Significance testing in nonparametric regression

Samuel Maistre,
supervised by V. Patilea & P. Lavergne

Crest-Ensai & Irmar

Jeunes probabilistes et statisticiens,
April 17, 2012

Outline

Outline

1. Introduction
   - Context
   - State of the art
2. Test statistic
   - The statistic
   - Power Against Local Alternatives
3. Monte-carlo study

Hypothesis tested

- $Y \in \mathbb{R}$,
- $W \in \mathbb{R}^p$ with density f,
- $X \in \mathbb{R}^q$ and the density of (W, X) is f_2,
- We want to test

$$\begin{aligned} & H_0 : E[Y \mid W, X] = E[Y \mid W] \quad a.s. \\ \text{versus} \quad & H_1 : \Pr\{E[Y \mid W, X] = E[Y \mid W]\} < 1. \end{aligned}$$

- Let $r(w) = E[Y \mid W = w]$ and $u = Y - r(W)$, then

$$H_0 \; : \; E[u \mid W, X] = 0 \quad a.s.$$

Fan et Li (1996), Lavergne et Vuong (2000)

- Estimate $E\left[u\,E\left[u \mid W, X\right] f^2(W)\, f_2(W, X)\right]$ by
  - $I_n^{FL} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \widehat{u_i f_i}\, \widehat{u_j f_j} K_{nij}$
    $= \frac{1}{n(n-1)^3} \sum_{i=1}^{n} \sum_{\substack{j \neq i \\ k \neq i \\ l \neq j}} (Y_i - Y_k)(Y_j - Y_l) L_{nik} L_{njl} K_{nij}$
  - $I_n^{LV} = \frac{1}{n^{(4)}} \sum_{a} (Y_i - Y_k)(Y_j - Y_l) L_{nik} L_{njl} K_{nij}$
- $L_{nik} = \frac{1}{g^p} L\left(\frac{W_i - W_k}{g}\right); \quad K_{nij} = \frac{1}{h^{p+q}} K\left(\frac{W_i - W_j}{h}, \frac{X_i - X_j}{h}\right).$

Delgado et Gonzalez Manteiga (2001)

- Estimate

$$T(w,x) = E\left[u\,f(W)\,\mathbf{1}_{w,x}(W,X)\right]$$

where

$$\mathbf{1}_{w,x}(W,X) = \mathbf{1}\{W \geq w, X \geq x\}$$

by

$$\begin{aligned} T_n(w,x) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{u_i f_i}\,\mathbf{1}_{w,x}(W_i,X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (Y_i - Y_j)\,L_{nij}\,\mathbf{1}_{w,x}(W_i,X_i) \end{aligned}$$

- Then use

$$I_n^{DGM} = \sum_{k=1}^{n} T_n^2(W_k)\,.$$

Outline

1. Introduction
   - Context
   - State of the art

2. Test statistic
   - The statistic
   - Power Against Local Alternatives

3. Monte-carlo study

Fundamental lemma (1/2)

Lemma

Let (W_1, X_1, Y_1) and (W_2, X_2, Y_2) be two independent draws of (W, X, Y) and

$$I(h) := \mathbb{E}\left[u_1 u_2 \omega(W_1)\omega(W_2) h^{-p} K((W_1 - W_2)/h)\psi(X_1 - X_2)\right]$$

where $h \in \mathbb{R}_+^*$, ω, K and ψ are functions such that $\omega, \mathcal{F}[K], \mathcal{F}[\psi] > 0$, then $\forall h \in \mathbb{R}_+^*$,

$$\mathbb{E}[Y \mid W, X] = \mathbb{E}[Y \mid W] \quad p.s. \quad \Leftrightarrow \quad I(h) = 0.$$

Fundamental lemma (2/2)

Proof.

$$
\begin{aligned}
I(h) &= \mathbb{E}\left[u_1 u_2 \omega(W_1)\omega(W_2)\int_{\mathbb{R}^p} e^{-it'(W_1-W_2)}\mathcal{F}[K](th)\,dt \right.\\
&\qquad \left. \times \int_{\mathbb{R}^q} e^{-iu'(X_1-X_2)}\mathcal{F}[\psi](u)\,du\right]\\
&= \int_{\mathbb{R}^q}\int_{\mathbb{R}^p}\left|\mathbb{E}\left[\mathbb{E}[u \mid W, X]\,\omega(W)\,e^{-i\{t'W+u'X\}}\right]\right|^2\\
&\qquad \times \mathcal{F}[K](th)\,\mathcal{F}[\phi](u)\,dtdu.
\end{aligned}
$$

□

The test statistic

- We take $\omega(W) = f(W)$ and we estimate $I(h)$ by

$$I_n = \frac{1}{n^{(4)}} \sum_a (Y_i - Y_k)(Y_j - Y_l) L_{nik} L_{njl} K_{nij} \psi_{ij}$$

- $L_{nik} = \frac{1}{g^p} L\left(\frac{W_i - W_k}{g}\right)$,
- $K_{nij} = \frac{1}{h^p} K\left(\frac{W_i - W_j}{h}\right)$,
- $\psi_{ij} = \psi(X_i - X_j)$.

Consistency theorem

- Power against H_{1n} : $E[Y \mid W, X] = E[Y \mid W] + \delta_n d(W, X)$ when $\delta_n \to 0$

Theorem

Under some regularity conditions and if

- $f(\cdot)$ *and* $r(\cdot) f(\cdot)$ *are of regularity* $s \geq 2$;
- $ng^p \to \infty$, $nh^p \to \infty$, $h/g \to 0$ *and* $nh^{p/2} g^{2s} \to 0$;

then

$$\text{(i)} \quad nh^{p/2} I_n \to \mathcal{N}\left(C\mu, \omega^2\right) \text{ if } \delta_n^2 nh^{p/2} \to C$$

$$\text{(ii)} \quad nh^{p/2} I_n \to \infty \text{ if } \delta_n^2 nh^{p/2} \to \infty.$$

Proof insights

- $$\mathbb{E}\left[\left(nh^{p/2}I_n - \frac{h^{p/2}}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} u_i u_j f_i f_j K_{nij}\psi_{ij}\right)^2\right] = O\left(n^{-1}g^{-p}\right) + O\left(n^{-1}h^{-p}\right) + O\left(h^p/g^p\right) + \ldots$$
- $\frac{h^{p/2}}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} u_i u_j f_i f_j K_{nij}\psi_{ij}$ is a U-statistic with
  - $\mathbb{E}\left[\frac{h^{p/2}}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} u_i u_j f_i f_j K_{nij}\psi_{ij}\right] \simeq nh^{p/2}\delta_n^2\mu$
  - $\text{Var}\left[\frac{h^{p/2}}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} u_i u_j f_i f_j K_{nij}\psi_{ij}\right] \to \omega^2$

Variance Estimation

- $\omega^2 = 2\mathbb{E}\left[\sigma^2(W,X_1)\,\sigma^2(W,X_2)\,f^4(W)\,\psi(X_1-X_2)\right] \times \int K^2(u)\,du,$
- $\widehat{\omega^2_{FL}} = \frac{2h^p}{n^{(2)}} \sum_{i=1}^{n} \sum_{j\neq i} \widehat{u_i f_i}^2 \widehat{u_j f_j}^2 K^2_{nij} \psi^2_{nij},$
- $\widehat{\omega^2_{LV}} = \frac{2h^p}{n^{(6)}} \sum_{a} (Y_i - Y_k)(Y_i - Y_{k'})(Y_i - Y_k)(Y_i - Y_{k'})$

  $\times L_{nik} L_{nik'} L_{njl} L_{njl'} K^2_{nij} \psi^2_{nij}.$

Comparison with other methods

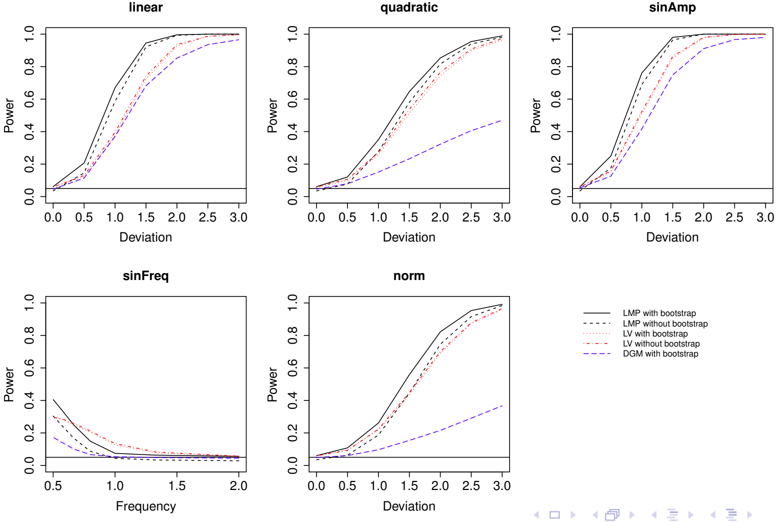
	Lavergne et Vuong	Delgado et Gonzalez Manteiga	Our statistic
Rate of convergence	$\sqrt{n}h^{(p+q)/2}$	$\sqrt{n}$	$\sqrt{n}h^{p/2}$
Known asymptotic distribution	Yes	No	Yes

A word on bootstrap

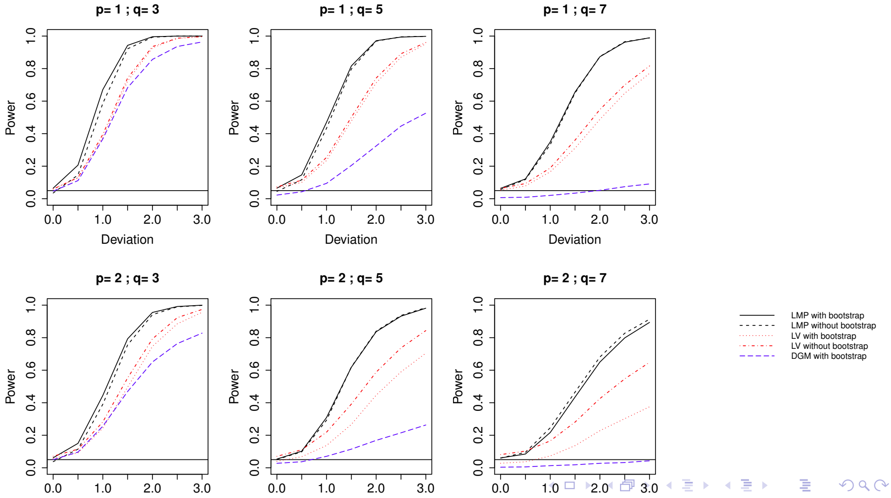
- Use wild bootstrap :
  - $Y_i^* = \hat{r}_i + \varepsilon_i^*$;
  - $\varepsilon_i^* = \hat{\varepsilon}_i \eta_i$;
  - $\mathbb{E}[\eta_i] = 0$, $\mathbb{E}\left[\eta_i^2\right] = 1$, $\mathbb{E}\left[\eta_i^3\right] = 1$ (two points distribution like Haerdle et Mammen, 1993).
- Shown to be consistent for Fan et Li's statistic by Gu et al. (2007).
- To be shown for Lavergne et Vuong's statistic and ours ...
- Non natural assumption of f bounded away from 0 :
  - implies bounded support for W and strictly positive density on the support ;
  - satisfied by no standard distributions but the uniform.

Outline

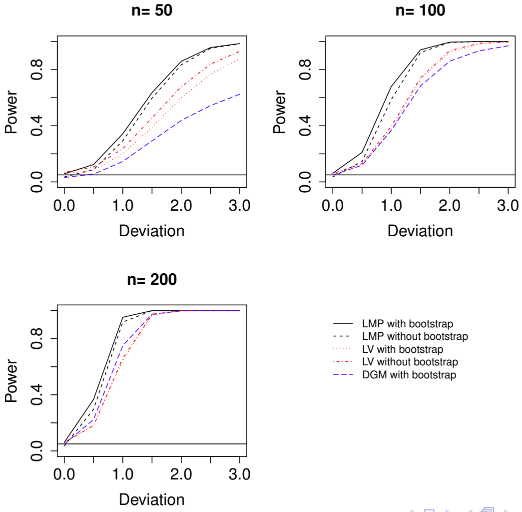
Various Alternatives



Various Dimensions



Various sizes



To be continued

- Proove the consistency of the bootstrap,
- Same approach for generalized linear modeling,
- Same approach for single-index modeling,
- Adaptative bandwidth.

Bibliography : significance testing

Yanqin Fan and Qi Li.
Consistent model specification tests: Omitted variables and semiparametric functional forms.
Econometrica, 64(4):865–890, July 1996.

Pascal Lavergne and Quang Vuong.
Nonparametric significance testing.
Econometric Theory, 16(4):576–601, August 2000.

Miguel A. Delgado, and Wenceslao González Manteiga.
Significance testing in nonparametric regression based on the bootstrap.
Annals of Statistics, 29(5):1469–1507, 2001.