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 – 2018»

(09–11 июля 2018 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
2018

Метод выбора модели для оценивания непараметрической регрессии с шумами Леви по дискретным наблюдениям*

Повзун М. А., Пчелинцев В. А., Пчелинцев Е. А.

Томский государственный университет, Томск
e-mail: evgen-pch@yandex.ru

Аннотация

В работе рассматривается задача оценивания непараметрического периодического сигнала в непрерывной модели регрессии с шумами Леви по наблюдениям в дискретные моменты времени. Построена адаптивная процедура выбора модели на основе улучшенных взвешенных оценок наименьших квадратов. Изучены свойства предложенной процедуры. Получено точное оракульное неравенство для робастного риска, и в адаптивной постановке установлено свойство асимптотической эффективности для улучшенной процедуры выбора модели.

Ключевые слова: неасимптотическое оценивание, робастный квадратический риск, непараметрический периодический сигнал, процесс Леви, выбор модели, оракульное неравенство, асимптотическая эффективность.

Введение. В настоящее время важными задачами являются идентификация динамических систем, обработка сигналов и анализ данных в информационно-телекоммуникационных комплексах. Для описания систем широкое применение находят непрерывные модели, задавамые стохастическими дифференциальными уравнениями. В этой статье рассматривается задача статистической идентификации стохастического периодического сигнала в предположении, что шумы в уравнении задаются процессами Леви.

Рассмотрим регрессионную модель с непрерывным временем

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

где $S(\cdot)$ – неизвестная 1- периодическая функция в $\mathcal{L}_2[0,1]$, $(\xi_t)_{0\le t\le n}$ – шумовой процесс Леви, описываемый уравнением

$$\xi_t = \varrho_1 w_t + \varrho_2 z_t \quad \text{and} \quad z_t = x * (\mu - \widetilde{\mu})_t\,, \tag{2}$$

*Работа выполнена при финансовой поддержке РНФ, проект No 17-11-01049.

где ϱ_1 и ϱ_2 – неизвестные постоянные, $(w_t)_{t\geq 0}$ – винеровский процесс, $\mu(\mathrm{d}s\,\mathrm{d}x)$ является скачкообразной мерой с детерминированным компенсатором $\widetilde{\mu}(\mathrm{d}s\,\mathrm{d}x) = \mathrm{d}s\Pi(\mathrm{d}x)$, $\Pi(\cdot)$ – мера Леви, т.е. некоторая положительная мера на $\mathbb{R}_* = \mathbb{R}\setminus\{0\}$, (см., например, [2, 7]) такая, что

$$\Pi(x^2) = 1 \quad \text{и} \quad \Pi(x^6) < \infty\,. \tag{3}$$

Здесь используем обозначение $\Pi(|x|^m) = \int_{\mathbb{R}_*} |z|^m\,\Pi(\mathrm{d}z)$. Отметим, что мера Леви $\Pi(\mathbb{R}_*)$ может быть равна $+\infty$. В дальнейшем будем обозначать через Q распределение процесса $(\xi_t)_{0\leq t\leq n}$ в пространстве Скорохода $\mathbf{D}[0,n]$ и через $\mathcal{Q}_n^*$ обозначим все распределения, для которых параметры ϱ_1 и ϱ_2 удовлетворяют следующим условиям

$$0 < \underline{\varrho} \leq \varrho_1^2 \quad \text{and} \quad \sigma_Q = \varrho_1^2 + \varrho_2^2 \leq \varsigma^*\,, \tag{4}$$

где границы $\underline{\varrho}$ и ς^* функции от n, т.е. $\underline{\varrho} = \underline{\varrho}_n$ и $\varsigma^* = \varsigma_n^*$ такие, что для любого $\check{\delta} > 0$

$$\liminf_{n\to\infty} n^{\check{\delta}}\,\underline{\varrho}_n > 0 \quad \text{and} \quad \lim_{n\to\infty} n^{-\check{\delta}}\,\varsigma_n^* = 0\,. \tag{5}$$

Задача состоит в том, чтобы оценить функцию S по дискретным наблюдениям $(y_{t_j})_{0\leq j\leq np}$, $t_j = j/p$, p – частота наблюдений, зависящая от n.

В настоящей работе рассматривается проблема оценивания в адаптивной постановке, т. е. когда гладкость функции регрессии S неизвестна. Более того, также предполагаем, что распределение Q шума $(\xi_t)_{0\leq t\leq n}$ в пространстве Скорохода $\mathcal{D}[0,n]$ неизвестно. Мы знаем только, что оно принадлежит классу распределений $\mathcal{Q}_n^*$, определенному в (4)–(5). По этим причинам используется подход робастного оценивания, разработанный для непараметрических задач в [4, 9, 10]. С этой целью определим робастный риск как

$$\mathcal{R}_n^*(\widehat{S}_n, S) = \sup_{Q\in\mathcal{Q}_n^*} \mathcal{R}_Q(\widehat{S}_n, S)\,, \tag{6}$$

где $\widehat{S}_n$ – некоторая оценка, т.е. измеримая функция относительно $(y_t)_{0\leq t\leq n}$, $\mathcal{R}_Q(\cdot,\cdot)$ – квадратичный риском, определяемый как

$$\mathcal{R}_Q(\widehat{S}_n, S) := \mathbf{E}_{Q,S}\,\|\widehat{S}_n - S\|^2 \quad \text{and} \quad \|S\|^2 = \int_0^1 S^2(t)\mathrm{d}t\,. \tag{7}$$

Целью данной работы является разработка метода выбора модели для оценки непрерывного сигнала в (1) по дискретным наблюдениям. Такое предположение о дискретных наблюдениях возникает, если невозможно обеспечить непрерывное наблюдение процесса (1). Существует множество работ, посвященных аналогичным задачам непараметрического оценивания для модели регрессии (1) и других

непрерывных процессов на основе дискретных наблюдений, проведенных Хоффманом и Рейсом и Контом, Генном-Каталотом и Розенхолком проблема оценки коэффициентов диффузионного процесса по дискретным данным. Хоффманн, Мунк и Шмидт-Хибер исследуют непараметрическое оценивание коэффициентов диффузии по дискретным данным, когда наблюдение размывается аддитивным шумом. Конт и Генон-Каталот изучали проблему непараметрического оценивания для чистых скачков процесса Леви в модели (1) на основе дискретного наблюдения времени. Процедуры выбора модели обеспечивают адаптивные решения непараметрических моделей методом резких неасимптотических оракульных неравенств.

Улучшенная оценка. Пусть $(\phi_j)_{j\geq 1}$ – ортонормированный базис в $\mathbf{L}_2[0,1]$. Эти функции продолжим периодическим способом на $\mathbb{R}$, т.е. $\phi_j(t)=\phi_j(t+1)$ для любого $t\in\mathbb{R}$ и предположим, что они равномерно ограничены, т. е. для некоторой постоянной $\bar{\phi_n}$, которая может зависеть от n, $\sup_{0\leq j\leq n}\sup_{0\leq t\leq 1}|\phi_j(t)|\leq\bar{\phi_n}<\infty$.

Для оценки неизвестной функции S в (1) рассмотрим ее разложение в ряд Фурье $S(t)=\sum_{j=1}^{\infty}\theta_j\phi_j(t)$.

Соответствующие коэффициенты Фурье

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

можно оценить как

$$\widehat{\theta}_{j,p}=\frac{1}{n}\int_0^n \psi_{j,p}(t)\,\mathrm{d}y_t\,.$$

Используя здесь (2), находим

$$\widehat{\theta}_{j,p}=\bar{\theta}_j+n^{-1/2}\xi_{j,p}(n),$$

где

$$\bar{\theta}_j=\theta_j+H_{j,p},$$
$$\xi_{j,p}(n)=n^{-1/2}I_n(\psi_{j,p}),$$
$$H_{j,p}=H_{j,p}(S)=\sum_{k=1}^{p}\int_{t_{k-1}}^{t_k}\phi_j(t_k)\,(S(t)-S(t_l))\;\mathrm{d}t.$$

Теперь определим класс взвешенных оценок наименьших квадратов для $S(t)$ как

$$\widehat{S}_\lambda=\sum_{j=1}^{n}\lambda(j)\widehat{\theta}_{j,p}\phi_j\,, \tag{8}$$

где $\lambda=(\lambda(j))_{1\leq j\leq n}\in\mathbb{R}^n$ принадлежат некоторому конечному мно-

жеству Λ из $[0,\,1]^n$.

Для первых $d \le n$ коэффициентов Фурье (8) будем использовать улучшенный метод оценивания, предложенный для параметрических моделей в [13]. С этой целью положим $\widetilde{\theta}_n = (\widehat{\theta}_{j,n})_{1\le j\le d}$. В дальнейшем будем использовать норму $|x|_d^2 = \sum_{j=1}^d x_j^2$ для любого вектора $x = (x_j)_{1\le j\le d}$ из $\mathbb{R}^d$. Теперь определим улучшенные оценки коэффициентов θ_j как

$$\theta^*_{j,n} = (1 - g(j))\,\widehat{\theta}_{j,n} \quad \text{and} \quad g(j) = \frac{\mathbf{c}_n}{|\widetilde{\theta}_n|_d}\mathbf{1}_{\{1\le j\le d\}}\,,$$

где $\mathbf{c}_n$ – некоторый положительный параметр.

Введем класс взвешенных улучшенных оценок для S как

$$S^*_\lambda = \sum_{j=1}^n \lambda(j)\theta^*_{j,n}\phi_j\,. \tag{9}$$

Обозначим разность квадратичных рисков оценок (8) и (9) через $\Delta_Q(S) := \mathcal{R}_Q(S^*_\lambda, S) - \mathcal{R}_Q(\widehat{S}_\lambda, S)$. Получим следующий результат.

Теорема 1. *Пусть наблюдаемый процесс $(y_t)_{0\le t\le n}$ описывается уравнениями (1)–(2). Пусть первые $d \le n$ компонент λ равны единицы. Тогда для любого $n \ge 1$ и $r^*_n > 0$*

$$\sup_{Q\in\mathcal{Q}^*_n}\sup_{\|S\|\le r^*_n}\Delta_Q(S) \le -\mathbf{c}_n^2 + \frac{4\phi_* r^*_n d}{\sqrt{3p}}c_n\,. \tag{10}$$

Следствие 1. *В условиях Теоремы 1 1) если $p > 4\phi_* r^*_n d/\sqrt{3}c_n$, то*

$$\sup_{Q\in\mathcal{Q}^*_n}\sup_{\|S\|\le r^*_n}\Delta_Q(S) \le 0;$$

2)

$$\lim_{p\to\infty}\sup_{Q\in\mathcal{Q}^*_n}\sup_{\|S\|\le r^*_n}\Delta_Q(S) \le -\mathbf{c}_n^2.$$

Процедура выбора модели. Процедура выбора модели для неизвестной функции S в (1) будет строиться на основе семейства оценок $(S^*_\lambda)_{\lambda\in\Lambda}$. Чтобы получить адаптивную оценку, необходимо написать правило выбора весов $\lambda \in \Lambda$ в (9). Очевидно, что наилучшим способом является минимизация эмпирической квадратичной ошибки $\mathrm{Err}_n(\lambda) = \|S^*_\lambda - S\|^2$. Используя определение оценки (9) и преобразование Фурье S, имеем

$$\mathrm{Err}_n(\lambda) \;=\; \sum_{j=1}^n \lambda^2(j)(\theta^*_{j,n})^2 \;-\; 2\sum_{j=1}^n \lambda(j)\theta^*_{j,n}\,\theta_j + \|S\|^2\,. \tag{11}$$

Поскольку коэффициенты Фурье $(\theta_j)_{j\ge1}$ неизвестны, весовые коэффициенты $(\lambda_j)_{j\ge1}$ не могут быть найдены путем минимизации этой

величины. Чтобы обойти эту трудность, нужно заменить $\theta^*_{j,n}\,\theta_j$ их оценками $\widetilde{\theta}_{j,n}$. Пусть

$$\widetilde{\theta}_{j,n} = \theta^*_{j,n}\,\widehat{\theta}_{j,n} - \frac{\widehat{\sigma}_n}{n}\,,$$

где $\widehat{\sigma}_n$ – оценка предельной дисперсии $\sigma_Q = \mathbf{E}_Q\,\xi^2_{j,n}$, которая определяется следующим образом

$$\widehat{\sigma}_n = \sum_{j=[\sqrt{n}]+1}^{n} \widehat{t}^{\,2}_{j,n} \quad \text{and} \quad \widehat{t}_{j,n} = \frac{1}{n}\int_0^n \mathrm{Tr}_j(t)\mathrm{d}y_t\,. \tag{12}$$

где $(\mathrm{Tr}_j(t))_{j\geq 1}$ – тригонометрический базис в $\mathcal{L}_2[0,1]$. За такую замену в эмпирической квадратичной ошибке приходится платить некоторый штраф. Таким образом, определим платежную функцию

$$J_n(\lambda) \,=\, \sum_{j=1}^{n} \lambda^2(j)(\theta^*_{j,n})^2 \,-\, 2\sum_{j=1}^{n} \lambda(j)\,\widetilde{\theta}_{j,n} \,+\, \delta\,\widehat{P}_n(\lambda)\,, \tag{13}$$

где δ – некоторая положительная постоянная, $\widehat{P}_n(\lambda)$ – пенализационное слагаемое и

$$\widehat{P}_n(\lambda) = \frac{\widehat{\sigma}_n\,|\lambda|^2_n}{n}\,. \tag{14}$$

Определим улучшенную процедуру выбора модели как

$$S^* = S^*_{\lambda^*} \quad \text{and} \quad \lambda^* = \mathrm{argmin}_{\lambda\in\Lambda}\, J_n(\lambda)\,. \tag{15}$$

Изучены неасимптотические свойства для процедуры (15).

Теорема 2. *Для любого $n \geq 1$ и $0 < \delta < 1/5$, риск (7) оценки (15) для S удовлетворяет следующему оракульному неравенству*

$$\mathcal{R}_Q(S^*_{\lambda^*}, S) \,\leq\, \frac{1-\delta}{1-5\delta}\min_{\lambda\in\Lambda}\mathcal{R}_Q(S^*_\lambda, S) + \frac{\Psi_{Q,n}}{n\delta}$$

$$+\,\frac{10|\Lambda|_*\mathbf{E}_Q|\widehat{\sigma}_n - \sigma_Q|}{n}\,. \tag{16}$$

где $\Psi_{Q,n}/n^\varepsilon \to 0$ для любого $\varepsilon > 0$ при $n \to \infty$.

Асимптотическая эффективность. Для изучения асимптотической эффективности определим следующий функциональный шар Соболева

$$W_{k,\mathbf{r}} = \{f \in \mathbf{C}^k_p[0,1] \,:\, \sum_{j=0}^{k} \|f^{(j)}\|^2 \leq \mathbf{r}\}\,,$$

где $\mathbf{r} > 0$ и $k \geq 1$ – неизвестные параметры, $\mathbf{C}^k_p[0,1]$ – пространство k дифференцируемых 1 - периодических $\mathbb{R} \to \mathbb{R}$ функций таких,

что для любого $0 \le i \le k-1$

$$f^{(i)}(0) = f^{(i)}(1)\,.$$

Чтобы сформулировать наши асимптотические результаты, определим

$$l_*(\mathbf{r}) \;=\; ((1+2k)\mathbf{r})^{1/(2k+1)} \left(\frac{k}{\pi(k+1)}\right)^{2k/(2k+1)}\,. \qquad (17)$$

$\mathbf{H}_1$) *Предположим, что существует* $\varepsilon > 0$ *такое, что*

$$\lim_{n\to\infty} \frac{n^{5/6+\varepsilon}}{p} = 0.$$

Более того, предполагаем, что в процедуре (15) *параметр* $\delta = \delta(n)$ *– функция от* n *такая, что*

$$\lim_{n\to\infty} \delta(n) = 0 \quad \textit{и} \quad \lim_{n\to\infty} n^{\varepsilon}\,\delta(n) = \infty$$

для всех $\varepsilon > 0$.

Теорема 3. *Робастный риск* (6) *при условии* $\mathbf{H}_1$) *допускает следующую асимптотическую нижнюю границу*

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

Теорема 4. *Квадратический риск*(6) *для процедуры* S^* *при условии* $\mathbf{H}_1$) *имеет следующую асимптотическую верхнюю границу*

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

Ясно, что Теорема4 и Теорема 3 влечет

Следствие 2. *Процедура выбора модели* S^* *асимптотически эффективна, т.е.*

$$\lim_{n\to\infty} (v_n)^{\frac{2k}{2k+1}} \sup_{S\in W_{k,\mathbf{r}}} \mathcal{R}_n^*(S^*, S) \;= l_*(\mathbf{r})\,. \qquad (18)$$

Замечание 1. *Заметим, что равенство* (18) *означает, что параметр* (17) *– постоянная Пинскера (см. [14]).*

Заключение. В заключении следует подчеркнуть, что в статье разрабатывается новый метод выбора модели, основанный на улучшенных оценках наименьших квадратов. Оказывается, что эффект улучшения в непараметрической оценке, приведенный в (10) существеннее, чем в задачах оценивания параметров, поскольку улучшение точности пропорционально размеру параметра d, который стремится к бесконечности в непараметрической постановке.

Литература

1. Akaike H. A new look at the statistical model identification. *IEEE Trans. on Automatic Control* **19** (1974) 716–723.

2. Cont R., Tankov P. *Financial Modelling with Jump Processes.* Chapman & Hall, 2004.

3. Fourdrinier D., Pergamenshchikov S. (2007) Improved selection model method for the regression with dependent noise. *Annals of the Institute of Statistical Mathematics*, **59**(3), 435–464.

4. Galtchouk L.I., Pergamenshchikov S.M. (2006) Asymptotically efficient estimates for non parametric regression models. *Statistics and Probability Letters*, **76** (8), 852–860.

5. Galtchouk L., Pergamenshchikov S. (2009) Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression.*Journal of Korean Statistical Society*, **38**(4), 305–322.

6. Ibragimov I. A., Khasminskii R. Z. *Statistical Estimation: Asymptotic Theory.* Springer, New York, 1981.

7. Jacod J., Shiryaev A.N. *Limit theorems for stochastic processes.* 2nd edition, Springer, Berlin, 2002.

8. Kassam S.A. Signal detection in non-Gaussian noise. – New York: Springer-Verlag Inc., IX, 1988.

9. Konev V. V., Pergamenshchikov S. M. Efficient robust nonparametric estimation in a semimartingale regression model. *Ann. Inst. Henri Poincaré Probab. Stat.*, **48** (4), 2012, 1217–1244.

10. Konev V. V., Pergamenshchikov S. M. Robust model selection for a semimartingale continuous time regression from discrete data. *Stochastic processes and their applications*, **125**, 2015, 294 – 326.

11. Massart P. *A non-asymptotic theory for model selection.* European Congress of Mathematics, Eur. Math. Soc., Z?rich, 2005.

12. Nussbaum M. (1985) Spline smoothing in regression models and asymptotic efficiency in $\mathbf{L}_2$.- *Ann. Statist.* **13**, 984–997.

13. Pchelintsev E. (2013) Improved estimation in a non-Gaussian parametric regression. *Stat. Inference Stoch. Process.*, **16** (1), 15 – 28.

14. Pinsker M.S. (1981) Optimal filtration of square integrable signals in gaussian white noise. *Problems of Transimission information* **17**, 120–133.

Povzun M. A., Pchelintsev V. A., Pchelintsev E. A. (Tomsk State University, Tomsk, 2018) **Model selection method for estimating a nonparametric regression with Lévy noise by discrete time observations.**

Abstract. This paper considers the problem of estimating a nonparametric periodic signal in a continuous regression model with Lévy noises from discrete time observations. An adaptive model selection procedure based on improved weighted least squares estimates was constructed. The properties of the proposed procedure were studied. The sharp oracle inequality for robust risk is obtained, and in the adaptive setting the asymptotic efficiency for the improved model selection procedure is established.

Key words: non-asymptotic estimation, robust quadratic risk, nonparametric periodic signal, Lévy process, model selection, oracle inequality, asymptotic efficiency.