquad \left. + \lambda \max_{3 \le j \le M_n} |\alpha_j - 2\alpha_{j-1} + \alpha_{j-2}| \right\}.
\end{aligned}
$$

Based on the $\widehat{\alpha}_j$'s, we can estimate the density of Z and then also the density of X using in addition $\widehat{\tau}$.

Asymptotic properties and simulations for these estimators : work in progress...

Conclusions

- We considered the model $Y = X + \varepsilon = \tau + Z + \varepsilon$, where $Z \geq 0, X \perp\!\!\!\perp \varepsilon$, and the density of ε is symmetric around 0, but otherwise unknown.
- To assure identifiability, we assumed that X is non-decomposable.
- We proposed two minimum distance estimators for τ (and the density of X), one based on the characteristic function of Y, and one based on the angle of the phase function of Y.

Work in progress

- Asymptotic theory
- Simulations and data analysis