Pointwise estimation of the density of regression errors by model selection

Sandra Plancade

MAP5 Université Paris 5

27 août 2008

Framework

- Let (X_i, Y_i) be a sample from the regression framework :

$$Y_i = b(X_i) + \epsilon_i$$

with

- The (X_i) i.i.d. variables from density μ supported on $[0, 1]$. Moreover, μ is lower bounded by $m_0 > 0$ and upper bounded by $m_1 < \infty$.
- The (ϵ_i) are i.i.d. variables from density f supported on $\mathbb{R}$, indépendent from the (X_i), with $\mathbb{E}[\epsilon_i] = 0$ and upper bounded by $\nu < \infty$.

Framework

- Let (X_i, Y_i) be a sample from the regression framework :

$$Y_i = b(X_i) + \epsilon_i$$

  with

  - The (X_i) i.i.d. variables from density μ supported on $[0, 1]$. Moreover, μ is lower bounded by $m_0 > 0$ and upper bounded by $m_1 < \infty$.
  - The (ϵ_i) are i.i.d. variables from density f supported on $\mathbb{R}$, indépendent from the (X_i), with $\mathbb{E}[\epsilon_i] = 0$ and upper bounded by $\nu < \infty$.

- This lecture propose an estimator $\tilde{f}$ of f adapted to the pointwise risk :

$$\mathbb{E}[(\tilde{f} - f)^2(x_0)]$$

  where x_0 is a fixed point in $\mathbb{R}$.

Principle of error estimation

- The (ϵ_i) are unobserved, so we construct proxies. More precisely, we observe a $2n$-sample $(X_i, Y_i)_{i=-n,\dots,n}$ that we split into two independent samples :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

Principle of error estimation

- The (ϵ_i) are unobserved, so we construct proxies. More precisely, we observe a $2n$-sample $(X_i, Y_i)_{i=-n,\dots,n}$ that we split into two independent samples :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \; , \; Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

  - From Z^- we build an estimator $\widehat{b}$ of b

Principle of error estimation

- The (ϵ_i) are unobserved, so we construct proxies. More precisely, we observe a $2n$-sample $(X_i, Y_i)_{i=-n,\dots,n}$ that we split into two independent samples :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

  - From Z^- we build an estimator $\widehat{b}$ of b
  - The residuals from the second sample :

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) \ , \ i = 1, \dots, n$$

are proxies from the (ϵ_i). Given Z^-, they are i.i.d. from density f^-.

Principle of error estimation

- The (ϵ_i) are unobserved, so we construct proxies. More precisely, we observe a $2n$-sample $(X_i, Y_i)_{i=-n,\dots,n}$ that we split into two independent samples :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

  - From Z^- we build an estimator $\widehat{b}$ of b
  - The residuals from the second sample :

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) \ , \ i = 1, \dots, n$$

    are proxies from the (ϵ_i). Given Z^-, they are i.i.d. from density f^-.
  - Finally, by applying a density estimation procedure to the $(\widehat{\epsilon}_i)$, we get an estimator $\tilde{f}$.

The pointwise risk

- $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$

The pointwise risk

- $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$

  - $\mathbb{E}[(\tilde{f} - f^-)^2(x_0)|Z^-]$ is a density estimation error

The pointwise risk

- $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$

  - $\mathbb{E}[(\tilde{f} - f^-)^2(x_0)|Z^-]$ is a density estimation error
  - As $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) = \epsilon_i + (b - \widehat{b})(X_i)$, we have :

$$f^-(x_0) = \int_0^1 f(x_0 - (b - \widehat{b})(x))\mu(x)dx$$

The pointwise risk

- $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$

- $\mathbb{E}[(\tilde{f} - f^-)^2(x_0)|Z^-]$ is a density estimation error
- As $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) = \epsilon_i + (b - \widehat{b})(X_i)$, we have :

$$f^-(x_0) = \int_0^1 f(x_0 - (b - \widehat{b})(x))\mu(x)dx$$

Then, if f is Lipschitz with constant L, we have :

$$\begin{aligned} \mathbb{E}[(f - f^-)^2(x_0)] &\leq \mathbb{E}[\textstyle\int_0^1 (f(x_0) - f(x_0 - (b - \widehat{b})(x))^2 \mu(x)dx] \\ &\leq L^2 \mathbb{E}[\textstyle\int_0^1 (b - \widehat{b})^2(x)\mu(x)dx] \end{aligned}$$

The pointwise risk

- $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$
  - $\mathbb{E}[(\tilde{f} - f^-)^2(x_0)|Z^-]$ is a density estimation error
  - As $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) = \epsilon_i + (b - \widehat{b})(X_i)$, we have :

$$f^-(x_0) = \int_0^1 f(x_0 - (b - \widehat{b})(x))\mu(x)dx$$

Then, if f is Lipschitz with constant L, we have :

$$\begin{aligned} \mathbb{E}[(f - f^-)^2(x_0)] &\leq \mathbb{E}[\int_0^1 (f(x_0) - f(x_0 - (b - \widehat{b})(x))^2 \mu(x)dx] \\ &\leq L^2 \mathbb{E}[\int_0^1 (b - \widehat{b})^2(x)\mu(x)dx] \end{aligned}$$

- So we need two estimators :
  - An estimator of the regression function with an upper bound for the integrated risk
  - An estimator of the density with an upper bound for the pointwise risk

I) Density estimation by pointwise model selection

Let $(U_1, \ldots, U_n)$ i.i.d. from density g on $\mathbb{R}$ with $\nu := \|g\|_\infty < \infty$, and x_0 a fixed point in $\mathbb{R}$. We want to build an estimator of g by pointwise model selection.

I.a) Principle of model selection

I.b) Set of models

I.c) Classes of regularity

I.d) Estimation procedure

I.e) Results

I.a) Principle of model selection

We consider a finite collection $\{S_m, m \in \mathcal{M}_n\}$ of linear subsets of $L^2(\mathbb{R})$.

I.a) Principle of model selection

We consider a finite collection $\{S_m, m \in \mathcal{M}_n\}$ of linear subsets of $L^2(\mathbb{R})$.

- For every $m \in \mathcal{M}_n$, given $\{\phi_\lambda, \lambda \in I_m\}$ an orthonormal basis of S_m, the orthogonal projection of g onto S_m is : $g_m = \sum_{\lambda \in I_m} \langle \phi_\lambda, g \rangle \phi_\lambda$. Then, we consider the projection estimator of g onto S_m :

$$\widehat{g}_m := \sum_{\lambda \in I_m} (\frac{1}{n} \sum_{i=1}^{n} \phi_\lambda(U_i))\phi_\lambda$$

I.a) Principle of model selection

We consider a finite collection $\{S_m, m \in \mathcal{M}_n\}$ of linear subsets of $L^2(\mathbb{R})$.

- For every $m \in \mathcal{M}_n$, given $\{\phi_\lambda, \lambda \in I_m\}$ an orthonormal basis of S_m, the orthogonal projection of g onto S_m is : $g_m = \sum_{\lambda \in I_m} \langle \phi_\lambda, g \rangle \phi_\lambda$. Then, we consider the projection estimator of g onto S_m :

$$\widehat{g}_m := \sum_{\lambda \in I_m} (\frac{1}{n} \sum_{i=1}^{n} \phi_\lambda(U_i))\phi_\lambda$$

- We get a collection of estimators $\{\widehat{g}_m, m \in \mathcal{M}_n\}$, from which we would like to select the best one. For every $m \in \mathcal{M}_n$:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = \underbrace{(g - g_m)^2(x_0)}_{\text{bias}} + \underbrace{\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)]}_{\text{variance}}$$

I.a) Principle of model selection

We consider a finite collection $\{S_m, m \in \mathcal{M}_n\}$ of linear subsets of $L^2(\mathbb{R})$.

- For every $m \in \mathcal{M}_n$, given $\{\phi_\lambda, \lambda \in I_m\}$ an orthonormal basis of S_m, the orthogonal projection of g onto S_m is : $g_m = \sum_{\lambda \in I_m} \langle \phi_\lambda, g \rangle \phi_\lambda$. Then, we consider the projection estimator of g onto S_m :

$$\widehat{g}_m := \sum_{\lambda \in I_m} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_\lambda(U_i)\right) \phi_\lambda$$

- We get a collection of estimators $\{\widehat{g}_m, m \in \mathcal{M}_n\}$, from which we would like to select the best one. For every $m \in \mathcal{M}_n$:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = \underbrace{(g - g_m)^2(x_0)}_{\text{bias}} + \underbrace{\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)]}_{\text{variance}}$$

  - We estimate the bias term with help of $\widehat{g}_m$.
  - We upper-bound the variance term by a deterministic term function of m and n, called the penalty.

I.b) Set of models

The models are built from the sine-cardinal function :

$$\phi(x) := \frac{\sin(\pi x)}{\pi x}$$

Fore every $m \in \mathbb{N}^*$, $k \in \mathbb{Z}$, we define :

$$\phi_{m,k} := \sqrt{m}\phi(mx - k)$$

$$S_m = Vect(\phi_{m,k}, k \in \mathbb{Z})$$

and we consider the collection of models $\mathcal{M}_n = \{S_m, m = 1, \ldots, M_n\}$, with $M_n \leq n$.

I.c) Classes of regularity

For every $\beta > 0$, $K > 0$, let's define :

$$W(\beta, K) := \{h : \mathbb{R} \to \mathbb{R}, \int h = 1, \int_{\mathbb{R}} |h^*(\lambda)|^2 \lambda^{2\beta} d\lambda \leq L^2\}$$

where $h^*(\lambda) = \int_{\mathbb{R}} h(x) e^{i\lambda x} dx$.

I.c) Classes of regularity

For every $\beta > 0$, $K > 0$, let's define :

$$W(\beta, K) := \{h : \mathbb{R} \to \mathbb{R}, \int h = 1, \int_{\mathbb{R}} |h^*(\lambda)|^2 \lambda^{2\beta} d\lambda \leq L^2\}$$

where $h^*(\lambda) = \int_{\mathbb{R}} h(x) e^{i\lambda x} dx$.

Proposition

Let $\beta > 0$, $K > 0$, then :

$$(h - h_m)^2(x) \leq C m^{-(2\beta - 1)}, \ \forall h \in W(\beta, K), \forall x \in \mathbb{R}$$

for some constant C.

I.d) Estimation procedure

For every $m \leq M_n$, $\widehat{g}_m = \sum_{k \in \mathbb{Z}} [(1/n) \sum_{i=1}^{n} \phi_{m,k}(U_i)] \phi_{m,k}$ and :

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = (g - g_m)^2(x_0) + \mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)]$$

I.d) Estimation procedure

For every $m \leq M_n$, $\widehat{g}_m = \sum_{k\in\mathbb{Z}}[(1/n)\sum_{i=1}^{n}\phi_{m,k}(U_i)]\phi_{m,k}$ and :

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = (g - g_m)^2(x_0) + \mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)]$$

- Upper-bound for the variance term : $\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \leq \frac{\nu m}{n}$

I.d) Estimation procedure

For every $m \leq M_n$, $\widehat{g}_m = \sum_{k\in\mathbb{Z}}[(1/n)\sum_{i=1}^n \phi_{m,k}(U_i)]\phi_{m,k}$ and :

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = (g - g_m)^2(x_0) + \mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)]$$

- Upper-bound for the variance term : $\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \leq \frac{\nu m}{n}$
- The bias term is difficult to estimate, we replace it by :

$$\sup_{m\leq j\leq M_n} (g_j - g_m)^2(x_0)$$

Indeed, if $f \in W(\beta, L)$ with $\beta > 1/2$:

$$\begin{aligned}
\sup_{m\leq j\leq M_n} (g_j - g_m)^2(x_0) &\leq 2\{\sup_{m\leq j\leq M_n} (g_j - g)^2(x_0) + (g_m - g)^2(x_0)\} \\
&\leq 2C\{\sup_{m\leq j\leq M_n} j^{-(2\beta-1)} + m^{-(2\beta-1)}\} \\
&= C' m^{-(2\beta-1)}
\end{aligned}$$

- The natural idea is to replace $(g_j - g_m)^2(x_0)$ by $(\widehat{g}_j - \widehat{g}_m)^2(x_0)$ but :

$$\mathbb{E}[(\widehat{g}_j - \widehat{g}_m)^2(x_0)] = \\ (g_j - g_m)^2(x_0) + \underbrace{\mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2]}_{\leq \nu(j+m)/n}$$

- The natural idea is to replace $(g_j - g_m)^2(x_0)$ by $(\widehat{g}_j - \widehat{g}_m)^2(x_0)$ but :

$$\mathbb{E}[(\widehat{g}_j - \widehat{g}_m)^2(x_0)] =$$
$$(g_j - g_m)^2(x_0) + \underbrace{\mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2]}_{\leq \nu(j+m)/n}$$

- We define for every $m \leq M_n$:

$$\widehat{Crit}(m) := \sup_{m \leq j \leq M_n} [(\widehat{g}_j - \widehat{g}_m)^2(x_0) - x_{j,m}\frac{\nu(j+m)}{n}] + x_m \frac{\nu m}{n}$$
$$\widehat{m} := \arg\min_{m=1,\dots,M_n} \widehat{Crit}(m)$$

where $(x_{j,m})$ and x_m are numbers of order $\ln(j+m)$ and $\ln m$. Then our estimator is $\widehat{g}_{\widehat{m}}$.

- The natural idea is to replace $(g_j - g_m)^2(x_0)$ by $(\widehat{g}_j - \widehat{g}_m)^2(x_0)$ but :

$$\mathbb{E}[(\widehat{g}_j - \widehat{g}_m)^2(x_0)] = \\ (g_j - g_m)^2(x_0) + \underbrace{\mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2]}_{\leq \nu(j+m)/n}$$

- We define for every $m \leq M_n$:

$$\widehat{Crit}(m) := \sup_{m \leq j \leq M_n} [(\widehat{g}_j - \widehat{g}_m)^2(x_0) - x_{j,m}\frac{\nu(j+m)}{n}] + x_m \frac{\nu m}{n} \\ \widehat{m} := \arg\min_{m=1,\ldots,M_n} \widehat{Crit}(m)$$

  where $(x_{j,m})$ and x_m are numbers of order $\ln(j+m)$ and $\ln m$. Then our estimator is $\widehat{g}_{\widehat{m}}$.

- **Remark** : ν can be replaced by an estimator $\widehat{\nu}_n$.

Theorem

If $g \in W(\beta, K)$ *with* $\beta > 1/2$ *then :*

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + \frac{C'}{n}$$

Theorem

If $g \in W(\beta, K)$ *with* $\beta > 1/2$ *then :*

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + \frac{C'}{n}$$

cf Butucea (2001)

- The minimax rate of convergence over $W(\beta, K)$ is $n^{-(2\beta-1)/(2\beta)}$
- The adaptative minimax rate of convergence over the classes $\{W(\beta, K), \beta > 1/2\}$ is $(n/\ln n)^{-(2\beta-1)/(2\beta)}$

II) The errors density

Let's consider $(X_i, Y_i)_{i=-n,\ldots,n}$ from the regression framework, and :

$$Z^- = \{(X_i, Y_i), i = -n, \ldots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \ldots, n\}$$

II) The errors density

Let's consider $(X_i, Y_i)_{i=-n,\dots,n}$ from the regression framework, and :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

- Let $\widehat{b}$ be any estimator of b built from the sample Z^-.

II) The errors density

Let's consider $(X_i, Y_i)_{i=-n,\dots,n}$ from the regression framework, and :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

- Let $\widehat{b}$ be any estimator of b built from the sample Z^-.
- Let consider :

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) \ , \ i = 1, \dots, n$$

the residuals from the sample Z^+. Given Z^-, the $(\widehat{\epsilon}_i)$ are i.i.d. of conditionnal density f^-, and $\nu^- := \|f^-\|_\infty$.

II) The errors density

Let's consider $(X_i, Y_i)_{i=-n,\dots,n}$ from the regression framework, and :

$$Z^- = \{(X_i, Y_i), i = -n, \dots, -1\} \ , \ Z^+ = \{(X_i, Y_i), i = 1, \dots, n\}$$

- Let $\widehat{b}$ be any estimator of b built from the sample Z^-.
- Let consider :

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i) \ , \ i = 1, \dots, n$$

  the residuals from the sample Z^+. Given Z^-, the $(\widehat{\epsilon}_i)$ are i.i.d. of conditionnal density f^-, and $\nu^- := \|f^-\|_\infty$.
- For every $m \leq M_n$:

$$\widehat{f}_m^- := \sum_{k\in\mathbb{Z}}((1/n)\sum_{i=1}^n \phi_{m,k}(\widehat{\epsilon}_i))\phi_{m,k}$$
$$\widehat{Crit}^-(m) = \sup_{m\leq j\leq M_n}[(\widehat{f}_j^- - \widehat{f}_m^-)^2(x_0) - x_{j,m}\frac{\nu^-(j+m)}{n}] + x_m\frac{\nu^- m}{n}$$
$$\widehat{m} = \arg\min_{m=1,\dots,M_n} \widehat{Crit^-}(m)$$

and our estimator of f is $\tilde{f} := \widehat{f}_{\widehat{m}}^-$.

Theorem

If $f \in W(\beta, K)$ *with* $\beta > 3/2$ *:*

$$\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + C'\mathbb{E}[\|\widehat{b} - b\|_\mu^2]$$

Theorem

If $f \in W(\beta, K)$ *with* $\beta > 3/2$ *:*

$$\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + C'\mathbb{E}[\|\widehat{b} - b\|_\mu^2]$$

Proof : $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$

- $\mathbb{E}[(\tilde{f} - f^-)^2(x_0)|Z-] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + \frac{C'}{n}$
  $\Rightarrow \mathbb{E}[(\tilde{f} - f^-)^2(x_0)] \leq C''(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}}$
- $\mathbb{E}[(f^- - f)^2(x_0)] \leq \mathbb{E}[\|\widehat{b} - b\|_\mu^2]$

Theorem

If $f \in W(\beta, K)$ with $\beta > 3/2$:

$$\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + C'\mathbb{E}[\|\widehat{b} - b\|_{\mu}^2]$$

Proof : $\mathbb{E}[(\tilde{f} - f)^2(x_0)] \leq 2\{\mathbb{E}[(\tilde{f} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}$

- $\mathbb{E}[(\tilde{f} - f^-)^2(x_0)|Z-] \leq C(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}} + \frac{C'}{n}$
  $\Rightarrow \mathbb{E}[(\tilde{f} - f^-)^2(x_0)] \leq C''(\frac{n}{\ln n})^{-\frac{2\beta-1}{2\beta}}$
- $\mathbb{E}[(f^- - f)^2(x_0)] \leq \mathbb{E}[\|\widehat{b} - b\|_{\mu}^2]$

- **Consequence** If we consider an adaptative estimator for b (cf Baraud, 2001), the rate of convergence for $\tilde{f}$ is the maximum of :

  - the minimax rate of convergence of b.

  - the minimax rate of convergence of f is the sample (ϵ_i) was observed.

Bibliography

- Y. Baraud, 'Model selection for regression on a random design', 2001, ESAIM.
- B. Laurent, C Ludena, C Prieur, 'Adaptative estimation of linear functionnals by model selection', 2007, ArXiv.
- C. Butucea, 'Exact adaptative pointwise estimation on Sobolev classes of densities', 2001, ESAIM.