Model selection for a semi - Markov continuous time regression observed in the discrete time moments

Serguei PERGAMENCHTCHIKOV
University of Rouen (France) and
Tomsk State University (Russia)

International Conference "SAPS XI",
Saint Petersbourg, Russia, July 20, 2017
Joint work with V. Barbu and S. Beltaief

Ce projet est cofinancé par l'Union européenne.
L'Europe s'engage en Normandie
avec le Fonds européen de Développement Régional (FEDER)

Model

Consider a regression model in continuous time

$$\mathrm{d}y_t = S(t)\mathrm{d}t + \mathrm{d}\xi_t\,, \quad 0 \le t \le n\,,$$

where S is an unknown 1 - periodic function; $\xi = (\xi_t)_{t\ge 0}$ is an unobservable noise such that for each square integrable function function f the stochastic integral

$$I_n(f) = \int_0^n f(s)\mathrm{d}\xi_s$$

is well defined and has the properties:

$$\mathbf{E}_Q I_n(f) = 0 \quad \text{and} \quad \mathbf{E}_Q I_n^2(f) \le \varkappa_Q \int_0^n f^2(s)\mathrm{d}s\,.$$

Model

If $(\xi_t)_{t\geq 0}$ is the Wiener process, then we obtain the well - known "signal + white noise model" introduced by Ibragimov and Khasminskii (1979) and Pinsker (1981).

If $(\xi_t)_{t\geq 0}$ is the Gaussian process, then we obtain "signal + color noise model", introduced by Kutoyants (1977) for the parametric estimation. Later, Konev and Pergamenshchikov (2003), Höpfner and Kutoyants (2009) use these models with the Ornstein-Uhlenbeck noises for the parametric estimation and, Konev and Pergamenshchikov (2010) for the non parametric estimation.

If $(\xi_t)_{t\geq 0}$ is the Non-Gaussian Ornstein-Uhlenbeck process (Barndorf - Nielsen and Shephard (2001)) then we obtain "signal + color noise model with jumps", introduced by Pchelintsev (2013) for the parametric estimation and Konev and Pergamenshchikov (2012) for the non parametric estimation.

Problem

The first problem: estimation of S on the basis of the continuous observations

$$(y_t)_{0 \le t \le n}$$

The second problem: estimation of S on the basis of the discret observations

$$(y_{t_j})_{0 \le j \le pn}$$

where $t_j = j/p$ and p is observation frequency. The main condition

$$p \geq n^{5/6} .$$

Quadratic risk

The quality of an estimate $\widetilde{S}_n$ with the quadratic risk

$$\mathcal{R}_Q(\widetilde{S}_n, S) = \mathbf{E}_Q \|\widetilde{S}_n - S\|^2, \quad \|f\|^2 = \int_0^1 f^2(t)\mathrm{d}t\,.$$

Since the noise distribution Q is unknown, it seems reasonable to introduce the robust risk of the form

$$\mathcal{R}_n^*(\widetilde{S}_n, S) = \sup_{Q \in \mathcal{Q}_n} \mathcal{R}_Q(\widetilde{S}_n, S)\,,$$

which enables one to take into account the information that $Q \in \mathcal{Q}_n$ and ensures the quality of an estimate $\widetilde{S}_n$ for all distributions in the family $\mathcal{Q}_n$.

Lévy process

We consider the noise $(\xi_t)_{t\geq 0}$ defined as

$$\xi_t = \varrho_1 L_t + \varrho_2 z_t \,.$$

Here $(L_t)_{t\geq 0}$ is a zero mean Lévy process, i.e.

$$L_t = \check{\varrho}\, w_t + \sqrt{1-\check{\varrho}^2}\, \check{L}_t \,, \quad \check{L}_t = x * (\mu - \widetilde{\mu})_t \,,$$

where, $0 \leq \check{\varrho} \leq 1$ is an unknown constant, $(w_t)_{t\geq 0}$ is a standard Brownian motion, $\mu(\mathrm{d}s, \mathrm{d}x)$ is the jump measure with the deterministic compensator $\widetilde{\mu}(\mathrm{d}s\,\mathrm{d}x) = \mathrm{d}s\Pi(\mathrm{d}x)$, where $\Pi(\cdot)$ is some positive measure on $\mathbb{R}$ which is called the Léve measure.
The parameters ϱ_1 and ϱ_2 are unknown constants.

Lévy measure

The jumps measure is defined on the Borel σ - field in $\mathbb{R}_+ \times \mathbb{R}_0$, where $\mathbb{R}_0 = \mathbb{R} \setminus \{0\}$. For any $t > 0$ and any Borel set $A \subseteq \mathbb{R}_0$

$$\mu([0,t] \times A) = \sum_{0 \le s \le t} \mathbf{1}_{\{\Delta \check{L}_s \in A\}}.$$

The Léve measure is defined on the σ - field in $\mathbb{R}_0$ and for any Borel set $A \subseteq \mathbb{R}_0$

$$\Pi(A) = \mathbf{E}\mu([0,1] \times A) = \mathbf{E} \sum_{0 \le s \le 1} \mathbf{1}_{\{\Delta \check{L}_s \in A\}}.$$

Semi-Markov processes

Here, z_t is the semi-Markov process (see, for example, Barbu and Liminiuos (2008)), defined as

$$z_t = \sum_{j=1}^{N_t} Y_j$$

where $(N_t)_{t \geq 0}$ is a renewal counting function

$$N_t = \sum_{k \geq 1} \mathbf{1}_{\{T_k \leq t\}},$$

the renewal seuqence $T_k = \sum_{j=1}^{k} \tau_j$ and $(\tau_j)_{j \geq 1}$ is the i.i.d. sequence of positive random variables with $\mathbf{E}\tau_1 < \infty$.

Semi-Markov processes

Moreover, $(Y_j)_{j\geq 1}$ is an i.i.d. sequence of random variables with

$$\mathbf{E}Y_1 = 0\,, \quad \mathbf{E}Y_1^2 = 1 \quad \text{and} \quad \mathbf{E}Y_1^4 < \infty\,.$$

Note that if τ_j is exponential random variables with the parameter $\lambda > 0$, then N_t is a standard homogeneous Poisson process with the intensity $\lambda > 0$, $(z_t)_{t\geq 0}$ is a compound Poisson process and ξ_t is the Levy process in this case.

Noise distributions

As to the parameters ϱ_1 and ϱ_2 we assume that

$$\sigma_Q = \varrho_1^2 + \varrho_2^2 / \mathbf{E}\,\tau_1 \le \sigma^* \,,$$

where the unknown bound σ^* is a function of n, i.e. $\sigma^* = \sigma^*(n)$, such that for any $\delta > 0$

$$\lim_{n\to\infty} n^\delta\, \sigma^*(n) = +\infty \quad \text{and} \quad \lim_{n\to\infty} \frac{\sigma^*(n)}{n^\delta} = 0 \,.$$

We denote by $\mathcal{Q}_n$ the family of all distributions of the process $(\xi_t)_{0\le t\le n}$ in $\mathbf{D}[0,n]$ satisfying these properties.

Construction of procedure

The model selection method introduced by Konev and Pergamenshchikov (2009) for the semimartingal regression models. The function $S \in \mathcal{L}_2[0,1]$ can be represented as

$$S(t) = \sum_{j \geq 1} \theta_j \, \phi_j(t) \, ,$$

where

$$\theta_j = (S, \phi_j) = \int_0^1 S(t) \phi(t) \, \mathrm{d}t \, .$$

The first step in constructing the model selection procedure consists in estimating the coefficients θ_j for S

$$\widehat{\theta}_j = \frac{1}{n} \int_0^n \phi_j(t) \, \mathrm{d} \, y_t \, .$$

Family of estimators

Now we introduce a weight least squares estimate for $S(t)$ as

$$\widehat{S}_\gamma(t) = \sum_{j=1}^{n} \gamma(j)\widehat{\theta}_j \phi_j(t)\,,$$

where $\gamma = (\gamma(j))_{1\le j\le n}$ is the vector of weight coefficients $0 \le \gamma(j) \le 1$. The model selection procedure will be chosen from a finite family of such estimates

$$(\widehat{S}_\gamma)_{\gamma\in\Gamma}\,.$$

Cost function

The empirical squared error of the estimator can be represented as

$$\mathrm{Err}(\gamma) = \|\widehat{S}_\gamma - S\|^2 = \sum_{j=1}^{n} \gamma^2(j)\widehat{\theta}_j^2 - 2\sum_{j=1}^{n} \gamma(j)\widehat{\theta}_j\theta_j + \|S\|^2 .$$

Since the Fourier coefficients $(\theta_j)_{j\geq 1}$ are unknown, the weight coefficients $(\gamma(j))_{1\leq j\leq n}$ cannot be determined by minimizing this quantity. To circumvent this difficulty we replace the terms $\widehat{\theta}_j\theta_j$ by

$$\widetilde{\theta}_j = \widehat{\theta}_j^2 - \frac{\sigma_Q}{n} ,$$

where σ_Q is the noise variance.

Cost function

For replacing the terms $\widehat{\theta}_j\,\theta_j$ by its estimates on the right-hand side of the empirical squared error, one has to pay some penalty. Thus, one comes to the cost function of the form

$$J(\gamma) = \sum_{j=1}^{n} \gamma^2(j)\widehat{\theta}_j^2 - 2\sum_{j=1}^{n} \gamma(j)\,\widetilde{\theta}_j + \rho\,P(\gamma)$$

where ρ is some positive constant and $\widehat{P}(\gamma)$ is the penalty term defined as

$$P_Q(\gamma) = \frac{\sigma_Q\,|\gamma|^2}{n}\,.$$

Model selection procedure

Minimizing the cost function

$$\widehat{\gamma} = \mathrm{argmin}_{\gamma \in \Gamma} J(\gamma)\,,$$

yields the model selection procedure

$$\widehat{S}_* = \widehat{S}_{\widehat{\gamma}}\,.$$

Oracle inequalities

Theorem

For any $n \geq 2$ *and* $0 < \rho < 1/3$

$$\mathcal{R}_Q(\widehat{S}_*, S) \leq \frac{1+3\rho}{1-3\rho} \min_{\gamma \in \Gamma} \mathcal{R}_Q(\widehat{S}_\gamma, S) + \frac{\mathbf{B}_Q(n)}{n\rho}$$

and

$$\mathcal{R}^*_n(\widehat{S}_*, S) \leq \frac{1+3\rho}{1-3\rho} \min_{\gamma \in \Gamma} \mathcal{R}^*_n(\widehat{S}_\gamma, S) + \frac{\mathbf{B}^*_n}{n\rho}\,.$$

Specification of weights

Consider a numerical grid of the form

$$\mathcal{A} = \{1, \ldots, k^*\} \times \{r_1, \ldots, r_m\},$$

where $r_i = i\varepsilon$ and $m = [1/\varepsilon^2]$. For each $\alpha = (\beta, t) \in \mathcal{A}$, we introduce the weight sequence

$$\gamma_\alpha = (\gamma_\alpha(j))_{1 \le j \le p}$$

with the elements

$$\gamma_\alpha(j) = \mathbf{1}_{\{1 \le j < j_*\}} + \left(1 - (j/\omega_\alpha)^\beta\right) \mathbf{1}_{\{j_* \le j \le \omega_\alpha\}}.$$

Now we define the set Γ as

$$\Gamma = \{\gamma_\alpha, \alpha \in \mathcal{A}\}.$$

Model

For the Monte Carlo simulations we chose a 1-periodic function which is defined as

$$S(t) = \begin{cases} |t - \frac{1}{2}| & \text{if} \quad \frac{1}{4} \leq t \leq \frac{3}{4}\,, \\ \\ \frac{1}{4} & \text{elsewhere,} \end{cases}$$

where $0 \leq t \leq 1$.

Model

We simulate the model

$$\mathrm{d}y_t = S(t)\mathrm{d}t + \mathrm{d}\xi_t \quad \text{and} \quad \xi_t = 0.5\mathrm{d}w_t + 0.5\mathrm{d}z_t\,.$$

Here z_t is the semi-Markov process defined through i.i.d. sequence $(Y_j)_{j\geq 1}$ and $(\tau_k)_{k\geq 1}$

$$Y_j \sim \mathcal{N}(0,1) \quad \text{and} \quad \tau_k \sim \chi^2_3\,.$$

Model

We use the model selection procedure with

$$\rho = (3 + \ln n)^{-2} .$$

The parameters of the weight coefficients :

$$r_i = \frac{i}{\ln n} , \quad m = [\ln^2 n] \quad \text{and} \quad k^* = 100 + \sqrt{\ln n} .$$

Empirical risks

We define the empirical risk as

$$\overline{\mathbf{R}} = \frac{1}{p}\sum_{j=1}^{p} \hat{\mathbf{E}}\left(\hat{S}_n(t_j) - S(t_j)\right)^2 ,$$

where the observation frequency $p = 100001$ and the expectations was taken as an average over $N = 10000$ replications, i.e.

$$\hat{\mathbf{E}}\left(\hat{S}_n(.) - S(.)\right)^2 = \frac{1}{N}\sum_{l=1}^{N}\left(\hat{S}_n^l(\cdot) - S(\cdot)\right)^2 .$$

Empirical risks

We set the relative quadratic risk as

$$\overline{\mathbf{R}_*} = \overline{\mathbf{R}} / \|S\|_p^2 \quad \text{and} \quad \|S\|_p^2 = \frac{1}{p} \sum_{j=0}^{p} S^2(t_j) \, .$$

In our case $\|S\|_p^2 = 0.1883601$.

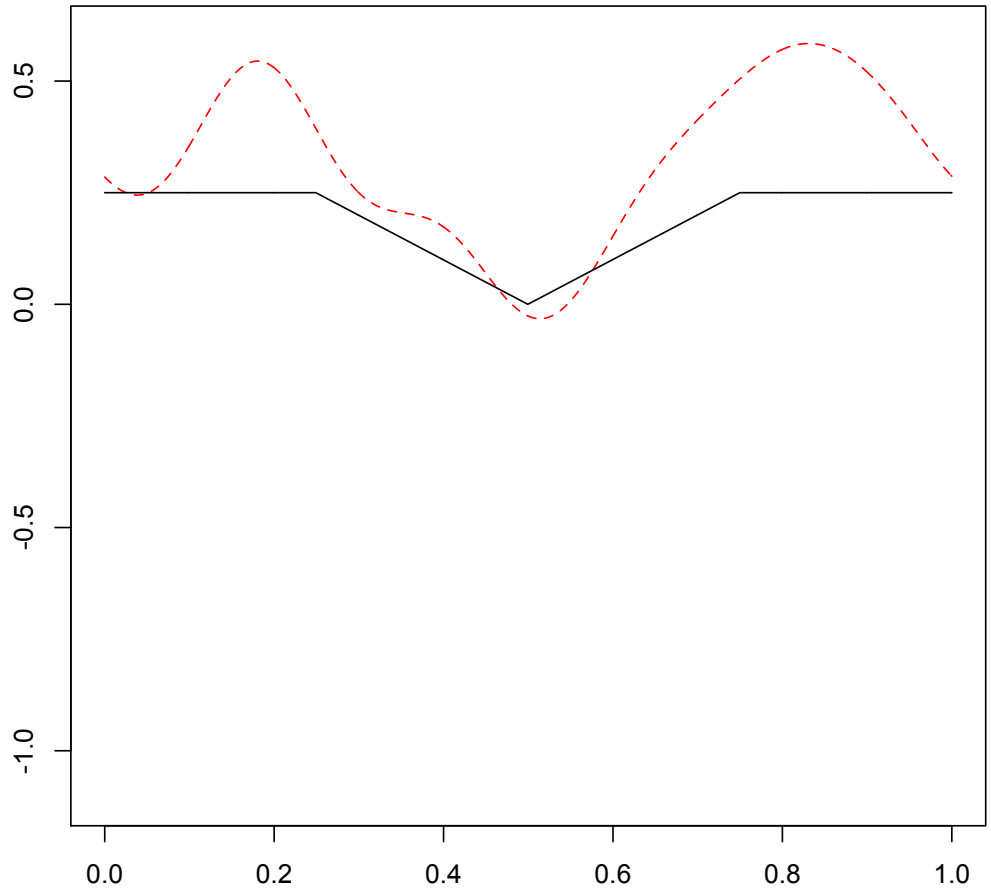
Model

The table below gives the values for the sample risks for different numbers of observations n.

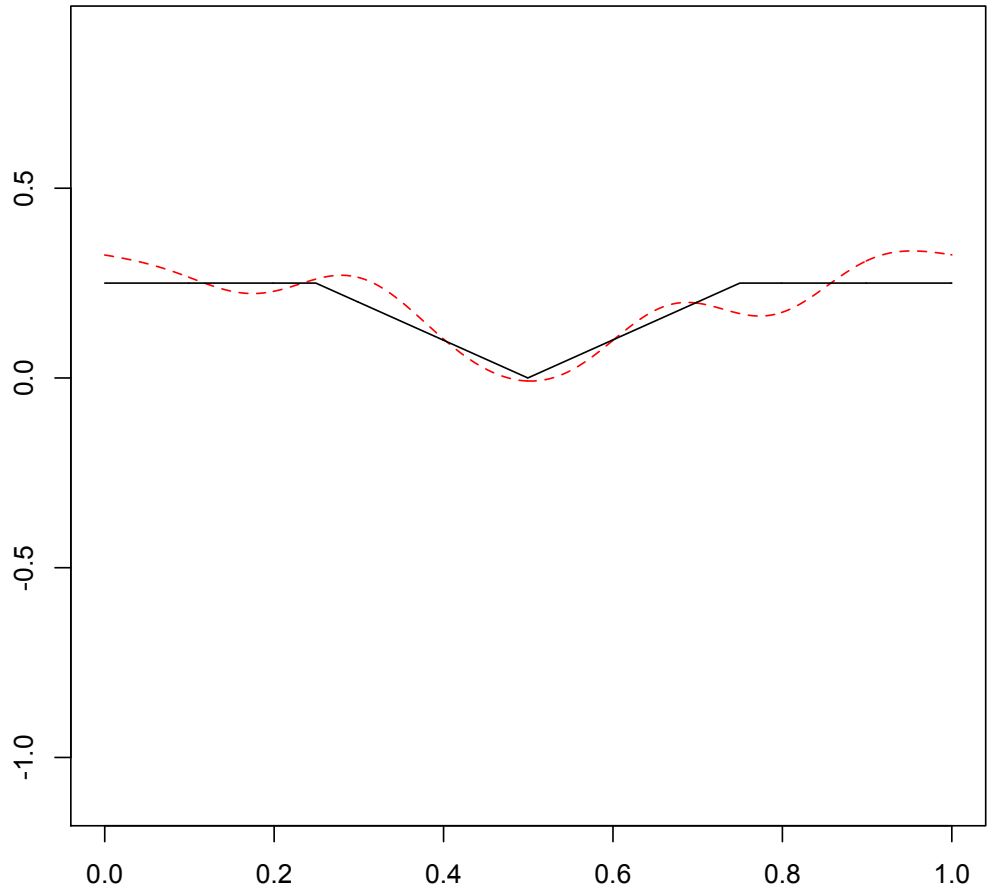
n	$\overline{\mathbf{R}}$	$\overline{\mathbf{R}_*}$
20	0.0398	0.211
100	0.0091	0.0483
200	0.0067	0.0355
1000	0.0022	0.0116

Table : Empirical risks

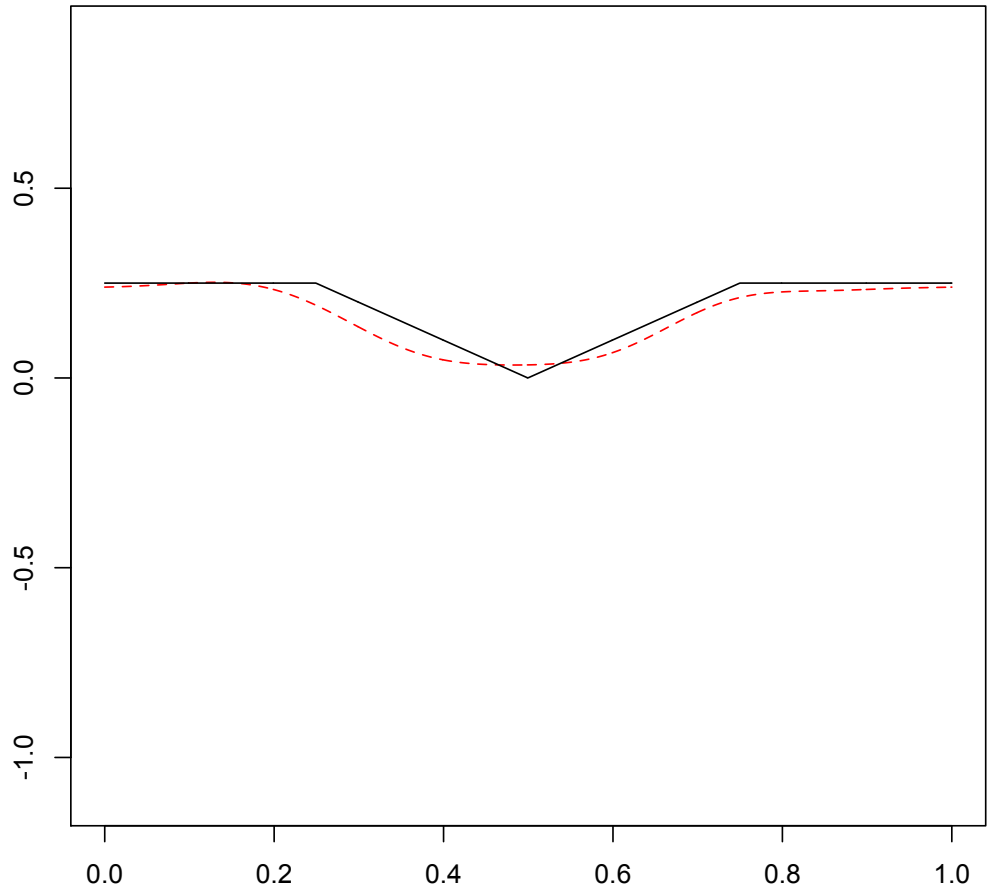
$n = 20$



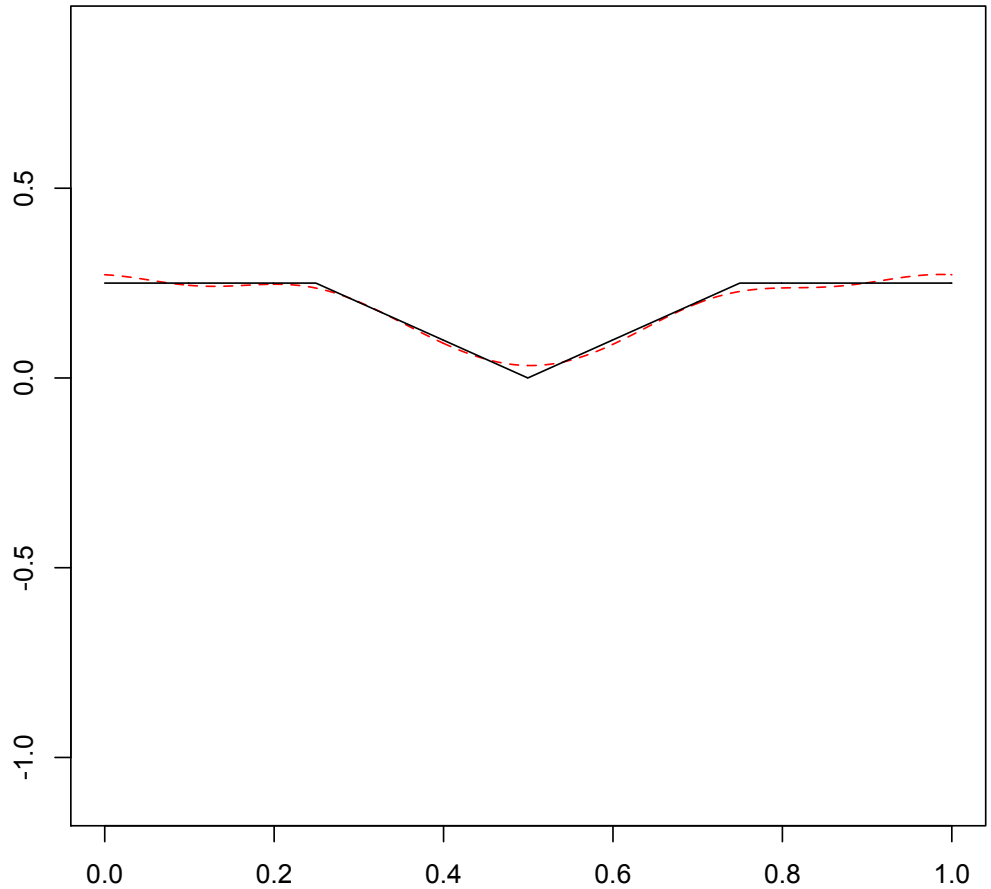
$n = 100$



$n = 200$



$n = 1000$



Lower bound

We show that the Pinsker constant in this case has the following form

$$R_k^* = ((2k+1)r)^{1/(2k+1)} \left(\frac{k}{(k+1)\pi} \right)^{2k/(2k+1)}.$$

To this end we denote by Π_n the set of all estimators $\widehat{S}_n$ measurable with respect to the sigma-algebra $\sigma\{y_t\,,\, 0 \le t \le n\}$.
It is well known that for the simple risks the optimal (minimax) estimation convergence rate for the functions from the set W_r^k is $n^{2k/(2k+1)}$ (see, for example, Pinsker (1981), Nussbaum (1985)).

Lower bound

Theorem

We obtain the following lower bound

$$\liminf_{n\to\infty} v_n^{\frac{2k}{2k+1}} \inf_{\widehat{S}_n\in\Pi_n} \sup_{S\in W_r^k} \mathcal{R}_n^*(\widehat{S}_n, S) \geq R_k^* ,$$

where $v_n = n/\sigma^*$.

Upper bound

Theorem

The robust risk for the model selection procedure $\widehat{S}_*$ *admits the following asymptotic upper bound*

$$\limsup_{n\to\infty} v_n^{\frac{2k}{2k+1}} \sup_{S\in W_r^k} \mathcal{R}_n^*(\widehat{S}_*, S) \le R_k^* .$$

Efficient estimation

Theorem

Under the conditions listed above

$$\lim_{n\to\infty} v_n^{\frac{2k}{2k+1}} \inf_{\widehat{S}_n\in\Pi_n} \sup_{S\in W_r^k} \mathcal{R}_n^*(\widehat{S}_n, S) = R_k^* .$$

We obtain the efficiency

$$\lim_{n\to\infty} \frac{\inf_{\widehat{S}_n\in\Pi_n} \sup_{S\in W_r^k} \mathcal{R}_n^*(\widehat{S}_n, S)}{\sup_{S\in W_r^k} \mathcal{R}_n^*(\widehat{S}_*, S)} = 1 .$$

Efficient estimation

It should be noted that the equality means the robust efficiency holds with the convergence rate

$$v_n^{\frac{2k}{2k+1}} \quad \text{and} \quad v_n = n/\sigma^* \, .$$

If the distribution upper bound $\sigma^* \to 0$ as $n \to \infty$ we obtain a faster rate with respect to $n^{2k/(2k+1)}$, and if $\sigma^* \to \infty$ as $n \to \infty$ we obtain a slower rate. In the case when σ^* is constant the robuste rate is the same as the classical non robuste convergence rate.

Renewal density

We recall that the process z_t is defined through the renewal process

$$N_t = \sum_{k \geq 1} \mathbf{1}_{\{\sum_{j=1}^{k} \tau_j \leq t\}} ,$$

where $(\tau_j)_{j \geq 1}$ arei.i.d. positive random variables with the density g. Let us denote by η the renewal density

$$\eta(x) = \sum_{l=1}^{\infty} g^{(l)}(x) ,$$

where $g^{(l)}$ is the lth convolution power of the density g.

Renewal density

Goldie (1991)

Theorem

Under some technical conditions the renewal density ρ *is such that*

$$\rho(x) = \frac{1}{\mathbf{E}\tau_1} + \Delta(x)\,,$$

where $\Delta(\cdot)$ *is some function defined on* $\mathbb{R}_+$ *with values in* $\mathbb{R}$ *such that*

$$\sup_{x \geq 0} x^\gamma |\Delta(x)| < \infty \quad \textit{for all} \quad \gamma > 0\,.$$

Conclusion

In the conclusion we would like to emphasize that in this talk :

- we construct a selection model procedure based on the weight least square estimators;
- we find conditions for which we obtained an sharp non asymptotic oracle inequalities for the simple quadratic risks and for the robust risks as well;
- using the Pinsker method we obtain a lower bound for the robust quadratic risks, then, through the obtained sharp oracle inequalities we show that the risk upper bound for the constructed procedure matters this lower bound, i.e. the procedure is efficient in the adaptative setting.

Conclusion

THANK YOU VERY MUCH

FOR YOUR ATTENTION !!!