Универсальные оценки в задачах нелинейной и непараметрической регрессии

Ю. Ю. Линке
Институт математики им. С.Л. Соболева СО РАН
лаборатория прикладных обратных задач

Современные проблемы обратных задач
Новосибирск, 19-23 декабря

Введение

Непараметрическая регрессия

Пусть наблюдения $X_1, \ldots, X_n$ имеют структуру

$$X_i = f(z_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

- $\{z_i\}$ (регрессоры, точки дизайна) известны
- $\{\varepsilon_i\}$ — ненаблюдаемые случайные ошибки с нулевыми средними
- регрессионная функция $f(t)$ неизвестна
- задача состоит в оценивании функции $f(t)$

Нелинейная регрессия

Пусть наблюдения $X_1, \ldots, X_n$ имеют структуру

$$X_i = f(\theta, z_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

...

- регрессионная функция $f(\cdot, \cdot)$ известна
- задача состоит в оценивании параметра θ

• В задачах нелинейной регрессии асимптотически оптимальные оценки как правило задаются неявно в виде решений тех или иных уравнений.

• Для задач нелинейной регрессии весьма типична ситуация, когда имеется несколько корней того или иного уравнения, определяющего оценку (см., например,
Small C.G., Yang Z. *Multiple Roots of Estimating Functions.* (1999);
Small C.G., Wang J. *Numerical Methods for Nonlinear Estimating Equations*. (2003).

$$X_i = f(\theta, z_i) + \varepsilon_i, \qquad \mathbb{E}\varepsilon_i = 0, \quad \mathbb{E}\varepsilon_i^2 = \sigma^2 / w_i(\theta), \quad i = 1, \ldots, n.$$

Оценка квазиправдоподобия (**Heyde**, 1997) определяется здесь уравнением

$$\psi_n(t) := \sum_{i=1}^{n} w_i(t) f'(t, z_i)\big(X_i - f(t, z_i)\big) = 0.$$

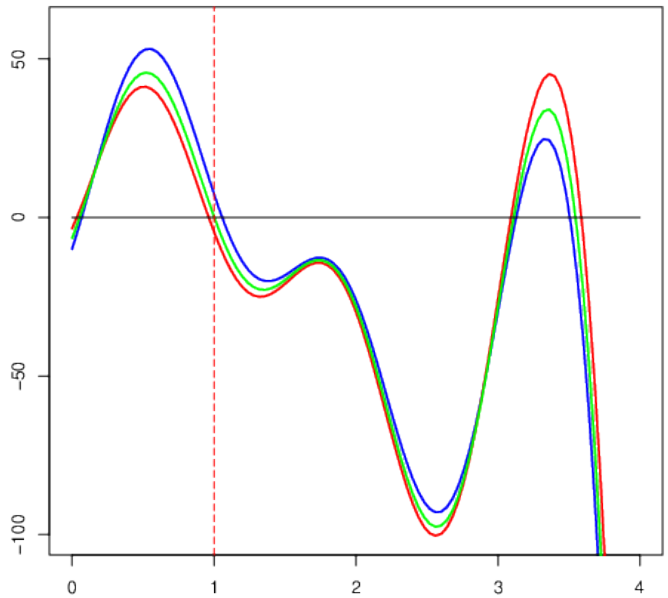


Рис.: **Графики функции $\psi_n(t)$ для некоторых f и $\{w_i\}$ и трех независимых реализаций выборки $\{X_i\}$.**
Какой из корней приближает неизвестный параметр?

Выход: одношаговые оценки (one-step estimators): $\theta_n^{**} = \theta_n^* - \psi_n(\theta_n^*)/\psi_n'(\theta_n^*)$.

Одношаговые оценки. Библиографические ссылки

Одношаговые процедуры являются весьма популярным современным инструментом статистического оценивания в различных задачах математической статистики, связанных с поиском корней тех или иных уравнений, и в последние годы интерес к одношаговому оцениванию в современной статистической литературе только нарастает.
Antoniadis, Fan (2001), **Zhao** (2000, 2001), **Bianco, Boente** (2002),
Welsh, Ronchetti (2002), **Fan, Li** (2002), **Xie, Yang** (2003),
Fan, Yao, Cai (2003), **Qian, Correa** (2003), **Yang** (2004),
Li, Marron (2005), **Giloni, Simonoff** (2005), **Jurečková, Picek** (2006),
Bianco, Boente, Martinez (2006), **Fan, Lin, Zhou** (2006),
Cai, Fan, H.Zhou, Y. Zhou (2007), **Linton, Xiao** (2007),
Li, Zheng (2007), **Zou, Li** (2008), **Johansen, Nielsen** (2009),
Chen, Li, Zhang (2010), **Bergesioa, Yohaia** (2011),
Bradic, Fan, Wang (2011), **Karunamuni, Wu** (2011), **Jureckova** (2012),
Cai, Chen, Fang (2012), **Li, Calder, Cressie** (2012), **Jureckova, Sen, Picek** (2012),
Acitas, Kasap, Senoglu, Arslan (2013), **Chen, Lin** (2013),
Hall, Ma (2014), **Fan, Xue, Zou** (2014),
Johansen, Nielsen (2016), **Wang, Wang** (2016), **Taddy** (2016),
Lipsitz, Fitzmaurice, Sinha, Hevelone, Hu, Nguyen (2017), **Jiao, Nielsen** (2017),
Ning, Liu (2017), **Huo, Huang, Ni** (2017),
Dattner, Gugushvili (2018), **Morgan** (2018),
Huang, Huo (2019), **Bradic, Guo** (2019),
Bassett, Deride (2020), **Duchesne, Micheaux, Tatsinkou** (2020),
Eisenach, Bunea, Ning, Dinicu (2020), **Fang, Zhao, Ahmed, Qu** (2020) и др.

Проблема в нелинейной регрессии: построение предварительных оценок.

Идея построения явных оценок в задачах нелинейной регрессии

Пусть наблюдения $X_1, \ldots, X_n$ представимы в виде $X_i = f(\theta, z_i) + \varepsilon_i$, где функция f известна, $\{z_i\}$ — наблюдаемые с. в. в схеме серий. Требуется оценить $\theta \in \Theta = (a, b) \subset \mathbb{R}$.

- Идея построения оценок состоит в использовании сумм специальным образом взвешенных наблюдений со структурой интегральных сумм Римана.
- При каждом фиксированном $n \geq 1$ упорядочим $z_1, \ldots, z_n$ по возрастанию: $z_{n:1} \leq \ldots \leq z_{n:n}$. Далее перенумеруем соответствующим образом отклики и погрешности, которые обозначим через X_{ni} и ε_{ni}. В этом случае $X_{ni} = f(\theta, z_{n:i}) + \varepsilon_{ni}, \quad i = 1, \ldots, n.$

($\mathbf{D}_0$) Пусть $z_i \in [c, d]$ п.н. и $\max\limits_{1 \leq i \leq n+1} \Delta z_{ni} \xrightarrow{p} 0$, где $\Delta z_{ni} = z_{n:i} - z_{n:i-1}$, $z_{n:0} = c$, $z_{n:n+1} = d$. При всех $t \in \Theta$ функция $f(t, z)$ интегрируема по Риману по z на отрезке $[c, d]$ и интеграл Римана $T(t) = \int_c^d f(t, z)dz$ — строго монотонная и непрерывная функция.

Оценку параметра θ определим равенством

$$\theta_n^* = T^{-1}\Big(\sum_{i=1}^n \Delta z_{ni} X_{ni}\Big),$$

поскольку при широких ограничениях

$$\sum_{i=1}^n \Delta z_{ni} X_{ni} = \sum_{i=1}^n \Delta z_{ni} f(\theta, z_{n:i}) + \sum_{i=1}^n \Delta z_{ni} \varepsilon_{ni} \xrightarrow{p} T(\theta) \quad , \quad \theta_n^* \xrightarrow{p} T^{-1}\big(T(\theta)\big) \equiv \theta.$$

$$X_{ni} = f(\theta, z_{n:i}) + \varepsilon_i, \quad \Delta z_{ni} = z_{n:i} - z_{n:i-1}, \quad T(t) = \int_c^d f(t,z)dz, \quad \theta_n^* = T^{-1}\Big(\sum_{i=1}^n \Delta z_{ni} X_{ni}\Big).$$

Приведем несколько примеров известных регрессионных моделей и соответствующих функций $T(\cdot)$, определяющих θ_n^*. При этом мы не задаемся вопросом явного представления обратной функции $T^{-1}(\cdot)$, поскольку ее значения легко могут быть вычислены с любой наперед заданной точностью. Более того, в общем случае не требуется чтобы и интеграл Римана $T(\cdot)$ вычислялся в элементарных функциях.

Примеры. Пусть $z_i \in [c,d]$, $c \geq 0$, $d < \infty$ и $\theta > 0$.

1) если $f(\theta, z_i) = (1+\theta z_i)^r$, $r \neq 0, -1, 1$, то $T(\theta) = \dfrac{(1+d\theta)^{r+1} - (1+c\theta)^{r+1}}{\theta(r+1)}$;
если $r < -1$, то при $c = 0$ и $d = \infty$ выполнено $T^{-1}(t) = t^{-1}|r+1|^{-1}$;

2) если $f(\theta, z_i) = 1/(1+\theta z_i)$, то $T(\theta) = \dfrac{1}{\theta}\log\dfrac{1+d\theta}{1+c\theta}$;

3) если $f(\theta, z_i) = 1/(1+e^{-\theta z_i})$, то $T(\theta) = (d-c) - \dfrac{1}{\theta}\log\dfrac{1+e^{-\theta c}}{1+e^{-\theta d}}$;

4) если $f(\theta, z_i) = 1/(\theta + z_i)$, то $T(\theta) = \log\dfrac{\theta+d}{\theta+c}$ and $T^{-1}(t) = \dfrac{d - ce^t}{e^t - 1}$;

5) если $f(\theta, z_i) = e^{-\theta z}$, то $T(\theta) = \theta^{-1}\big(e^{-\theta c} - e^{-\theta d}\big)$; $T^{-1}(t) = 1/t$ при $c = 0$ и $d = \infty$;

6) если $f(\theta, z_i) = \log(1+\theta z_i)$, $T(\theta) = \big((1+d\theta)\log(1+\theta d) - (1+c\theta)\log(1+\theta c)\big)/\theta - (d-c)$;

7) если $f(\theta, z_i) = z^\theta$ $(c, d \leq 1)$, то $T(\theta) = \big(d^{\theta+1} - c^{\theta+1}\big)/(\theta+1)$; $T^{-1}(t) = 1/t - 1$ при $c = 0$ и $d = 1$.

Многофакторные модели

$$X_i = f(\theta, \mathbf{z}_i) + \varepsilon_i, \qquad \mathbf{z}_i \in \mathcal{P} \subset \mathbb{R}^k, \quad i = 1, \ldots, n.$$

Предлагаемый подход основан на конструкции кратного интеграла Римана.

($\mathbf{D}_{0k}$) Существует такое разбиение $\mathcal{P}$ на n подмножеств $\{\mathcal{P}_i,\ i = 1, \ldots, n\}$, что каждый элемент разбиения содержит только одну точку из $\{\mathbf{z}_i\}$ (занумеруем $\mathcal{P}_i$ так, чтобы $\mathbf{z}_i \in \mathcal{P}_i$) и $\max_{i \le n} d(\mathcal{P}_i) \xrightarrow{p} 0$, где $d(A) = \sup_{\mathbf{x},\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|$ — диаметр множества, $\|\cdot\|$ — евклидова норма в $\mathbb{R}^k$.
Функция $f(\theta, \mathbf{z})$ как функция k-мерного аргумента $\mathbf{z}$ интегрируема по Риману на множестве $\mathcal{P}$ и интеграл Римана $T(t) = \int_{\mathcal{P}} f(t, \mathbf{z}) d\mathbf{z}$ строго монотонная и непрерывная функция.

Оценку θ_n^* определим равенством

$$\theta_n^* = T^{-1}\Big(\sum_{i=1}^{n} X_i \Lambda_k(\mathcal{P}_i)\Big) \xrightarrow{p} T^{-1}\big(T(\theta)\big) \equiv \theta,$$

где $\Lambda_k(\cdot)$ — мера Лебега в $\mathbb{R}^k$. Здесь мы учли, что

$$\sum_{i=1}^{n} X_i \Lambda_k(\mathcal{P}_i) = \sum_{i=1}^{n} f(\theta, \mathbf{z}_i)\Lambda_k(\mathcal{P}_i) + \sum_{i=1}^{n} \varepsilon_i \Lambda_k(\mathcal{P}_i), \qquad \sum_{i=1}^{n} f(\theta, \mathbf{z}_i)\Lambda_k(\mathcal{P}_i) \xrightarrow{p} T(\theta).$$

Результаты компьютерного моделирования

Для каждой модели генерируется 1000 независимых выборок объема $n = 100$ вида
$X_i = f_i(\theta) + \varepsilon_i$ при $f_i(\theta) = f(\theta, \mathbf{z}_i)$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2/w_i(\theta))$, $i = 1, \ldots, n$.

Одношаговые процедуры применяются для приближения оценок квазиправдоподобия, которые являются наилучшими в некотором классе и определяются здесь уравнением

$$\psi_n(t) := \sum_{i=1}^{n} \psi_i(t, X_i) = 0 \quad \text{при} \quad \psi_i(t, X_i) = w_i(t) f_i'(t)(X_i - f_i(t))$$

Таблица: Метод квазиправдоподобия. Результаты компьютерного моделирования при использовании одношаговых процедур для приближения оценок квазиправдоподобия.

	$f_i(t)$	$\sigma^2/w_i(t)$	$z_i = (z_{1i}, z_{2i})$	θ	θ_n^*	θ_n^{**}
1.	$z_{1i}t + z_{2i}\log(t)$	$(1 + tz_{1i} + z_{1i}\log(t))^{-2}$	$z_{1i} = 0.5 + \frac{i}{n}$ $z_{2i} = 1.5 - \frac{i}{n}$	2	1.895 2.381 1.746	2.067 1.966 2.182
2.	$z_{1i}t + z_{2i}t^{-1/5}$	$1/(z_{1i}t + z_{2i}t^{-1/5})$	$z_{1i} = \frac{i}{n}$ $z_{2i} = 2 - \frac{i}{n}$	3	1.804 3.941 2.074	3.204 2.928 3.017
3.	$z_{1i}t + z_{2i}\cos(t)$	$0.4(1 + tz_{1i} + z_{2i}\cos(t))^{-2}$	$z_{1i} = i/n$ $z_{2i} = 2 - i/n$	1	0.662 0.726 1.589	1.065 1.030 0.973
4.	$z_{1i}t + z_{2i}\cos(t)$	$0.25(tz_{1i} + z_{2i}\cos(t))^{-1}$	$z_{1i} = i/n$ $z_{2i} = 2 - i/n$	1	0.805 1.703	1.032 1.025

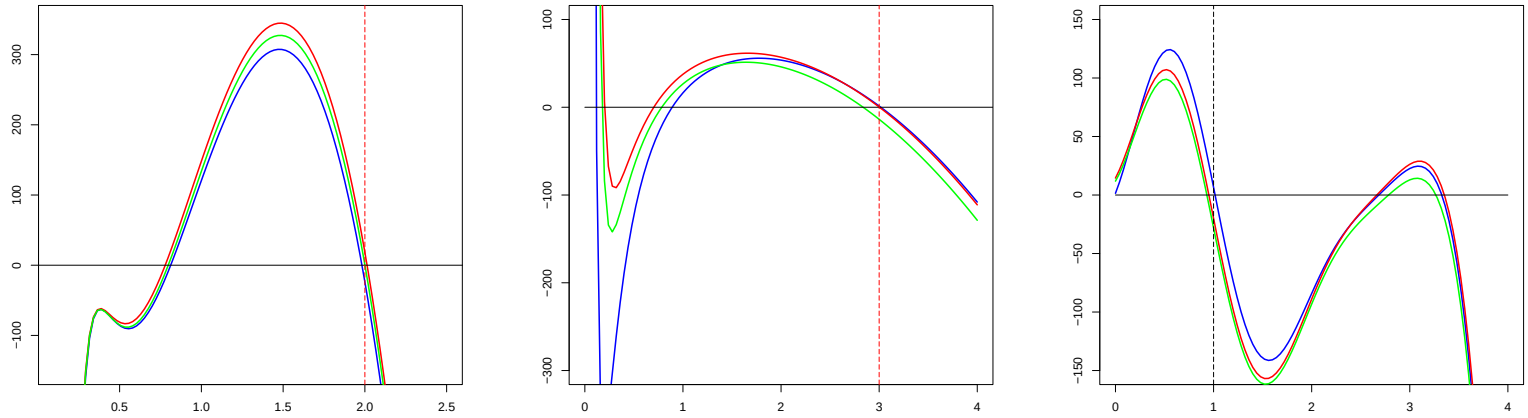


Рис.: Метод квазиправдоподобия. Иллюстрируется поведение функции $\psi_{\boldsymbol{n}}(x) = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} w_{\boldsymbol{i}}(x) f_{\boldsymbol{i}}'(x)(X_{\boldsymbol{i}} - f_{\boldsymbol{i}}(x))$ в окрестности истинного значения параметра при трех различных реализациях выборки.

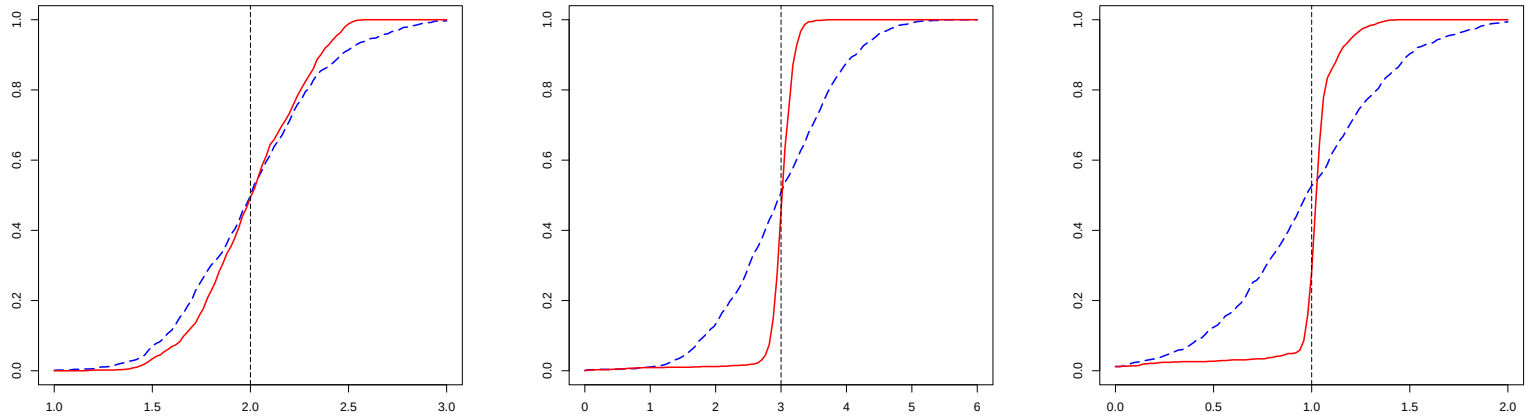


Рис.: Одношаговая аппроксимация оценок квазиправдоподобия. Эмпирические функции распределения для $\theta_{\boldsymbol{n}}^{*}$ (пунктирная линия) и $\theta_{\boldsymbol{n}}^{**}$, построенные по 1000 независимых выборок.

Непараметрическая регрессия

Пусть наблюдения $X_1, \ldots, X_n$ имеют структуру

$$X_i = f(z_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

где $\{z_i\}$ (регрессоры, точки дизайна) известны, $\{\varepsilon_i\}$ — ненаблюдаемые случайные ошибки. Задача состоит в оценивании неизвестной непрерывной регрессионной функции $f(t)$.

Выберем $h > 0$ и назовем его шириной окна. Простейший вариант оценки для f:

$$f^*_{SA}(t) = \frac{\sum_{i=1}^n X_i \cdot \mathbb{I}(t - h \le z_i \le t + h)}{\sum_{i=1}^n \mathbb{I}(t - h \le z_i \le t + h)}.$$

Недостатки $f^*_{SA}(t)$:

- функция $f^*_{SA}(t)$ имеет разрывы;
- все наблюдений X_i, для которых $z_i \in [t - h, t + h]$, учитываются в оценке в равной степени.

Оценка Надарая–Ватсона: $$f^*_{NW}(t) = \frac{\sum_{i=1}^n X_i K_h(t - z_i)}{\sum_{i=1}^n K_h(t - z_i)}, \qquad K_h(u) = h^{-1} K(u/h),$$

ядро $K(t)$ — плотность симметричного распределения (с носителем $[-1, 1]$ или $\mathbb{R}$).

Оценки Надарая–Ватсона относят к локально-постоянными, поскольку

$$f^*_{NW}(t) = \frac{\sum_{i=1}^n X_i K_h(t - z_i)}{\sum_{i=1}^n K_h(t - z_i)}, \qquad f^*_{NW}(t) = \arg\min_\theta \sum_{i=1}^n (X_i - \theta)^2 K_h(t - z_i).$$

Классический вариант локально–линейных оценок $f^*_{LL}(t)$ определяется соотношением

$$f^*_{LL}(t) = (1, 0)\arg \min_{(\theta_0, \theta_1)'} \sum_{i=1}^n (X_i - \theta_0 - \theta_1(z_i - t))^2 K_h(t - z_i).$$

(в малой окрестности точки t выполнено $f(z) \approx f(t) + f'(t)(z - t) \equiv \theta_0 - \theta_1(z - t)$)
Классические локально–полиномиальные оценки $f^*_{LP}(t)$ порядка p

$$f^*_{LP}(t) = (1, 0, \ldots, 0)\arg \min_{(\theta_0, \theta_1, \ldots, \theta_p)'} \sum_{i=1}^n (X_i - \theta_0 - \ldots - \theta_p(z_i - t)^p)^2 K_h(t - z_i).$$

Оценки Пристли–Чжао и Гассера–Мюллера

$$f^*_{PC}(t) = \sum_{i=1}^n (z_i - z_{i-1}) K_h(t - z_i) X_i,$$

$$f^*_{GM}(t) = \sum_{i=1}^n \left[\int_{s_{i-1}}^{s_i} K_h(t - z_i) \right] X_i \text{ при } s_i = \frac{z_i + z_{i-1}}{2}$$

(здесь точки дизайна неслучайны, одномерны и упорядочены).

Библиографические ссылки

• Наиболее популярные процедуры оценивания в непараметрической регрессии — оценки ядерного типа, среди которых можно выделить оценки Надарая–Ватсона, Пристли–Чжао, Гассера–Мюллера, локально–полиномиальные оценки (**Müller**, 1988; **Härdle**, 1990; **Wand, Jones**, 1995; **Fan, Gijbels**, 1996; **Fan, Yao**, 2003; **Györfi, Kohler, Krzyzak, Walk**, 2002; **Young**, 2017; **Wakefield**, 2013; **Panik**, 2009; **Klemela**, 2014; **Fahrmeir, Kneib, Lang, Marx**, 2013).

• Публикации в этой области можно условно разделить на две группы (в зависимости от условий на элементы дизайна $\{z_i\}$). К одной относятся работы с фиксированным дизайном, к другой — со случайным.

• В случае фиксированного дизайна в подавляющем большинстве работ предполагаются те или иные условия регулярности вида

$$z_i = i/n, \qquad z_i = g(i/n) + o(1/n) \qquad \text{или} \qquad \max_{i \le n}(z_i - z_{i-1}) = O(1/n)$$

(см., например, **Zhou, Zhu**, 2020; **Tang, Xi, Wu, Wang**, 2018; **Eagleson, Muller**, 1997; **Gu, Roussas, Tran**, 2007; **Beran, Feng**, 2001; **Hardle, Luckhaus**, 1984, **Wu, Chu**, 1994; **Hansen**, 2008; **Benhenni** et. al., 2010; **Ahmad, Lin**, 1984, **Georgiev**, 1989, 1990 и др.). Более общее условие $\max_{i \le n}(z_i - z_{i-1}) \to 0$ используются в работах **Wu, Wang, Balakrishnan** (2020), **He** (2019), **Yang, Yang** (2016), при этом исследуется поточечная сходимость.

• В работах, имеющих дело со случайным дизайном, рассматриваются или независимые одинаково распределеннные величины, или величины, удовлетворяющие тем или иным известным формам зависимости. Например, используются различные варианты условий перемешивания, схемы скользящих средних, ассоциированных случайных величин, марковские или мартингальные свойства, авторегрессия и др. (**Roussas**, 1990, 1991; **Györfi** et. al., 2002; **Masry**, 2005; **Hansen**, 2008; **Honda**, 2010; **Laib, Louani**, 2010; **Kulik, Lorek**, 2011; **Kulik, Wichelhaus**, 2011; **Hardle, Luckhaus**, 1984; **Mack, Silvermann**, 1982; **Muller**, 1997; **Chu, Deng**, 2003; **Linton, Jacho-Chavez**, 2010; **Li, Yang, Hu**, 2016; **Hong, Linton**, 2016; **Shen, Xie**, 2013; **Миллионщиков**, 2005; **Jiang, Mack**, 2001 **Linton, Wang**, 2016; **Chan, Wang**, 2014; **Gao** et. al., 2015; **Chen, Gao, Li**, 2012; **Karlsen, Myklebust, Tjostheim**, 2007; **Wang, Phillips**, 2009; **Chen, Li, Zhang**, 2010) и др.

• Задача равномерной аппроксимации оценок ядерного типа изучалась многими авторами (**Nadaraya**, 1970; **Devroye**, 1979; **Mack, Silvermann**, 1982; **Liero**, 1989; **Ioannides**, 1993; **Liang, Jing**, 2005; **Einmahl, Mason**, 2005; **Hansen**, 2008, **Shen, Xie**, 2013; **Wang, Chan** , 2014; **Gao, Kanaya, Li, Tjostheim**, 2015; **Li, Yang, Hu**, 2016 и др.).

Можно ли построить равномерно состоятельные оценки ядерного типа для неизвестной регрессионной функции без использования традиционных условий на элементы дизайна (при более общих, близких к минимальным и наглядных условиях на точки дизайна)?

• равномерная состоятельность оценки: $\sup_t |f^*_{n,h}(t) - f(t)| \xrightarrow{p} 0$ при $n \to \infty$

Универсальные локально-постоянные оценки

- пусть $X_i = f(z_i) + \varepsilon_i$, $i = 1, \ldots, n$, неизвестная функция $f(t)$, $t \in [0, 1]$, непрерывна;
- $\{z_i;\ i = 1, \ldots, n\}$ — это набор наблюдаемых случайных величин с, вообще говоря, неизвестными распределениями со значениями на $[0, 1]$, не обязательно независимых или одинаково распределенных; случайные величины $\{z_i\}$ могут зависеть от n (т.е. данная схема включает в себя модели с фиксированным дизайном);

По элементам $\{z_i;\ i \le n\}$ образуем вариационный ряд:

$$z_{n:1} \le \ldots \le z_{n:n}.$$

Положим $z_{n:0} = 0$, $z_{n:n+1} = 1$, $\Delta z_{ni} = z_{n:i} - z_{n:i-1}$, $i = 1, \ldots, n+1$.
Наблюдения, относящиеся к $z_{n:i}$, обозначим X_{ni}, $i = \overline{1, n}$.

Оценку $f^*_{ULC}(t)$ для $f(t)$ определим равенством

$$f^*_{ULC}(t) = \frac{\sum_{i=1}^n X_{ni} K_h(t - z_{n:i}) \Delta z_{ni}}{\sum_{i=1}^n K_h(t - z_{n:i}) \Delta z_{ni}}, \qquad K_h(t) = h^{-1} K(h^{-1} t).$$

Заметим, что $f^*_{ULC}(t) = \arg\min_\theta \sum_{i=1}^n (X_{ni} - \theta)^2 K_h(t - z_{n:i}) \Delta z_{ni}$.

$$f^*_{NW}(t) = \frac{\sum_{i=1}^n X_i K_h(t - z_i)}{\sum_{i=1}^n K_h(t - z_i)}, \qquad f^*_{ULC}(t) = \frac{\sum_{i=1}^n X_{ni} K_h(t - z_{n:i})\Delta z_{ni}}{\sum_{i=1}^n K_h(t - z_{n:i})\Delta z_{ni}}.$$

Условие на элементы дизайна, гарантирующее равномерную состоятельность $f^*_{ULC}(t)$:

(D) *При $n \to \infty$ имеет место предельное соотношение*

$$\max_{1 \le i \le n+1} \Delta z_{ni} \xrightarrow{p} 0, \qquad \Delta z_{ni} = z_{n:i} - z_{n:i-1}.$$

Например, условие (D) выполнено в следующих случаях:

- для любого регулярного дизайна;
- если $\{z_i\}$ н.о.р. и отрезок $[0, 1]$ является носителем распределения z_1 (в случае существования отделенной от нуля на $[0, 1]$ плотности распределения $\delta_n = O\left(\frac{\log n}{n}\right)$ п.н.);

...

- зависимость с.в. $\{z_i\}$ в условии (D) может быть более сильной.

- Новые оценки обладают свойством универсальности относительно структуры дизайна: дизайн может быть как фиксированным и не обязательно регулярным, так и случайным, при этом не обязательно удовлетворяющим условиям слабой зависимости.
- Относительно дизайна требуется лишь некоторое условие плотного заполнения области определения регрессионной функции, что по сути является необходимым для восстановления функции на области задания элементов дизайна.

Теорема. *Для любого фиксированного $h \in (0,1)$ с вероятностью 1 выполнено*

$$\sup_{t\in[0,1]} |f^*_{ULC}(t) - f(t)| \le \omega_f(h) + \zeta_n(h), \tag{1}$$

где $\omega_f(h) = \sup_{t,s\in[0,1]:|t-s|\le h} |f(t) - f(s)|$, а с.в. $\zeta_n(h)$ такова, что

$$\mathbb{P}(\zeta_n(h) > y) \le C_0\sigma^2 L^2 y^{-2} h^{-2} \mathbb{E}\delta_n + \mathbb{P}(\delta_n > h/(8L).$$

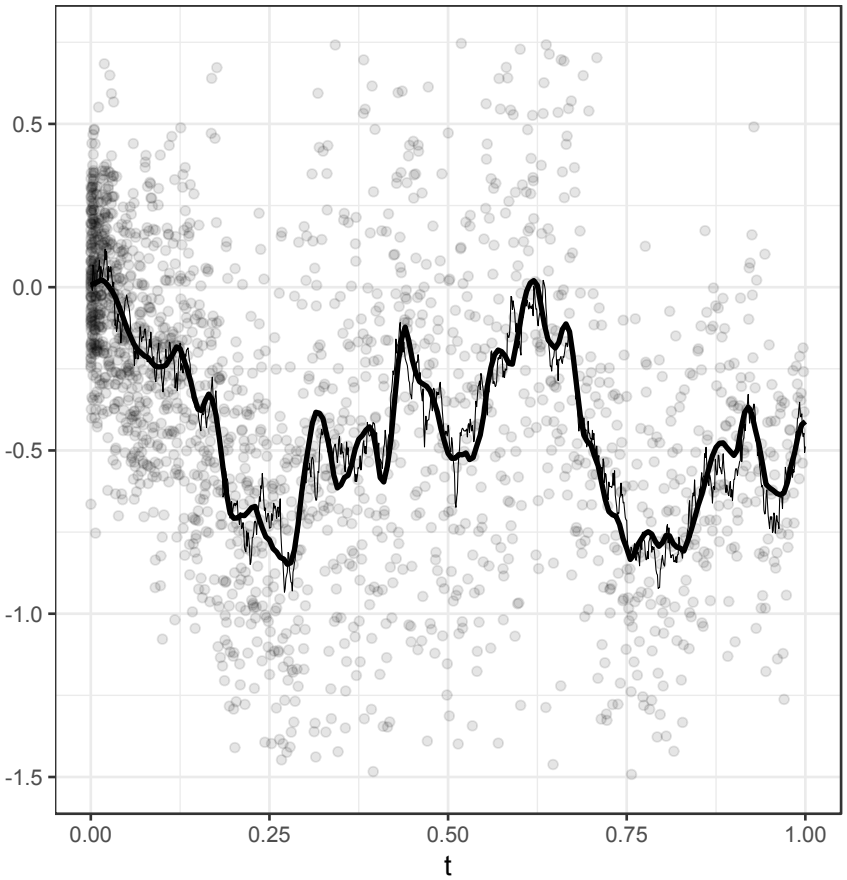
Если выполнено (D), то $\zeta_n(h) = O_p(h^{-1}(\mathbb{E}\delta_n)^{1/2})$.[1] В качестве $h = h_n$ можно взять решение уравнения $h^{-1}(\mathbb{E}\delta_n)^{1/2} = \omega_f(h)$ (фактически так выбранный размер окна уравнивает порядок малости (по h) обоих слагаемых в правой части (1)).

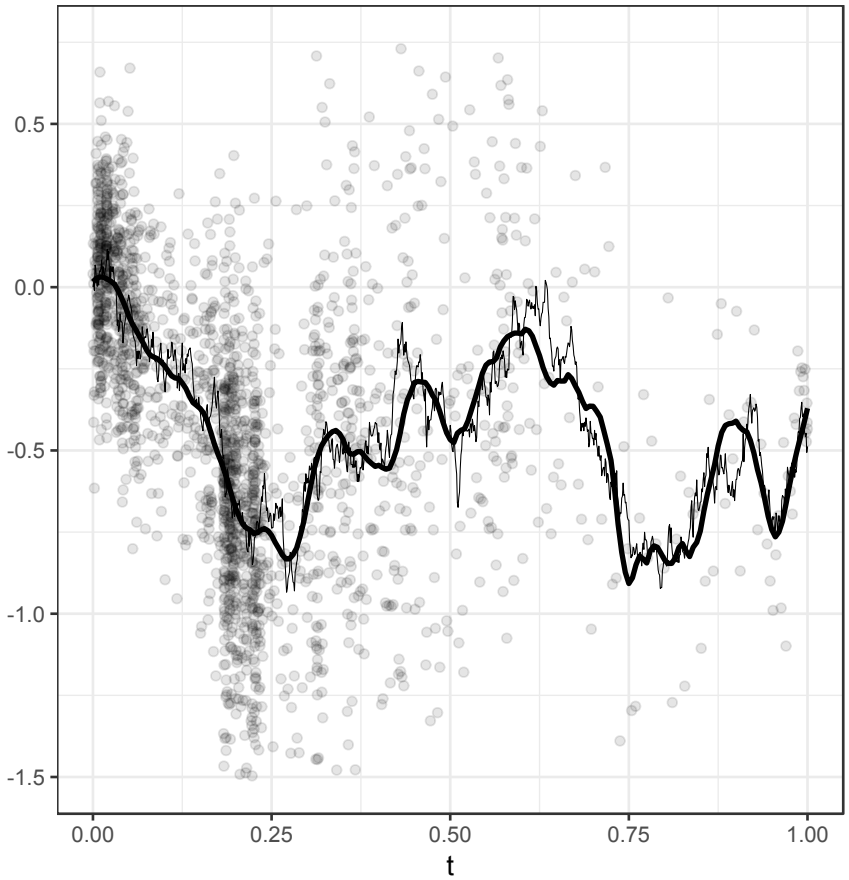
Следствие. *Пусть выполнено условие* (D) *и $\mathcal{C}$ есть произвольное подмножество равностепенно непрерывных функций в $C[0,1]$. Тогда $\gamma_n(\mathcal{C}) = \sup_{f\in\mathcal{C}} \sup_{t\in[0,1]} |f^*_{ULC}(t) - f(t)| \xrightarrow{p} 0$,*
где $h = h_n$ есть решение уравнения $h_n^{-1}(\mathbb{E}\delta_n)^{1/2} = \omega_{\mathcal{C}}(h_n)$ при $\omega_{\mathcal{C}}(h) = \sup_{f\in\mathcal{C}} \omega_f(h)$.
Кроме того, выполнено $\gamma_n(\mathcal{C}) = O_p(\omega_{\mathcal{C}}(h_n))$.

Если $\mathbb{E}\delta_n = O(1/n)$ и $\mathcal{C}$ состоит из гельд. функций f с пок. $\alpha \in (0,1]$ и унив. конст., то $h_n = O\left(n^{-\frac{1}{2(1+\alpha)}}\right)$ и $\omega_{\mathcal{C}}(h_n) = O\left(n^{-\frac{\alpha}{2(1+\alpha)}}\right)$. В частности, $\gamma_n(\mathcal{C}) = O_p(n^{-\frac{1}{4}})$ при $\alpha = 1$.

[1] $\zeta_n = O_p(\eta_n)$, если для всех чисел $M > 0$ выполнено $\limsup \mathbb{P}(|\zeta_n|/\eta_n > M) \le \beta(M)$, где $\{\eta_n\}$ – положительные (возможно, случайные) величины, $\lim_{M\to\infty} \beta(M) = 0$ ($\beta(M)$ может зависеть от K и σ^2).

Пример





- серые точки — выборка двумерных наблюдений (z_i, X_i)
- тонкая линия — график функции $f(t)$
- жирная линия — график функции $f^*_{ULC}(t)$

Сравнение с оценками Н.-В. в случае н.о.р. величин

- $\{z_i\}$ н.о.р., $\{\varepsilon_i\}$ н.о.р. и не зависят от $\{z_i\}$,
- $f(t)$ дважды непрерывно дифференцируема на $[0,1]$,
- ф.р. $F(t)$ с.в. z_1 имеет непр. дифф. положит. на $(0,1)$ плотность $p(t)$.

$$f^*_{NW}(t) = \frac{\sum_{i=1}^n X_i K_h(t-z_i)}{\sum_{i=1}^n K_h(t-z_i)}, \qquad f^*_{ULC}(t) = \frac{\sum_{i=1}^n X_{ni} K_h(t-z_{n:i})\Delta z_{ni}}{\sum_{i=1}^n K_h(t-z_{n:i})\Delta z_{ni}}.$$

Лемма 1 (Rosenblatt, 1969). Если $n \to \infty$ и $h \to 0$ так, что $h^3 n \to \infty$, то $\forall t \in (0,1)$

$$\mathrm{Bias}\, f^*_{NW}(t) = \frac{h^2\kappa_2}{2p(t)}\left(f''(t)p(t) + 2f'(t)p'(t)\right) + o(h^2), \qquad \mathbb{D} f^*_{NW}(t) \sim \frac{\sigma^2}{hnp(t)}\|K\|^2,$$

где $\kappa_2 = \int\limits_{-1}^{1} u^2 K(u)du$, $\|K\|^2 = \int\limits_{-1}^{1} K^2(u)du$.

Лемма 2. Пусть $\inf_{t\in[0,1]} p(t) > 0$ и $n \to \infty$ и $h \to 0$ так, что $(\log n)^{-1} h\sqrt{n} \to \infty$, $h^{-2}\mathbb{E}\delta_n \to 0$ и $h^{-3}\mathbb{E}\delta_n^2 \to 0$. Тогда при любом $t \in (0,1)$

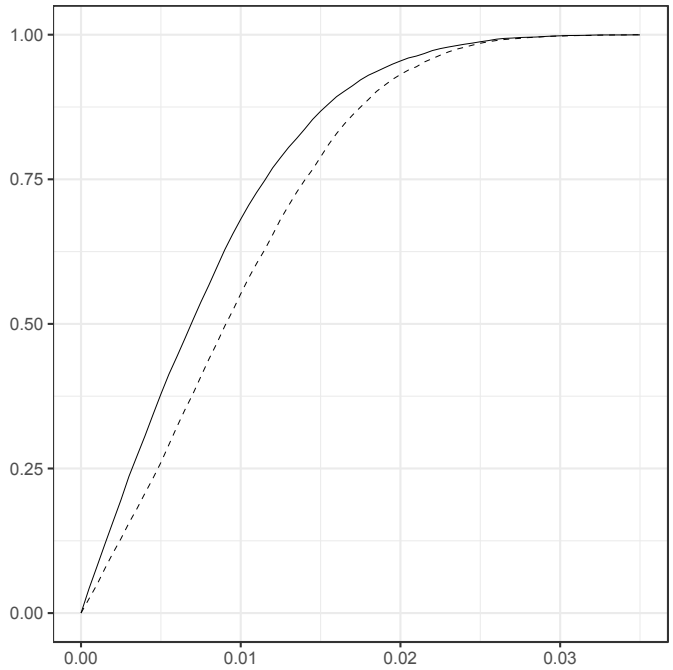
$$\mathrm{Bias}\, f^*_{ULC}(t) = \frac{h^2\kappa_2}{2} f''(t) + o(h^2), \qquad \mathbb{D} f^*_{ULC}(t) \sim \frac{2\sigma^2}{hnp(t)}\|K\|^2.$$

Вывод: Если стандартное отклонение погрешностей σ не сильно велико и

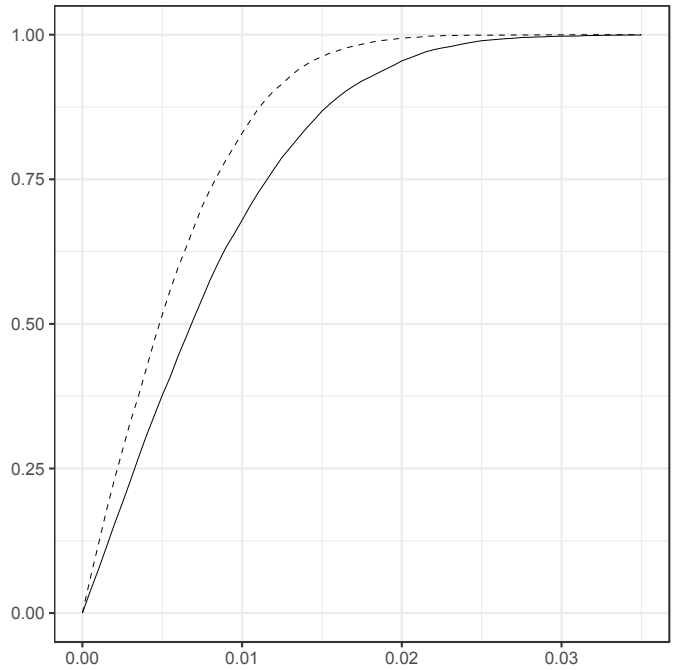
$$\left|f''(t)p(t) + 2f'(t)p'(t)\right| > \left|f''(t)p(t)\right|, \qquad (*)$$

то оценка $f^*_{ULC}(t)$ может быть лучше, чем оценка Надарая–Ватсона $f^*_{NW}(t)$.

Приведенные графики иллюстрируют влияние $(*)$ на точность аппроксимации.



$p(t) = 0.5 + t$



$p(t) = 1.5 - t$

Графики изображают эмпирические функции распределения для $|f^*_{ULC}(0.5) - f(0.5)|$ (сплошная линия) и $|f^*_{NW}(0.5) - f(0.5)|$ (пунктирная линия), построенные по 10000 симуляционных прогонок, при этом $f(t) = t^2$, $h = 0.15$, $K(\cdot)$ – ядро Епанечникова, погрешности нормально распределены со средним 0 и $\sigma = 0.1$, размер выборки $n = 1000$, плотность дизайна $p(t) = 0.5 + t$ или $p(t) = 1.5 - t$.
Для плотности $p(t) = 0.5 + t$ неравенство $(*)$ выполнено и оценка $f^*_{ULC}(t)$ оказывается лучше, нежели $f^*_{NW}(t)$. Если $p(t) = 1.5 - t$, то ситуация обратная.

Многофакторные модели

$$X_i = f(\mathbf{z}_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

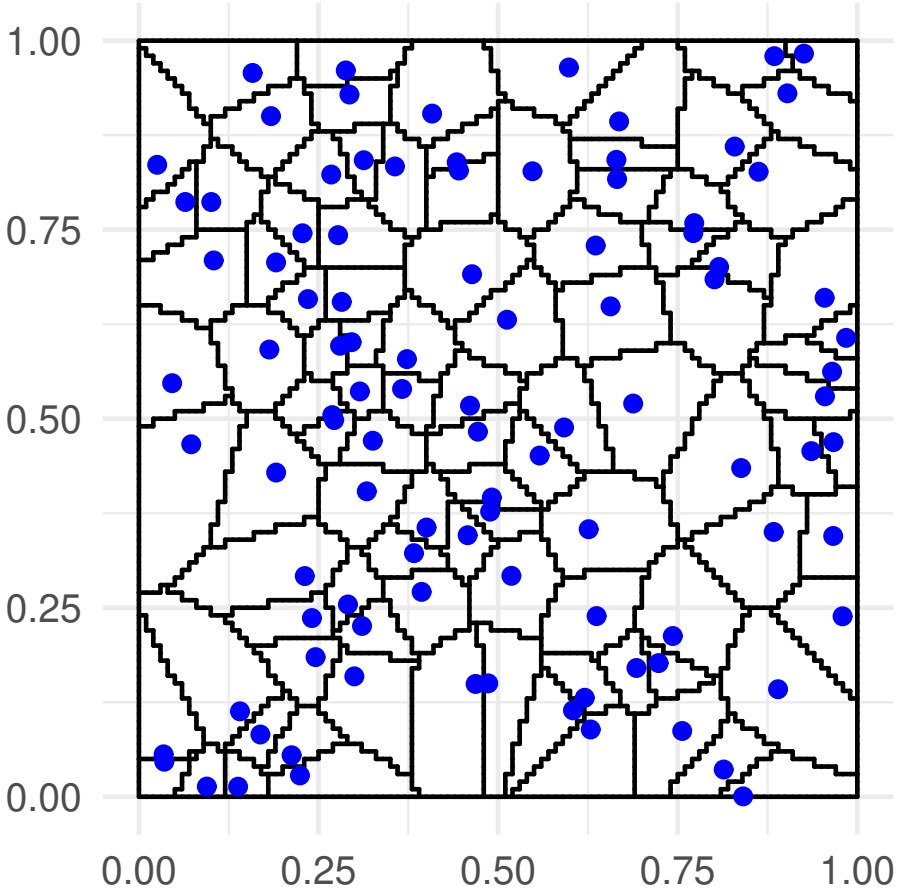
- $f : [0,1]^k \to \mathbb{R}$ неизвестна и непрерывна;
- $\{\mathbf{z}_i\}$ — набор наблюдаемых k-мерных векторов в схеме серий.

($\mathbf{D}_k$) Для каждого n существует такое разбиение множества $[0,1]^k$ на n подмножеств $\{\mathcal{P}_i,\ i = 1, \ldots, n\}$, что каждый элемент разбиения содержит только одну точку из $\{\mathbf{z}_i\}$ (занумеруем $\mathcal{P}_i$ так, чтобы $\mathbf{z}_i \in \mathcal{P}_i$) и $\max\limits_{i \le n} d(\mathcal{P}_i) \xrightarrow{p} 0$, где $d(A) = \sup\limits_{\mathbf{x},\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|$ — диаметр множества, $\|\cdot\|$ — супремальная норма в $\mathbb{R}^k$.

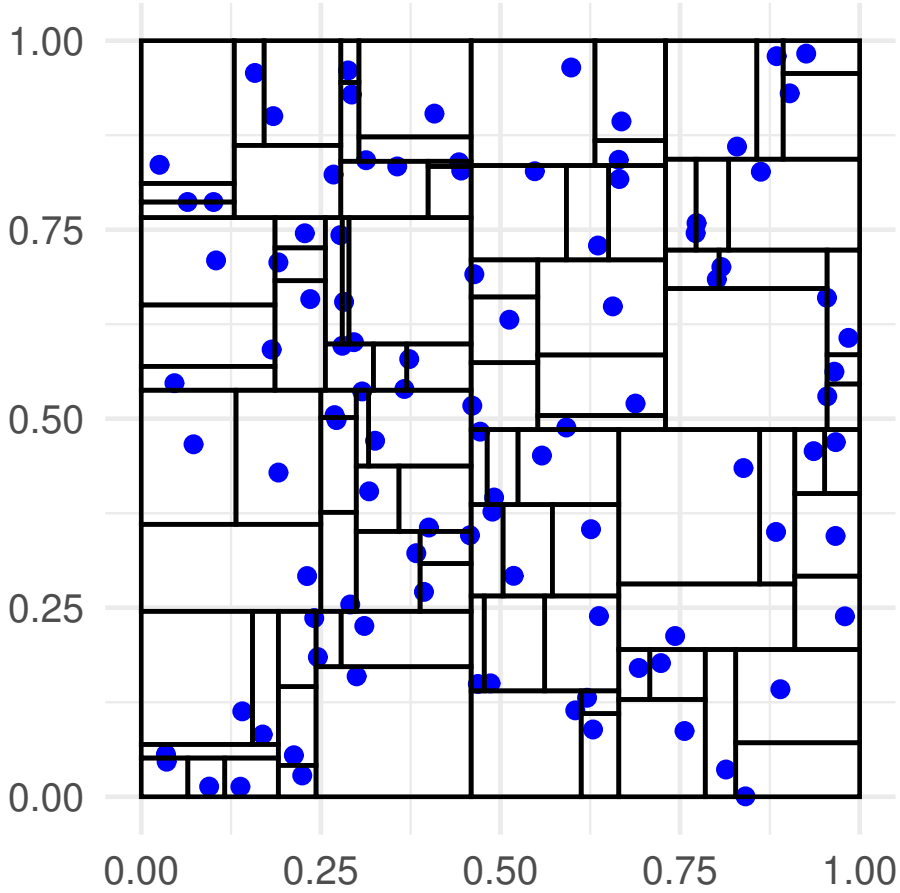
$$f^*_{ULC}(\mathbf{t}) = \frac{\sum_{i=1}^n X_i K_h(\mathbf{t} - \mathbf{z}_i)\Lambda_k(\mathcal{P}_i)}{\sum_{i=1}^n K_h(\mathbf{t} - \mathbf{z}_i)\Lambda_k(\mathcal{P}_i)},$$

где $\Lambda_k(\cdot)$ есть мера Лебега в $\mathbb{R}^k$, $K_h(\mathbf{s}) = h^{-k}K(h^{-1}\mathbf{s})$, $K(\mathbf{s})$, $\mathbf{s} \in \mathbb{R}^k$ — ядерная функция.

Построение разбиений.



Метод Вороного



Метод покоординатно-медианных сечений

Рис.: Примеры разбиения $\mathcal{P} = [0, 1]^2$

Оценивание функций среднего и ковариации случайного процесса

Рассмотрим набор $f_1(t), \ldots, f_n(t)$ независимых реализаций непрерывного случайного процесса $f(t)$, определенного на $[0,1]$. Задача состоит в оценивании функций среднего $\mu(t) = \mathbb{E}f(t)$ и ковариации $\psi(t,s) = \mathrm{cov}\{f(t), f(s)\}$ по парам наблюдений $\{(z_{ij}, X_{ij}),\ i = 1, \ldots n,\ j = 1, \ldots, m_i\}$ со структурой

$$X_{ij} = f_i(z_{ij}) + \varepsilon_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m_i.$$

- Относительно разреженного дизайна (количество точек дизайна для каждой из функций равномерно ограничено конечной константой) мы требуем лишь, чтобы вся совокупность точек дизайна образовывала измельчающееся разбиение отрезка $[0,1]$, а для плотного (количество наблюдений для каждой функции растет с ростом n) дизайна подобное условие должно быть выполнено для элементов дизайна каждой из функций.
- Новые оценки обладают свойством универсальности относительно структуры дизайна: он может быть как фиксированным и не обязательно регулярным, так и случайным, при этом не обязательно состоящим из слабо зависимых случайных величин.

Rice, Wu (2001); James, Hastie (2001); Lin, Carroll (2000), Wang (2003), Yao, Müller, Wang (2005a, 2005b); Muller (2005); Ramsay, Silverman (2005), Yao, Lee (2006); Wu, Zhang (2006); Hall, Müller, Wang (2006); Gervini (2006); Zhang, Chen (2007); Yao (2007); Degras (2008); Li, Hsing (2010), Cai, Yuan (2011); Bunea, Ivanescu, Wegkamp (2011); Cai, Yuan (2011); Ma, Yang, Carroll (2012); Cao, Yang, Todem (2012); Kim, Zhao (2013); Song, Liu, Shao, Yang (2014); Zheng, Yang, Hardle (2014); Cuevas (2014); Hsing, Eubank (2015); Cao, Wang, Li, Yang (2016); Wang, Chiou, Muller (2016); Kokoszka, Reimherr (2017); Zhang, Wang (2016, 2018); Zhou, Lin, Liang (2018); Wang J. Cao, Wang L., Yang (2020); Lin, Wang (2022).

К вопросу о состоятельности оценок Надарая–Ватсона

Пусть $X_i = f(z_i) + \varepsilon_i$, $i = 1, \ldots, n$; неизвестная функция $f(t)$, $t \in [0,1]$, непрерывна.

$$f^*_{NW}(t) = \frac{\sum_{i=1}^n X_i K_h(t - z_i)}{\sum_{i=1}^n K_h(t - z_i)} \quad \text{при} \quad K_h(\mathbf{t}) = h^{-1} K(h^{-1} t).$$

Положим $A_{n,h}(t) = \{i : |t - z_i| \le h,\ i \le n\}$.

Условие на $\{z_i\}$, обеспечивающее поточечную состоятельность оценок Н.-В.:

($\mathbf{D}_1$) При всех фиксированных $t \in [0,1]$ и $h \in (0,1)$ $\quad \#(A_{n,h}(t)) \xrightarrow{p} \infty$ при $n \to \infty$.

Условие на $\{z_i\}$, обеспечивающее равномерную состоятельность оценок Н.-В.:

($\mathbf{D}_2$) *Для всех положительных достаточно малых h выполнено*

$$\sup_{t \in [0,1]} \frac{\sup_{|s| \le h} \#^3(A_{n,h}(t+s))}{\#^4(A_{n,\delta h}(t))} \xrightarrow{p} 0.$$

Публикации

[1] Linke Yu. Yu. (2017) Asymptotic normality of one-step M-estimators based on non-identically distributed observations // **Statist. Probab. Lett.** 129, 216–221.
[2] Линке Ю. Ю. (2017) Асимптотические свойства одношаговых взвешенных M-оценок с приложениями к задачам регрессии // **Теория вероятн. и ее примен.** 62(3), 468–498.
[3] Linke Yu. Yu. (2019) Asymptotic properties of one-step M-estimators // **Communications in Statistics – Theory and Methods** 48, 4096–4118.
[4] Линке Ю. Ю., Борисов И. С. (2018) Построение явных оценок в задачах нелинейной регрессии // **Теория вероятн. и ее примен.** 63(1) 29–56.
[5] Linke Yu. Yu., Borisov I. S. (2017) Constructing initial estimators in one-step estimation procedures of nonlinear regression // **Statist. Probab. Lett.** 120(1), 87–94.
[6] Linke Yu. Yu., Borisov I. S. (2019) Toward the notion of intrinsically linear models in nonlinear regression // **Siberian Adv. Math.** 29(3), 210–216.

[7] Линке Ю. Ю. К вопросу о нечувствительности оценок Надарая–Ватсона относительно корреляции элементов дизайна. // **Теория вероятн. и ее примен.** (в печати)
[8] Linke Yu. Yu. (2022) Kernel estimators for the mean function of a stochastic process under sparse design conditions // **Siberian Adv. Math.**, 32, 269–276.
[9] Linke Yu. Yu., Borisov I. S. (2022) Insensitivity of Nadaraya–Watson estimators to design correlation // **Communications in Statistics – Theory and Methods**, 51, 6909–6918.
[10] Borisov I. S., Linke Yu. Yu., Ruzankin P.S. (2021) Universal weighted kernel-type estimators for some class of regression models // **Metrika**. 84(2), 141–166.
[11] Linke Yu., Borisov I., Ruzankin P., Kutsenko V., Yarovaya E., Shalnova S.(2022) Universal local linear kernel estimators in nonparametric regression // **Mathematics**. 10(15).