МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО
ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Международная лаборатория статистики случайных
процессов и количественного финансового анализа

Международная научная конференция «Робастная статистика и финансовая математика – 2019»

(04–06 июля 2019 г.)

Сборник статей

Под редакцией
д-ра физ.-мат. наук, профессора С.М. Пергаменщикова,
канд. физ.-мат. наук, доцента Е.А. Пчелинцева

Томск
Издательский Дом Томского государственного университета
2019

Improved model selection for estimation in semimartingale regression from discrete data [1]

Pchelintsev E. A.

Tomsk State University, Tomsk
e-mail: evgen-pch@yandex.ru

Abstract

We consider the adaptive nonparametric estimation problem for a function of a continuous time regression model with semimartingale noises by incomplete observations. We consider the regression model defined by non-Gaussian Ornstein–Uhlenbeck processes. A model selection procedure, based on the shrinkage weighted least squares estimates, is proposed. The sharp oracle inequalities for the robust risks are obtained. the robust efficiency property has been established in adaptive setting.

Keywords: improved estimation, least squares estimates, robust quadratic risk, Ornstein–Uhlenbeck process, semimartingale regression, model selection, sharp oracle inequality, asymptotic efficiency.

Introduction. Statement of a problem. In this paper we consider the following continuous time regression model

$$\mathrm{d}y_t = S(t)\mathrm{d}t + \mathrm{d}\xi_t\,, \quad 0 \le t \le n\,, \tag{1}$$

where S is an unknown 1-periodic $\mathbb{R} \to \mathbb{R}$ function from $\mathbf{L}_2[0,1]$, $(\xi_t)_{t\ge0}$ is an unobservable noise which is defined by a non-Gaussian Ornstein–Uhlenbeck process with the Lévy subordinator

$$\mathrm{d}\xi_t = a\xi_t\mathrm{d}t + \mathrm{d}u_t\,, \quad \xi_0 = 0\,, \tag{2}$$

where

$$u_t = \varrho_1\, w_t + \varrho_2\, z_t \quad \text{and} \quad z_t = x * (\mu - \widetilde{\mu})_t\,. \tag{3}$$

Here $(w_t)_{t\ge0}$ is a standard Brownian motion, "$*$" denotes the stochastic integral with respect to the compensated jump measure $\mu(\mathrm{d}s\,,\,\mathrm{d}x)$ with deterministic compensator $\widetilde{\mu}(\mathrm{d}s\,\mathrm{d}x) = \mathrm{d}s\Pi(\mathrm{d}x)$, i.e.

$$z_t = \int_0^t \int_{\mathbb{R}_*} v\,(\mu - \widetilde{\mu})(\mathrm{d}s\,\mathrm{d}v) \quad \text{and} \quad \mathbb{R}_* = \mathbb{R} \setminus \{0\}\,,$$

$\Pi(\cdot)$ is the Lévy measure on $\mathbb{R}_* = \mathbb{R} \setminus \{0\}$, (see, for example in [1]), such that

$$\Pi(x^2) = 1 \quad \text{and} \quad \Pi(x^8) < \infty\,. \tag{4}$$

[1] The work was supported by the RSF, the project No 17-11-01049.

We use the notation $\Pi(|x|^m) = \int_{\mathbb{R}_*} |z|^m \, \Pi(\mathrm{d}z)$. Moreover, we assume that the nuisance parameters $a \le 0$, ϱ_1 and ϱ_2 satisfy the conditions

$$-a_{max} \le a \le 0\,, \quad 0 < \underline{\varrho} \le \varrho_1^2 \quad \text{and} \quad \sigma_Q = \varrho_1^2 + \varrho_2^2 \le \varsigma^*\,, \tag{5}$$

where the bounds a_{max}, $\underline{\varrho}$ and ς^* are functions of n, i.e. $a_{max} = a_{max}(n)$, $\underline{\varrho} = \underline{\varrho}_n$ and $\varsigma^* = \varsigma^*_n$, such that for any $\epsilon > 0$

$$\lim_{n\to\infty} \frac{a_{max}(n) + \varsigma^*_n}{n^\epsilon} = 0 \quad \text{and} \quad \liminf_{n\to\infty} n^\epsilon \, \underline{\varrho}_n > 0\,. \tag{6}$$

We denote by $\mathcal{Q}_n$ the family of all distributions of process (1) – (2) on the Skorokhod space $\mathbf{D}[0,n]$ satisfying the conditions (5) – (6). It should be noted that the process (2) is conditionally-Gaussian square integrated semimartingale with respect to σ-algebra $\mathcal{G} = \sigma\{z_t\,,\, t \ge 0\}$ which is generated by jump process $(z_t)_{t\ge0}$.

The problem is to estimate the unknown function S in the model (1) on the basis of observations

$$(y_{t_j})_{0\le j\le np}\,, \quad t_j = \frac{j}{p}\,, \tag{7}$$

where the observations frequency p is some fixed integer number. For this problem we use the quadratic risk, which for any estimate $\widehat{S}$, is defined as

$$\mathcal{R}_Q(\widehat{S}, S) := \mathbf{E}_{Q,S} \, \|\widehat{S} - S\|^2 \quad \text{and} \quad \|f\|^2 := \int_0^1 f^2(t)\mathrm{d}t\,, \tag{8}$$

where $\mathbf{E}_{Q,S}$ stands for the expectation with respect to the distribution $\mathbf{P}_{Q,S}$ of the process (1) with a fixed distribution Q of the noise $(\xi_t)_{0\le t\le n}$ and a given function S. Moreover, in the case when the distribution Q is unknown we use also the robust risk

$$\mathcal{R}^*(\widehat{S}, S) = \sup_{Q\in\mathcal{Q}_n} \mathcal{R}_Q(\widehat{S}, S)\,. \tag{9}$$

To study the estimation problem for non-Gaussian observations (1) in the papers [2, 8, 7, 5, 6] it was introduced impulse noises defined through the semi-Markov or compound Poisson processes with unknown impulse distributions. However, the semi-Markov or compound Poisson processes can describe the impulse influence of only one fixed frequency. It should be noted that in the telecommunication systems, the noise impulses are without limitations on frequencies and, therefore, such models are too restricted for practical applications. To include all possible impulse noises, in [3, 4] it was proposed to use general non-Gaussian semimartingale processes. Later, for semimartingale models in the papers [9, 10, 11, 12] the authors developed the improved (shrinkage) nonparametric estimation methods. It should be emphasized, that in all these papers the improved estimation problems are studied only for the complete observations cases, i.e. when the all trajectory $(y_t)_{0\le t\le n}$ is accessed to be observed.

Our main goal in this paper is to develop improved estimation meth-

ods for the incomplete observations, i.e. when the process (1) can be observed only in the fixed time moments (7). As an example, we consider the regression model (1) with the noise defined by non-Gaussian Ornstein–Uhlenbeck process with unknown distribution.

Improved estimation. For estimating the unknown function S in (1) we will use it's Fourier expansion with respect to an orthonormal basis $(\phi_j)_{j\geq 1}$ in $\mathbf{L}_2[0,1]$. We extend these functions by the periodic way on $\mathbb{R}$, i.e. $\phi_j(t)=\phi_j(t+1)$ for any $t\in\mathbb{R}$. Assume that the basis functions are uniformly bounded, i.e. for some constant $\phi_*\geq 1$, which may be depend on n,

$$\sup_{0\leq j\leq n}\sup_{0\leq t\leq 1}|\phi_j(t)|\leq\phi_*<\infty\,. \tag{10}$$

Moreover we will use such basis that the restrictions of the functions $(\phi_j)_{1\leq j\leq p}$, on the sampling lattice

$$\mathcal{T}_p=\{t_1,...,t_p\},\quad t_j=j/p,$$

form an orthonormal basis in the Hilbert space $\mathbb{R}^{\mathcal{T}_p}$ with the inner product

$$(x,y)_p=\frac{1}{p}\sum_{j=1}^{p}x(t_j)y(t_j)\quad\text{for}\quad x,y\in\mathbb{R}^{\mathcal{T}_p}, \tag{11}$$

i.e. $(\phi_i,\phi_j)_p=\mathbf{1}_{\{i=j\}}$. We put the norm $\|x\|_p=\sqrt{(x,x)_p}$.

We write the discrete Fourier expansion of the unknown function S on the lattice $\mathcal{T}_p$ in the form

$$S(t)=\sum_{j=1}^{p}\theta_{j,p}\phi_j(t),$$

where the corresponding Fourier coefficients

$$\theta_{j,p}=(S,\phi_j)_p \tag{12}$$

can be estimated from the discrete data by the formulae

$$\widehat{\theta}_{j,p}=\frac{1}{n}\int_0^n\psi_{j,p}(t)\,\mathrm{d}y_t\,,\quad \psi_{j,p}(t)=\sum_{k=1}^{np}\phi_j(t_k)\mathbf{1}_{(t_{k-1},t_k]}(t)\,. \tag{13}$$

As in [6] we define a class of weighted least squares estimates for $S(t)$ as

$$\widehat{S}_\gamma(t)=\sum_{j=1}^{p}\gamma(j)\widehat{\theta}_{j,n}\psi_{j,p}(t)\,, \tag{14}$$

where the weights $\gamma=(\gamma(j))_{1\leq j\leq p}\in\mathbb{R}^p$ belong to some finite set Γ from $[0,\,1]^p$. We will use here the set Γ from [10].

For the first $d>6$ Fourier coefficients in (14) we will use the improved estimation method proposed for parametric models in [8]. To this end we set $\widetilde{\theta}_p=(\widehat{\theta}_{j,p})_{1\leq j\leq d}$. In the sequel we will use the norm $|x|_d^2=\sum_{j=1}^{d}x_j^2$ for any vector $x=(x_j)_{1\leq j\leq d}$ from $\mathbb{R}^d$. Now we define the

shrinkage estimators as

$$\theta^*_{j,p} = (1 - g(j))\ \widehat{\theta}_{j,p}\,, \tag{15}$$

where $g(j) = (\mathbf{c}_n/|\widetilde{\theta}_p|_d)\mathbf{1}_{\{1\le j\le d\}}$,

$$\mathbf{c}_n = \frac{\underline{\varrho}_n\,(d-6)}{2\left(r + \sqrt{d\varkappa/n}\right)\,n} \quad \text{and} \quad \varkappa = \sup_{Q\in\mathcal{Q}_n} \varkappa_Q\,.$$

The positive parameter r may be dependent of n, i.e. $r = r_n$, and such that

$$\lim_{n\to\infty} n^{-\epsilon}\, r_n = 0 \quad \text{for any} \quad \epsilon > 0\,. \tag{16}$$

Now we set shrinkage estimates for S

$$S^*_\gamma(t) = \sum_{j=1}^{p} \gamma(j)\theta^*_{j,p}\psi_{j,p}(t)\,. \tag{17}$$

We compare the estimators (14) and (17) through the difference

$$\Delta_Q(S) := \mathcal{R}_Q(S^*_\gamma, S) - \mathcal{R}_Q(\widehat{S}_\gamma, S)\,.$$

Now we obtain the non asymptotic bound for this comparative risk. Let now we set

$$p_0 = \frac{\sqrt{d}\phi_* L}{\mathbf{c}_n} + 1\,,\ L = \sup_{0\le s,t\le 1} \frac{|S(t) - S(s)|}{|t-s|}\,. \tag{18}$$

Theorem 1. *Assume that the conditions* $\mathbf{D}_1$) – $\mathbf{D}_2$) *hold. Moreover, assume that the function* S *is Lipschitzian. Then for any* $p \ge p_0$

$$\sup_{Q\in\mathcal{Q}_n}\ \sup_{\|S\|\le r} \Delta_Q(S) < 0\,. \tag{19}$$

Model selection. Main results. In order to obtain a good estimate, we will use a rule to choose a weight vector $\gamma \in \Gamma$ in (17) as in [10]. We denote the improved model selection procedure

$$S^* = S^*_{\gamma^*}\,, \tag{20}$$

where γ^* is defined in [10].

Theorem 2. *Under some conditions for any* $n \ge 2$ *and* $0 < \rho < 1/2$

$$\mathcal{R}^*(S^*, S) \ \le\ \frac{1+5\rho}{1-\rho} \min_{\gamma\in\Gamma} \mathcal{R}^*_n(S^*_\gamma, S) + \frac{1}{\rho n}\,\mathbf{U}_n\,,$$

where the coefficient $\mathbf{U}_n$ *is such that for any* $\epsilon > 0$

$$\lim_{n\to\infty} \frac{\mathbf{U}_n}{n^\epsilon} = 0\,. \tag{21}$$

This sharp oracle inequality allows us to prove the asymptotic efficiency property for proposed model selection procedure.

In order to study the asymptotic efficiency we need the functional Sobolev ball $W_{k,\mathbf{r}}$ that is defined in [10]. We denote by Σ_n all estimators $\widehat{S}_n$ i.e. any $\sigma\{y_t\,,\, 0 \le t \le n\}$ measurable functions. In the sequel we

denote by Q^* the distribution of the process $(y_t)_{0\le t\le n}$ with $\xi_t = \varsigma^* w_t$, i.e. white noise model with the intensity ς^*.

Theorem 3. *Assume that $Q^* \in \mathcal{Q}_n$. The robust risk (9) admits the following lower bound*

$$\liminf_{n\to\infty} \inf_{\widehat{S}_n\in\Sigma_n} v_n^{2k/(2k+1)} \sup_{S\in W_{k,\mathbf{r}}} \mathcal{R}^*(\widehat{S}_n, S) \ge l_k(\mathbf{r}),$$

where $l_k(\mathbf{r}) = ((2k+1)\mathbf{r})^{1/(2k+1)} \left(k/(\pi(k+1))\right)^{2k/(2k+1)}$ and $v_n = n/\varsigma^*$.

We show that this lower bound is sharp in the following sense.

Theorem 4. *Assume that $Q^* \in \mathcal{Q}_n$ and there exists $\epsilon > 0$ such that $\lim_{n\to\infty} n^{5/6+\epsilon}/p = 0$. Then the robust risk of the model selection procedure (20) satisfies the following upper bound*

$$\limsup_{n\to\infty} v_n^{2k/(2k+1)} \sup_{S\in W_{k,\mathbf{r}}} \mathcal{R}^*(S^*, S) \le l_k(\mathbf{r}).$$

It is clear that these theorems imply the following efficient property.

Corollary 1. *Assume that $Q* \in \mathcal{Q}_n$. Then the model selection procedure (20) is asymptotically efficient, i.e.*

$$\lim_{n\to\infty} v_n^{2k/(2k+1)} \sup_{S\in W_{k,\mathbf{r}}} \mathcal{R}^*(S^*, S) = l_k(\mathbf{r}).$$

References

[1] Cont R., Tankov P. Financial Modelling with Jump Processes. London: Chapman & Hall, 2004.

[2] Barbu V., Beltaief S., Pergamenshchikov S. M. Robust adaptive efficient estimation for semi - Markov nonparametric regression models // Statistical inference for stochastic processes. 2019. Vol. 22, No 2. P. 187–231.

[3] Konev V.V., Pergamenshchikov S.M. Nonparametric estimation in a semimartingale regression model. Part 1. Oracle Inequalities // Tomsk State University Journal of Mathematics and Mechanics. 2009. No 7. P. 23–41.

[4] Konev V.V., Pergamenshchikov S.M. Nonparametric estimation in a semimartingale regression model. Part 2. Robust asymptotic efficiency // Tomsk State University Journal of Mathematics and Mechanics. 2009. No 8. P. 31–45.

[5] Konev V.V., Pergamenshchikov S.M. Efficient robust nonparametric in a semimartingale regression model // Annals of the Institute of Henri Poincaré. Probab. and Stat. 2012. Vol. 48, No 4. P. 1217–1244.

[6] Konev V.V., Pergamenshchikov S.M. Robust model selection for a semimartingale continuous time regression from discrete data // Stochastic processes and their applications. 2015. Vol. 125. P. 294–326.

[7] Konev V., Pergamenshchikov S. and Pchelintsev E. Estimation of a regression with the pulse type noise from discrete data // Theory Probab. Appl. 2014. Vol. 58, No 3. P. 442–457.

[8] Pchelintsev E. Improved estimation in a non-Gaussian parametric regression // Stat. Inference Stoch. Process. 2013. Vol. 16, No 1. P. 15–28.

[9] Pchelintsev E., Pergamenshchikov S. Oracle inequalities for the stochastic differential equations // Stat. Inference Stoch. Process. 2018. Vol. 21 No 2. P. 469–483.

[10] Pchelintsev E., Pergamenshchikov S. Adaptive model selection method for a conditionally Gaussian semimartingale regression in continuous time // Tomsk State University Journal of Mathematics and Mechanics. 2019. No 58. P. 14 – 31.

[11] Pchelintsev E.A., Pchelintsev V.A., Pergamenshchikov S. M. Non asymptotic sharp oracle inequality for the improved model selection procedures for the adaptive nonparametric signal estimation problem // Komunikacie. 2018. Vol. 20, No 1. P. 72–76.

[12] Pchelintsev E.A., Pchelintsev V.A., Pergamenshchikov S. M. Improved robust model selection methods for a Lévy nonparametric regression in continuous time // Journal of Nonparametric Statistics. 2019. Vol. 31, No 3. P. 612–628.

[13] Pinsker, M.S. Optimal filtration of square integrable signals in gaussian white noise // Problems of Transimission information. 1981. Vol. 17. P. 120–133.

Пчелинцев Е. А. (Томский государственный университет, Томск, 2019) **Улучшенный метод выбора модели для оценивания в семимартингальной регрессии по дискретным данным.**

Аннотация. Рассматривается задача адаптивного непараметрического оценивания функции в модели непрерывной семимартингальной регрессии по неполным наблюдениям. Рассматривается модель регрессии с шумами, определяемами негауссовскими процессами Орнштейна - Уленбека. Предложена процедура выбора модели, основанная на улучшенных взвешенных оценках наименьших квадратов. Получены точные оракульные неравенства для робастных рисков. Установлено свойство робастной эффективности в адаптивной обстановке.

Ключевые слова: улучшенное оценивание, оценка наименьших квадратов, робастный квадратичный риск, процесс Орнштейна - Уленбека, семимартингальная регрессия, выбор модели, точное оракульное неравенство, асимптотическая эффективность.