ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 56 **2020** **Вып. 2**

УДК 621.391.1 : 519.2

© 2020 г. **Г.К. Голубев**

ОБ АДАПТИВНОМ ОЦЕНИВАНИИ ЛИНЕЙНЫХ ФУНКЦИОНАЛОВ ПО НАБЛЮДЕНИЯМ В БЕЛОМ ШУМЕ

Рассматривается задача оценивания линейного функционала от неизвестного многомерного вектора по его наблюдениям в гауссовском белом шуме. В качестве семейства оценок функционала используются оценки, порождаемые проекционными оценками неизвестного вектора, и основная задача состоит в том, чтобы выбрать наилучшую оценку в этом семействе. Цель статьи объяснить и обосновать математически простую статистическую идею, которая используется при адаптивном, т.е. основанном на наблюдениях, выборе наилучшей оценки линейного функционала из заданного семейства оценок. Обсуждаются также обобщения рассматриваемой статистической модели и предлагаемого метода оценивания, которые позволяют охватить широкий класс статистических задач.

Ключевые слова: линейный функционал, белый гауссовский шум, винеровский процесс, проекционная оценка, огибающая риска, адаптивная оценка, метод Акаике, мягкое пороговое ограничение, метод главных компонент, спектральная регуляризация.

DOI: 10.31857/S0555292320020047

§ 1. Введение

В этой статье рассматривается задача оценивания линейного функционала

$$L(\theta) = \sum_{k=1}^{\infty} \theta_k$$

от неизвестного вектора $\theta = (\theta_1, \theta_2, \ldots)^\top$ по наблюдениям

$$Y_k = \theta_k + \sigma\xi_k, \quad k = 1, 2, \ldots, \tag{1}$$

где ξ_k – случайные шумы, точнее, независимые стандартные гауссовские случайные величины, а $\sigma > 0$ – уровень шума, который далее для простоты считается известным. При этом естественно предполагается, что $\theta \in \ell_2(\mathbb{Z}^+)$.

В качестве оценок $L(\theta)$ будут использоваться оценки

$$\widehat{L}(\omega; Y) = \sum_{k=1}^{\omega} Y_k, \quad \omega \in \mathbb{Z}^+, \tag{2}$$

где величина ω, называемая далее частотой среза, может выбираться на основе наблюдений $Y = \{Y_1, Y_2, \ldots\}$. За идеей использовать это семейство оценок лежит эвристическая гипотеза, что θ_k становятся малыми, начиная с некоторого $k_\circ$, и они не вносят ощутимый вклад в значение $L(\theta)$, но при этом значение $k_\circ$ неизвестно.

Скажем сразу же несколько слов об используемых в статье обозначениях. В (2) и всюду далее аргументы функций разделяются символом ";" на два класса. До этого символа находятся фактические аргументы, т.е. те, которые меняются, а после него аргументы, которые рассматриваются как "замороженные", т.е. параметры.

Основная задача в этой статье – выбрать ω так, чтобы минимизировать ошибку оценивания

$$R(\omega;\theta) = \mathbf{E}\big|L(\theta) - \widehat{L}(\omega;Y)\big|;$$

здесь и далее $\mathbf{E}$ – усреднение по мере, порожденной наблюдениями (1) при фиксированном θ.

Ответ на вопрос, почему рассматривается именно эта статистическая модель, прост. Цель этой статьи – объяснить на элементарном уровне очень простую идею, которая лежит в основе адаптивного выбора ω, сведя при этом к минимуму технические математические детали. Возможные обобщения рассматриваемой модели обсуждаются в § 4.

Очевидно, что принципиальная проблема при выборе хорошей оценки из семейства $\widehat{L}(\omega;Y)$, $\omega \in \mathbb{Z}^+$, заключается в том, что θ_k неизвестны. Ранние подходы к ее решению (см., например, [1]) основывались на предположении, что эти величины принадлежат некоторому известному множеству Θ и частота среза выбирается так, чтобы минимизировать $\sup\limits_{\theta\in\Theta} R(\omega;\theta)$. Очевидно, что с практической точки зрения такой метод является излишне пессимистичным, так как ориентируется на самые "плохие" векторы в Θ. Кроме того, гипотеза о том, что множество Θ известно точно, является мало правдоподобной с практической точки зрения. Однако с математической точки зрения значение минимаксного подхода невозможно переоценить, поскольку только он позволяет определить достаточно узкий класс оценок, в котором имеет смысл искать наилучшую. В качестве такого класса оценок можно использовать, например, проекционные оценки из (2). Конечно, это справедливо отнюдь не для всех множеств Θ, но их достаточно много, например, такими являются

$$\Theta = \left\{\theta_k : \sum_{k=1}^{\infty} a_k|\theta_k| \leqslant 1\right\}, \tag{3}$$

где a_k – некоторая возрастающая по k последовательность.

Очевидно, что если у нас нет никакой априорной информации о векторе θ, единственное, что остается, это выбирать оценку $\widehat{L}(\omega;Y)$ или, что эквивалентно, ω на основе наблюдений. Сама по себе эта идея в статистике, конечно, не нова, но ее первая простая и эффективная реализация появилась относительно недавно в [2]. Причем она касалась не оценивания линейных функционалов, а восстановления всего вектора θ при квадратичном критерии качества. В данном случае речь идет о проекционных оценках θ

$$\widehat{\theta}_k(\omega;Y) = Y_k\mathbf{1}\{k \leqslant \omega\},$$

и задача состоит в том, чтобы выбрать ω на основе наблюдений так, чтобы среднеквадратичный риск

$$r(\omega;\theta) = \mathbf{E}\sum_{k=1}^{\infty}[\theta_k - \widehat{\theta}_k(\omega;Y)]^2$$

был минимален. Ее решение основано на следующих простых соображениях:

- При фиксированной частоте среза ω

$$r(\omega;\theta) = \sum_{k=\omega+1}^{\infty} \theta_k^2 + \sigma^2\omega,$$

и если бы мы знали θ_k, то выбрали бы

$$\omega(\theta) = \arg\min\left\{ \sum_{k=\omega+1}^{\infty} \theta_k^2 + \sigma^2\omega \right\}. \tag{4}$$

- Очевидно, что

$$\sum_{k=\omega+1}^{\infty} \theta_k^2 = \|\theta\|^2 - \sum_{k=1}^{\omega} \theta_k^2,$$

и поэтому

$$\omega(\theta) = \arg\min\left\{ -\sum_{k=1}^{\omega} \theta_k^2 + \sigma^2\omega \right\}. \tag{5}$$

- Для величины $\sum\limits_{k=1}^{\omega} \theta_k^2$ можно использовать ее несмещенную оценку $\sum\limits_{k=1}^{\omega} (Y_k^2 - \sigma^2)$. Эти аргументы приводят к методу Акаике

$$\widehat{\omega}_{\mathrm{A}}(Y) = \arg\min_{\omega}\left\{ -\sum_{k=1}^{\omega} Y_k^2 + 2\sigma^2\omega \right\}. \tag{6}$$

Несмотря на простоту этой мотивации, ее строгое математическое обоснование и практически важные обобщения появились лишь спустя 20 лет в работе [3]. Общая форма этого метода часто называется принципом несмещенного оценивания риска.

Ключевым элементом в методе Акаике и его понимании является эквивалентность формул (4) и (5). Это свойство присуще исключительно задачам, в которых риск оценивания измеряется аддитивными квадратичными потерями. Рассматриваемая в этой статье задача таковой, очевидно, не является, и поэтому для ее решения нужны принципиально другие методы.

По-видимому, работа [4] была первой, в которой предлагался математически обоснованный подход к адаптивному выбору сглаживающих параметров (в нашем случае это частота среза ω) в задачах, в которых не применим принцип несмещенного оценивания риска. Эта работа была безусловно революционной в математической статистике. За ней, естественно, последовало много работ, в которых предложенный метод применялся в различных статистических моделях, и ссылки на которые мы приводить не будем ввиду их многочисленности. К сожалению, ни из оригинальной работы, ни из последующих совсем не просто извлечь простые для понимания статистические аргументы, поясняющие, почему надо делать так, а не иначе. Оптимальность предложенного метода доказывалась с помощью довольно непростых вычислений, как правило, нагруженных многочисленными техническими условиями и деталями. Поэтому совсем не удивительно, что позднее оказалось, что можно адаптивно выбирать ω несколько проще [5, 6]. Но опять же, вычленить из этих статей простые для понимания аргументы, объясняющие статистическую суть метода, довольно сложно, поскольку она скрывается в доказательствах, переполненных важными, но по сути второстепенными математическими деталями.

Цель этой статьи – объяснить на элементарном уровне без несущественных математических деталей несколько простых идей, которые позволяют адаптивно вы-

бирать частоту среза. Как мы увидим, эти идеи и их математическое обоснование оказываются не сложнее тех, которые лежат в методе несмещенного оценивания риска.

§ 2. Основные результаты

Заметим, что для риска $R(\omega;\theta)$ справедлива следующая тривиальная граница сверху:

$$R(\omega;\theta) \leqslant \mathbf{E}\left|\sum_{k=\omega+1}^{\infty}\theta_k\right| + \sigma\,\mathbf{E}\left|\sum_{k=1}^{\omega}\xi_k\right|.$$

Поэтому мы хотели бы выбрать ω с помощью наблюдений Y так, чтобы правая часть в этом неравенстве была как можно меньше. При этом понятно, что желательно решить две следующие задачи:

1. Поскольку случайные величины ξ_k ненаблюдаемы, а выбираемое ω от них зависит, то нужно ограничить сверху $\mathbf{E}\,|W(\omega)|$, где

   $$W(\omega) = \sum_{k=1}^{\omega}\xi_k,$$

   при любых ω, зависящих от ξ_k. (Для кумулятивной суммы мы использовали обозначение $W(\cdot)$, чтобы подчеркнуть, что это винеровский процесс.)
2. Так как θ_k неизвестны, то ясно, что необходимо оценить по наблюдениям абсолютную величину смещения $|B(\omega;\theta)|$, где

   $$B(\omega;\theta) = \sum_{k=\omega+1}^{\infty}\theta_k.$$

Хотя на первый взгляд эти две задачи кажутся разными, в действительности для их решения используется одна и та же идея. Она состоит в замене случайных процессов некоторыми детерминированными функциями, которые их ограничивают либо сверху, либо снизу.

Проще всего пояснить этот подход на примере вычисления верхней границы для $\mathbf{E}\,|W(\omega)|$.

Чтобы максимально упростить технические детали, будем далее считать, что ω лежит на геометрической решетке

$$\Omega_h = \{1, \omega_2, \omega_3, \ldots\},$$

где

$$\omega_{k+1} = \min\bigl\{k \in \mathbb{Z}^+ :\ k \geqslant (1+h)\omega_k\bigr\},$$

а величина $h > 0$ является фиксированной.

Отметим, что в принципе, вместо геометрической решетки можно использовать множество положительных целых чисел. Для этого нужно немного модифицировать леммы 1 и 3, в доказательстве которых эта решетка реально применяется. Сделать это несложно, если воспользоваться стандартным методом, который применяется при доказательстве закона повторного логарифма для винеровского процесса. С другой стороны, геометрическая решетка может быть реально полезной, поскольку она позволяет существенно снизить вычислительную сложность предлагаемого далее метода.

Предположим, что найдена некоторая детерминированная функция $V_h(\omega) > 0$, такая что

$$\mathbf{E} \sup_{x \in \Omega_h} \big[|W(\omega)| - V_h(\omega)\big]_+ \leqslant K_h, \tag{7}$$

где K_h – некоторая постоянная. Тогда очевидно, что для любой частоты среза $\widehat{\omega}$, зависящей от ξ_k, $k = 1, \ldots,$ выполнено неравенство

$$\mathbf{E}\,|W(\widehat{\omega})| \leqslant \mathbf{E}\,V_h(\widehat{\omega}) + K_h.$$

При этом ясно также, что чем меньше будет функция $V_h(\cdot)$, тем лучше будет эта граница.

Хотя на первый взгляд кажется, что задача вычисления минимальной функции $V_h(\cdot)$, удовлетворяющей (7) при заданной постоянной K_h, является простой, ее точное решение, по-видимому, не известно. Близкую к минимальной функции дает следующая лемма. В ней и далее для краткости будем обозначать

$$\log[1 + \log(x)] = \log^*(x).$$

Л е м м а 1. *Если*

$$V_h(t) = \sqrt{t v_h(t)}, \tag{8}$$

где

$$v_h(t) = \log(t+1) + \frac{2\log^*(t+1)}{\log(1+1/h)},$$

то неравенство (7) *выполняется с*

$$K_h = \frac{K}{h}\log\Big(1 + \frac{1}{h}\Big), \tag{9}$$

а K – универсальная постоянная.

Доказательство этой леммы и нижеследующих лемм 2, 3 приведены в § 5.

Таким образом, мы приходим к следующей верхней границе для риска:

$$R(\omega;\theta) = \mathbf{E}\big|L(\theta) - \widehat{L}(\omega;Y)\big| \leqslant \mathbf{E}\big[|B(\omega;\theta)| + \sigma V_h(\omega)\big] + \sigma K_h. \tag{10}$$

В данной статье мы будем использовать правую часть этого неравенства для выбора частоты среза, т.е. пытаться приблизиться к выбору, который сделал бы оракул, знающий все θ_k, а именно

$$\omega_\circ(\theta) = \arg\min_{\omega \in \Omega_h}\big\{|B(\omega;\theta)| + \sigma V_h(\omega)\big\}. \tag{11}$$

Тогда ясно, что нам потребуется оценка для абсолютной величины смещения $|B(\omega;\theta)|$, построенная на основе наблюдений Y. Если такая оценка $\widehat{B}(\omega;Y)$ найдена, то заменив в (11) неизвестное смещение на его оценку, придем к следующему методу выбора частоты среза:

$$\widehat{\omega}(Y) = \arg\min_{\omega \in \Omega_h}\big\{\widehat{B}(\omega;Y) + \sigma V_h(\omega)\big\}.$$

Задача оценивания $|B(\omega;\theta)|$ является ключевой в данной статье. Ее сложность связана прежде всего с тем, что построить хорошую оценку для этой величины

невозможно. Достаточно надежно можно оценивать только лишь модули конечных сумм

$$\left|\sum_{k=w}^{w'}\theta_k\right| = |L(w';\theta) - L(w;\theta)|,$$

да и то лишь в случае, когда они существенно превосходят уровень шума $\sigma\sqrt{w'-w}$.

Чтобы пояснить, как можно трансформировать этот простой факт в оценку для $|B(\omega;\theta)|$, рассмотрим следующую вспомогательную прокси-задачу. Предположим, что мы хотим минимизировать по ω функцию

$$r(\omega) = |b(\omega)| + p(\omega),$$

где $p(\omega) \geqslant 0$ – известная неубывающая функция. При этом функцию $b(\omega)$ мы не знаем полностью, а знаем только лишь величины

$$\Delta(\omega,\omega') = |b(\omega) - b(\omega')|\mathbf{1}\big\{|b(\omega) - b(\omega')| \geqslant u(\omega'-\omega)\big\}; \tag{12}$$

здесь $u(\omega) \geqslant 0$ – известная неубывающая функция. Ясно, что без ограничения общности можно считать, что $u(0) = 0$. Чтобы избежать излишних математических формальностей, будем считать для простоты, что ω принадлежит некоторому конечному множеству.

Обозначим через $\mathcal{W}^N(\omega,\omega')$ подмножество векторов $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)^\top \in \Omega_h^N$, у которых первый и последний элементы фиксированы и равны, соответственно, ω и ω', а остальные упорядочены:

$$\omega = w_1 \leqslant w_2 \leqslant \ldots \leqslant w_N = \omega'.$$

Подмножество $\mathcal{W}^N(\omega)$ определяется аналогично, за исключением того, что последний элемент в нем не фиксирован.

Кроме того, нам потребуется любая монотонная огибающая функции $|b(\omega)|$, т.е. невозрастающая функция $\bar{b}(\omega)$, такая что

$$\bar{b}(\omega) \geqslant |b(\omega)|.$$

Оценим сверху $|b(\omega)|$ с помощью следующего тривиального неравенства:

$$|b(\omega)| \leqslant \sum_{k=1}^{N-1} |b(w_{k+1}) - b(w_k)| + |b(\omega'+1)|,$$

которое справедливо для любого вектора $\boldsymbol{w} \in \mathcal{W}^N(\omega,\omega')$. Поэтому ясно, что

$$|b(\omega)| \leqslant \min_{\boldsymbol{w}\in\mathcal{W}^N(\omega,\omega')} \sum_{k=1}^{N-1} |b(w_{k+1}) - b(w_k)| + \bar{b}(\omega'+1), \tag{13}$$

и наш следующий шаг – ограничить сверху правую часть этого неравенства с помощью функций $\Delta(\cdot\,,\cdot)$ из (12).

Обозначим $[x]_+ = \max\{0, x\}$, и воспользовавшись элементарным неравенством

$$\min_x[f(x) + g(x)] \leqslant \max_x f(x) + \min_x g(x),$$

продолжим (13) следующим образом:

$$|b(\omega)| \leqslant \min_{\boldsymbol{w}\in\mathcal{W}^N(\omega,\omega')}\left\{\sum_{k=1}^{N-1}\big[|b(w_{k+1})-b(w_k)|-u(w_{k+1}-w_k)\big]_+ + \right.$$
$$\left. + \bar{b}(\omega'+1) + \sum_{k=1}^{N-1} u(w_{k+1}-w_k)\right\} \leqslant$$
$$\leqslant \max_{\boldsymbol{w}\in\mathcal{W}^N(\omega,\omega')}\left\{\sum_{k=1}^{N-1}\big[|b(w_{k+1})-b(w_k)|-u(w_{k+1}-w_k)\big]_+\right\} +$$
$$+ \min_{\boldsymbol{w}\in\mathcal{W}^N(\omega,\omega')}\left\{\bar{b}(\omega'+1) + \sum_{k=1}^{N-1} u(w_{k+1}-w_k)\right\} \leqslant$$
$$\leqslant \max_{\omega'\geqslant\omega}\max_{\boldsymbol{w}\in\mathcal{W}^N(\omega,\omega')}\left\{\sum_{k=1}^{N-1}\big[|b(w_{k+1})-b(w_k)|-u(w_{k+1}-w_k)\big]_+\right\} +$$
$$+ \min_{\omega'\geqslant 1}\big\{\bar{b}(\omega+\omega') + u(\omega')\big\}. \tag{14}$$

Заметим, что первое слагаемое в правой части этого неравенства можно выразить через величины $\Delta(\omega_{k+1},\omega_k)$, а второе – нет. Поэтому единственное, что можно сделать в такой ситуации, это минимизировать по ω функцию

$$\max_{\omega'>\omega}\big\{\widetilde{b}^N(\omega)+p(\omega)\big\},$$

где

$$\widetilde{b}^N(\omega) \stackrel{\text{def}}{=} \max_{\boldsymbol{w}\in\mathcal{W}^N(\omega)}\sum_{k=1}^{N-1}\big[|b(w_{k+1})-b(w_k)|-u(w_{k+1}-w_k)\big]_+.$$

Таким образом, приходим к следующему методу минимизации $|b(\omega)|+p(\omega)$:

$$\omega^* = \arg\min_{\omega}\big\{\widetilde{b}^N(\omega)+p(\omega)\big\}.$$

Для этого алгоритма в силу (14) справедливо

Предложение. *Для ω^* справедливо неравенство*

$$|b(\omega^*)|+p(\omega^*) \leqslant$$
$$\leqslant \min_{\omega}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\omega)}\sum_{k=1}^{N-1}|b(w_{k+1})-b(w_k)|+p(\omega)\right\} + \min_{\omega}\big\{\bar{b}(\omega+1)+u(\omega)\big\}.$$

Доказательство. При $\omega=\omega^*$ для первого слагаемого в правой части (14) справедлива тривиальная граница сверху

$$\max_{\boldsymbol{w}\in\mathcal{W}^N(\omega^*)}\sum_{k=1}^{N-1}\big[|b(w_{k+1})-b(w_k)|-u(w_{k+1}-w_k)\big]_+ \leqslant$$
$$\leqslant \max_{\boldsymbol{w}\in\mathcal{W}^N(\omega^*)}\sum_{k=1}^{N-1}\big[|b(w_{k+1})-b(w_k)|,$$

а последнее слагаемое в (14) в силу монотонности $\bar{b}(\omega)$ оценивается сверху как

$$\min_{\omega' \geqslant 1} \bigl\{\bar{b}(\omega^* + \omega') + u(\omega')\bigr\} \leqslant \min_{\omega} \bigl\{\bar{b}(\omega + 1) + u(\omega)\bigr\}. \quad \blacktriangle$$

Чтобы использовать этот подход для минимизации $|B(\omega;\theta)| + \sigma V_h(\omega)$ по наблюдениям Y, возьмем

$$b(\omega) = B(\omega;\theta), \quad p(\omega) = \sigma V_h(\omega),$$

и пусть $\bar{B}(\omega;\theta)$ – любая невозрастающая огибающая $|B(\omega;\theta)|$.

Тогда в силу сказанного выше справедлива следующая граница сверху:

$$|B(\omega;\theta)| \leqslant \max_{\boldsymbol{w} \in \mathcal{W}^N(\omega)} \sum_{s=1}^{N-1} \varphi\bigl[\Delta L(w_s;\theta); u(\Delta w_s)\bigr] + \min_{\omega} \bigl\{\bar{B}(\omega;\theta) + u(\omega)\bigr\}; \tag{15}$$

здесь

- $\Delta L(w_s;\theta) = L(w_{s+1};\theta) - L(w_s;\theta)$,
- $\Delta w_s = w_{s+1} - w_s$,
- $u(x)\colon \mathbb{Z}^+ \to \mathbb{R}^+$ – любая неубывающая функция, такая что $u(0) = 0$.

Функцию

$$\varphi(x,t) = [|x| - t]_+, \quad x \in \mathbb{R},$$

часто называют мягким пороговым ограничением (soft thresholding), а параметр $t > 0$ – порогом. В статистике она обычно возникает и используется при оценивании разреженных векторов. Эта функция обладает простыми, но полезными свойствами. Например,

$$\min_{|\xi| \leqslant t} \varphi(x + \xi; t) = \varphi(x; 2t),$$
$$\max_{|\xi| \leqslant t} \varphi(x + \xi; t) = \varphi(x; 0) = |x|.$$

Далее потребуется несколько более общий факт, обобщающий эти тождества, а именно следующая

Лемма 2. *Справедливы неравенства*

$$\varphi(x; 2t) - \varphi(\xi; t) \leqslant \varphi(x + \xi; t) \leqslant \varphi(x; 0) + \varphi(\xi; t).$$

Из вероятностных свойств мягкого порогового ограничения будет нужен только один простой результат.

Лемма 3. *Пусть $W(\cdot)$ – стандартный винеровский процесс. Тогда*

$$\mathbf{E} \sup_{\substack{\omega_2, \omega_1 \in \Omega_h \\ \omega_2 \geqslant \omega_1}} \varphi\bigl[W(\omega_2) - W(\omega_1); V_h(\omega_2 - \omega_1)\bigr] \leqslant C K_h^2,$$

где величина K_h определена в (9), *а C – некоторая константа.*

Основная идея в этой статье – использовать первое слагаемое в правой части неравенства (15) для выбора частоты среза на основе наблюдений. Взяв $u(\omega) = \sigma V_h(\omega)$ и заменив величины $\Delta L(w_s;\theta)$ их несмещенными оценками

$$\Delta\widehat{L}(w_s;\theta) = \widehat{L}(w_{s+1};Y) - \widehat{L}(w_s;Y) = \Delta L[w_s;\theta] + \sigma[W(w_{s+1}) - W(w_s)],$$

придем к

$$\widehat{\omega}^N(Y) = \arg\min_{\omega\in\Omega_h}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\omega)}\sum_{s=1}^{N-1}\varphi\big[\Delta\widehat{L}(w_s;Y);\sigma V_h(\Delta w_s)\big] + \sigma V_h(\omega)\right\}. \tag{16}$$

Оценка линейного функционала $L(\theta)$ вычисляется, естественно, как

$$\widehat{L}^N(Y) = \widehat{L}[\widehat{\omega}^N(Y);Y]. \tag{17}$$

Задача контроля риска этого метода имеет довольно простое решение. Дело в том, что с помощью лемм 2 и 3 она сводится к детерминированному случаю. Точнее, из этих результатов сразу же вытекают следующие неравенства:

$$\begin{aligned}
&\mathbf{E}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\widehat{\omega}^N)}\sum_{s=1}^{N-1}\varphi\big[\Delta\widehat{L}(w_s;\theta);\sigma V_h(\Delta w_s)\big] + \sigma V_h(\widehat{\omega}^N)\right\} \geqslant\\
&\geqslant \mathbf{E}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\widehat{\omega}^N)}\sum_{s=1}^{N-1}\varphi\big[\Delta L(w_s;\theta);2\sigma V_h(\Delta w_s)\big] + \sigma V_h(\widehat{\omega}^N)\right\} - C(N-1)\sigma K_h^2
\end{aligned}$$

и в силу (16) для любого $\omega\in\Omega_h$

$$\begin{aligned}
&\mathbf{E}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\widehat{\omega}^N)}\sum_{s=1}^{N-1}\varphi\big[\Delta\widehat{L}(w_s;\theta);\sigma V_h(\Delta w_s)\big] + \sigma V_h(\widehat{\omega}^N)\right\} \leqslant\\
&\leqslant \max_{\boldsymbol{w}\in\mathcal{W}^N(\omega)}\sum_{s=1}^{N-1}\big|\Delta L(w_s;\theta)\big| + \sigma V_h(\omega) + C(N-1)\sigma K_h^2.
\end{aligned}$$

Поэтому очевидно, что для любого $\omega\in\Omega_h$

$$\begin{aligned}
&\mathbf{E}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\widehat{\omega}^N)}\sum_{s=1}^{N-1}\varphi\big[\Delta L(w_s;\theta);2\sigma V_h(\Delta w_s)\big] + \sigma V_h(\widehat{\omega}^N)\right\} \leqslant\\
&\leqslant \max_{\boldsymbol{w}\in\mathcal{W}^N(\omega)}\sum_{s=1}^{N-1}\big|\Delta L(w_s;\theta)\big| + \sigma V_h(\omega) + CN\sigma K_h^2.
\end{aligned}$$

Отсюда и из (10) и (15) с $u(\omega) = 2\sigma V_h(\omega)$ получаем

$$\begin{aligned}
&\mathbf{E}\big|L(\theta) - \widehat{L}^N(Y)\big| \leqslant\\
&\leqslant \mathbf{E}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\widehat{\omega}^N)}\sum_{s=1}^{N-1}\varphi\big[\Delta L(w_s;\theta);2\sigma V_h(\Delta w_s)\big] + \sigma V_h(\widehat{\omega}^N)\right\} +\\
&+ \min_{\omega\in\Omega_h}\big\{\bar{B}(\omega;\theta) + 2\sigma V_h(\omega)\big\} + \sigma K_h \leqslant\\
&\leqslant \min_{\omega\in\Omega_h}\left\{\max_{\boldsymbol{w}\in\mathcal{W}^N(\omega)}\sum_{s=1}^{N-1}\big|\Delta L(w_s;\theta)\big| + \sigma V_h(\omega)\right\} +\\
&+ \min_{\omega\in\Omega_h}\big\{\bar{B}(\omega;\theta) + 2\sigma V_h(\omega)\big\} + CN\sigma K_h^2.
\end{aligned} \tag{18}$$

Более компактная, но несколько более грубая версия этого неравенства представлена в следующей теореме. Определим огибающую $\bar{B}^N(\omega;\theta)$ следующим образом:

$$\bar{B}^N(\omega;\theta) \stackrel{\text{def}}{=} \max_{\omega''>\omega'>\omega} \left\{ \max_{\boldsymbol{w}\in\mathcal{W}^N(\omega',\omega'')} \sum_{s=1}^{N-1} \bigl|\Delta L(w_s;\theta)\bigr| + |B(w'';\theta)| \right\}.$$

Теорема 1. *Для риска оценки* $\widehat{L}^N(Y)$, *определенной в* (16), (17), *справедливо неравенство*

$$\mathbf{E}\bigl|L(\theta) - \widehat{L}^N(Y)\bigr| \leqslant 3 \min_{\omega\in\Omega_h} \bigl\{\bar{B}^N(\omega;\theta) + \sigma V_h(\omega)\bigr\} + CN\sigma K_h^2. \tag{19}$$

Доказательство вытекает практически непосредственно из (18). ▲

Замечание 1. По-видимому, константу 3 в (19) можно уменьшить (сделать близкой к 1), но для этого потребуются более сложные вероятностные методы, чем используемые в этой статье.

Замечание 2. Если формально интерпретировать неравенство (19), то оно кажется в некоторой степени абсурдным, потому что чем больше N, тем хуже граница сверху, и самая лучшая граница получается при $N = 2$. На самом деле ситуация не столь очевидна. Дело в том, что эта граница является заведомо завышенной, и если бы мы попытались ее улучшить (уменьшить постоянную 3 в (19)), то увидели бы "правильную" зависимость от N. Результаты моделирования в следующем параграфе подтверждают эту гипотезу.

Замечание 3. Для огибающей $\bar{B}^N(\omega;\theta)$ справедлива простая граница сверху

$$\bar{B}^N(\omega;\theta) \leqslant \sum_{k=\omega+1}^{\infty} |\theta_k|.$$

Ее достаточно для доказательства многих классических минимаксных теорем, например, для множеств Θ, определенных в (3). Но если рассматривать байесовскую постановку задачи, т.е. считать θ_k случайными величинами с нулевым средним, то она может привести к плохой верхней границе для риска.

§ 3. Моделирование

Практическое сравнение непараметрических методов оценивания является безусловно сложной задачей, не имеющей однозначно хорошего решения. Ее трудность связана прежде всего с тем, что рассматриваемая статистическая модель описывается многомерным параметром.

Грубо говоря, подход, который наиболее часто встречается в литературе по математической статистике, состоит в том, что выбирается от двух до двенадцати многомерных параметров θ, и для них методом Монте-Карло (как правило, с небольшим объемом выборки) вычисляются риски сравниваемых методов оценок. При таком подходе очевидно, что сказать что-либо определенное о том, как поведут себя сравниваемые методы при других параметрах, довольно затруднительно.

Чтобы охватить как можно более широкий класс неизвестных параметров, в этой статье для сравнения оценок будем использовать байесовский подход. Как хорошо известно, только он позволяет сравнивать статистические методы математически. Наряду с этим неоспоримым преимуществом существенный недостаток байесовского подхода заключается в том, что он зависит от априорного распределения многомерного параметра. Поскольку очевидно, что никаких сколько-нибудь существенных

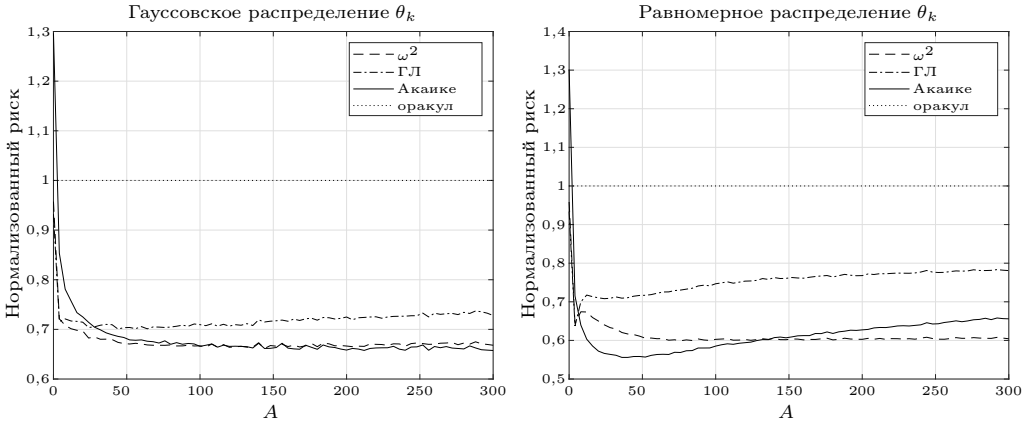

Рис. 1. Нормированные риски адаптивных частот среза при $\beta = 2$. Левый график – гауссовское распределение θ_k, правый – равномерное

аргументов в пользу выбора того или иного априорного распределения не существует, мы будем рассматривать классы априорных распределений.

Конкретно, в этой статье для сравнения статистических методов используется следующая стохастическая модель для θ_k: эти величины предполагаются независимыми и представимыми в виде

$$\theta_k = \frac{A}{k^\beta}\zeta_k,$$

где ζ_k – случайные величины, имеющие либо стандартное гауссовское, либо равномерное распределение. Параметр β характеризует скорость убывания θ_k, а A – их амплитуду. Таким образом, мы описываем бесконечномерный параметр θ с помощью пары положительных чисел $A \geqslant 0$ и $\beta > 1$. Для простоты будем считать, что $\sigma = 1$, но при этом изменять величину A, которая в этом случае играет роль отношения сигнал/шум, будем в достаточно широком диапазоне.

Далее сравним с помощью байесовского подхода три следующих метода адаптивного, т.е. основанного на наблюдениях, выбора частоты среза:

1. Метод Акаике $\widehat{\omega}_{\mathrm{A}}(Y)$ из (6).
2. Метод, предложенный Гольденшлюгером и Лепским (ГЛ) в [5].
3. Оценки $\widehat{L}^2(Y)$ и $\widehat{L}^3(Y)$ из (16), (17).

То, что метод Акаике можно применять не только для оценивания векторов, но также и линейных функционалов, было показано в [7]. При этом принципиально важно, чтобы θ_k были случайными величинами с нулевым средним.

В методе ГЛ частота среза вычисляется как

$$\widehat{\omega}_{GL}(Y) = \arg\min_w \Big\{ \max_{w'>w} \big[|L(w';Y) - L(w;Y)| - \sigma V_h(w') \big]_+ + \sigma V_h(w) \Big\}.$$

Отметим, что мы немного упростили и оптимизировали оригинальный метод из [5]. При этом была использована работа [8].

На рис. 1 показаны нормированные риски трех описанных выше методов выбора частоты среза, а именно $\widehat{\omega}_A(Y)$, $\widehat{\omega}_{GL}(Y)$ и $\widehat{\omega}^2(Y)$ как функции от амплитуды A при $\beta = 2$. В качестве нормировки был использован риск частоты среза $\omega_\circ(\theta)$ из (11), выбираемой оракулом. Другими словами, для каждого из описанных выше методов

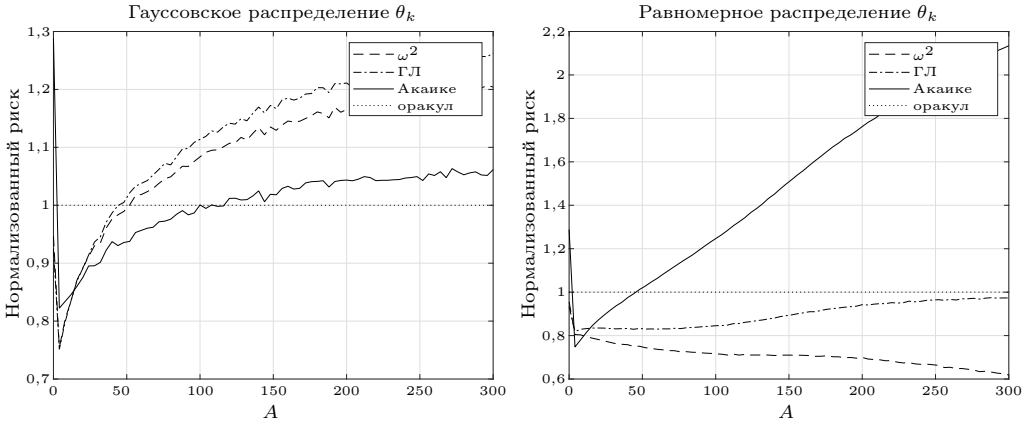

Рис. 2. Нормированные риски адаптивных частот среза при $\beta = 1{,}1$. Левый график – гауссовское распределение θ_k, правый – равномерное

на этом рисунке показаны графики функций

$$\rho(A) = \frac{\mathbf{E}\big|L(\theta) - \widehat{L}[\widehat{\omega}(Y); Y]\big|}{\mathbf{E}\big|L(\theta) - \widehat{L}[\omega_\circ(\theta); Y]\big|};$$

здесь математическое ожидание вычисляется по совместному распределению случайных величин Y_k, θ_k, $k = 1, 2, \ldots$, с помощью метода Монте-Карло с объемом выборки $3 \cdot 10^4$. Из этого рисунка, в частности, видно, что все три метода практически эквивалентны. Они в некотором смысле являются суперэффективными, так как выбирают частоту среза лучше, чем это делает оракул. К сожалению, этот эффект невозможно объяснить на основе результатов статьи. Он связан с тем, что θ_k случайны.

На рис. 2 показаны те же самые нормированные риски, но при $\beta = 1{,}1$. Мы видим, что при гауссовских θ_k метод Акаике является лучшим, что в силу результатов [7] неудивительно. В случае же равномерного распределения его оптимальность естественно теряется, так как в этом случае θ_k уже не имеют нулевого среднего.

Завершим этот параграф кратким сравнением оценок $\widehat{L}^2(Y)$ и $\widehat{L}^3(Y)$. Из правого графика на рис. 2 видно, что оценка $\widehat{L}^2(Y)$ проигрывает методу Акаике при гауссовских ζ_k и $\beta = 1{,}1$. Естественный вопрос – что будет происходить, если вместо этой оценки использовать $\widehat{L}^3(Y)$? Как видно из левого графика на рис. 3, оценка $\widehat{L}^3(Y)$ оказывается лучше, чем $\widehat{L}^2(Y)$. При этом надо отметить, что, во-первых, реальное улучшение не очень велико, а во-вторых, вычислительная сложность $\widehat{L}^3(Y)$ существенно выше. Поэтому оценки $\widehat{L}^N(Y)$ при $N > 3$ представляют скорее теоретический интерес. Для большинства практических задач $\widehat{L}^2(Y)$ и $\widehat{L}^3(Y)$ являются разумными компромиссами между статистическим качеством и вычислительной сложностью.

§ 4. Некоторые возможные обобщения

В этом параграфе кратко обсудим некоторые практически очевидные обобщения рассмотренной задачи, которые не влекут кардинального изменения предлагаемого метода оценивания.

1. Статистическая модель. Сама по себе статистическая модель (1) практически никогда не возникает в реальных статистических задачах. По существу, она пред-

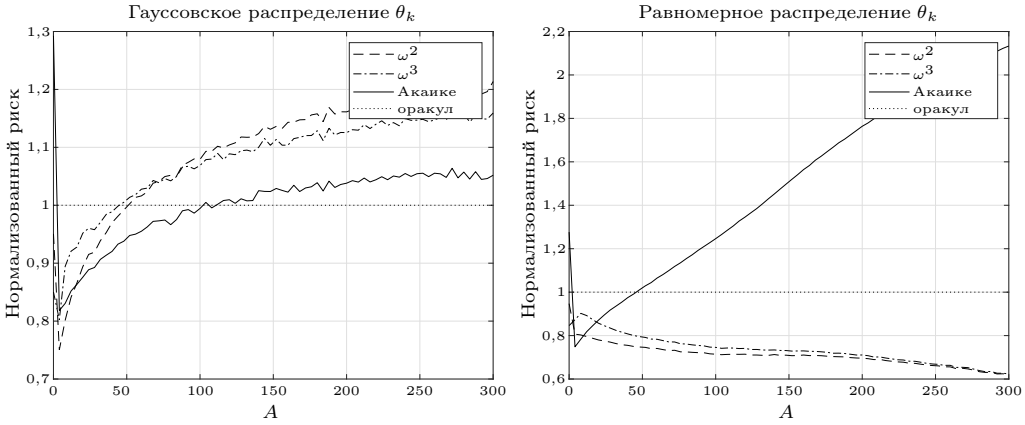

Рис. 3. Сравнение рисков оценок $\widehat{L}^2(Y)$, $\widehat{L}^3(Y)$ и метода Акаике при $\beta = 1{,}1$. Левый график – гауссовское распределение θ_k, правый – равномерное

ставляет собой так называемое спектральное представление линейных моделей большой размерности, имеющих очень широкие практические применения. Речь идет о моделях, в которых наблюдаемые данные $Z \in \mathbb{R}^n$ описываются следующей вероятностной моделью:

$$Z = X\beta + \sigma\xi, \tag{20}$$

где X – известная $(n \times m)$-матрица, $\beta \in \mathbb{R}^m$ – неизвестный вектор, а ξ – стандартный дискретный белый гауссовский шум. При этом размерности n и m, как правило, велики и таковы, что $n \geqslant m$. В частности, они могут быть равны ∞.

Эта модель приводится к (1) с помощью метода главных компонент. А именно, пусть $e_k \in \mathbb{R}^m$ и $\lambda_k \in \mathbb{R}^+$ – соответственно, собственные векторы и собственные числа матрицы $X^\top X$:

$$X^\top X e_k = \lambda_k e_k, \quad k = 1, \ldots, m.$$

Для определенности будем считать, что

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m.$$

Тогда

$$X^\top X = \sum_{k=1}^{m} \lambda_k e_k e_k^\top$$

и

$$X^\top Z = \sum_{k=1}^{m} \lambda_k e_k \langle e_k, \beta\rangle + \sigma\xi;$$

здесь $\langle \cdot, \cdot \rangle$ – скалярное произведение в $\mathbb{R}^m$.

Поэтому для $Y_k = \langle e_k, X^\top Z\rangle$ получаем следующее представление:

$$Y_k = \lambda_k \langle e_k, \beta\rangle + \sigma\xi_k, \quad k = 1, \ldots, m,$$

и положив

$$\theta_k = \lambda_k \langle e_k, \beta\rangle, \tag{21}$$

приходим к (1).

2. Линейные функционалы. Прежде всего заметим, что в доказательстве теоремы 1 практически ничего не изменится, если вместо линейного функционала $L(\theta) = \sum\limits_{k=1}^{\infty} \theta_k$ оценивать линейный функционал

$$L(\theta) = \sum_{k=1}^{\infty} l_k \theta_k, \tag{22}$$

где l_k таковы, что для всех ω

$$c\omega \leqslant \sum_{k=1}^{\omega} l_k^2 \leqslant C\omega; \tag{23}$$

здесь и далее c, C – некоторые строго положительные постоянные. При этом надо, естественно, сделать замену (см. (8))

$$V_h(t) \to V_h\Big(\sum_{k=1}^{t} l_k^2\Big).$$

Для линейной модели (20) семейство линейных функционалов из (22), (23) допускает следующее представление (см. (21)):

$$L_X(\beta) = \sum_{k=1}^{m} l_k \lambda_k \langle e_k, \beta\rangle = \sum_{k=1}^{m} l_k \langle e_k, X^\top X\beta\rangle = \langle l, X^\top X\beta\rangle,$$

где вектор $l \in \mathbb{R}^+$ таков, что

$$cs \leqslant \sum_{k=1}^{s} \langle l, e_k\rangle^2 \leqslant Cs.$$

3. Семейства оценок. Чтобы использовать проекционные оценки для оценивания линейных функционалов в модели (20), необходимо вычислять собственные векторы большой матрицы $X^\top X$. Эта и так достаточно непростая в вычислительном отношении задача станет очень сложной, если матрица $X^\top X$ окажется плохо обусловленной.

Один из возможных подходов к решению этой проблемы состоит в замене проекционных оценок на упорядоченные [3]. Это несколько более общий класс статистических методов, и в нем уже содержатся оценки, которые по своим статистическим свойствам так же хороши, как и проекционные, но не требуют применения метода главных компонент.

Очень кратко, упорядоченная оценка вектора β имеет вид

$$\widehat{\beta}(\alpha; Y) = H(X^\top X, \alpha)(X^\top X)_+^{-1} X^\top Y;$$

здесь

- $(X^\top X)_+^{-1}$ – псевдообратная матрица к $(X^\top X)$;
- $H(X^\top X, \alpha)$ – некоторая специальная матрица, которая зависит от сглаживающего параметра $\alpha \in \mathbb{R}^+$ и допускает следующее представление:

$$H_\alpha(X^\top X, \alpha) = \sum_{k=1}^{m} H(\lambda_k, \alpha) e_k e_k^\top.$$

При этом функция $H(\cdot\,,\alpha)\colon \mathbb{R}^+ \to [0,1]$ такова, что при любых фиксированных $\alpha_1, \alpha_2 \in \mathbb{R}^+$ и всех $\lambda \in \mathbb{R}^+$

$$\text{либо } H(\lambda,\alpha_1) \leqslant H(\lambda,\alpha_2), \text{ либо } H(\lambda,\alpha_1) \geqslant H(\lambda,\alpha_2).$$

Кроме того, как правило,

$$\lim_{\lambda\to 0} H(\lambda,\alpha) = 0, \quad \alpha > 0,$$
$$\lim_{\alpha\to 0} H(\lambda,\alpha) = 1, \quad \lambda > 0.$$

Оценки $\widehat{\beta}(\alpha;Y)$ порождают, естественно, следующее семейство оценок линейного функционала $L_X(\beta)$:

$$\widehat{L}_X(\alpha;Y) = \bigl\langle \widehat{\beta}(Y;\alpha), X^\top X l \bigr\rangle, \quad \alpha \in \mathbb{R}^+.$$

Метод выбора наилучшей оценки в этом семействе совершенно аналогичен рассмотренному ранее. Для его статистического анализа требуются некоторые дополнительные свойства упорядоченных оценок, которые можно найти, например, в [3] или [9].

4. Функция потерь. В этой статье качество оценивания линейного функционала измерялось величиной $\mathbf{E}\,|\widehat{L}(Y) - L(\theta)|$. Переход к другим потерям, например, к $\bigl[\mathbf{E}\,|\widehat{L}(Y) - L(\theta)|^p\bigr]^{1/p}$, $p \geqslant 1$, не влечет принципиальных и больших изменений. Единственное, что меняется, – это функция $V_h(t)$. Новая функция $V_{h,p}(t)$ будет теперь определяться из условия (см. лемму 3)

$$\mathbf{E} \sup_{\substack{\omega_2,\omega_1\in\Omega_h \\ \omega_2\geqslant\omega_1}} \Bigl\{\varphi\bigl[W(\omega_2) - W(\omega_1); V_{h,p}(\omega_2-\omega_1)\bigr]\Bigr\}^p \leqslant (CK_h^2)^p.$$

Ее вычисление очень просто, и мы его опустим.

§ 5. Доказательства

Доказательство леммы 1. Воспользуемся тривиальным неравенством

$$\mathbf{E} \sup_{t\in\Omega_h} \bigl[|W(t)| - V_h(t)\bigr]_+ \leqslant \sum_{t\in\Omega_h} \mathbf{E}\bigl[|W(t)| - V_h(t)\bigr]_+ \leqslant$$
$$\leqslant \sum_{k=1}^{\infty} \sqrt{\omega_k}\,\mathbf{E}\bigl[|\xi| - \sqrt{v_h(\omega_k)}\bigr]_+; \tag{24}$$

здесь ξ – стандартная гауссовская случайная величина.

Далее применим известное неравенство

$$\mathbf{E}[|\xi| - x]_+ \leqslant \frac{2}{x^2\sqrt{2\pi}} \exp\Bigl(-\frac{x^2}{2}\Bigr),$$

которое нетрудно проверить с помощью интегрирования по частям.

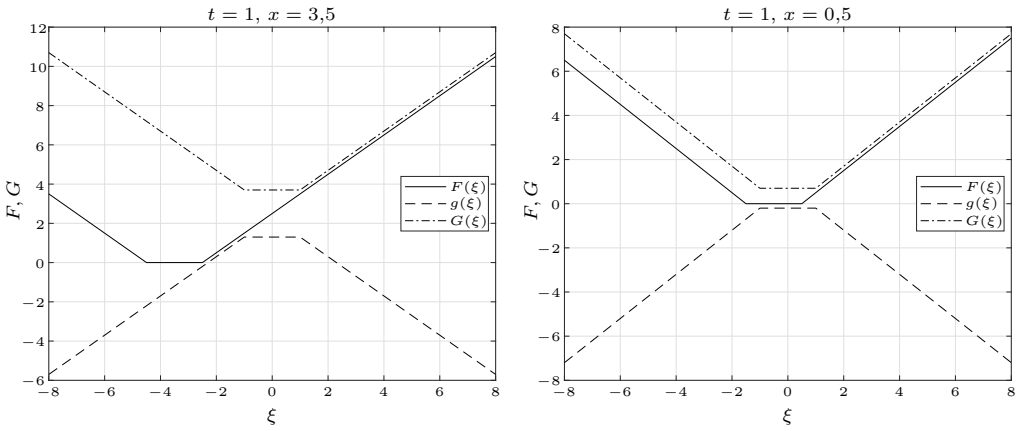

Рис. 4. Графики функций $F(\xi)$, $g(\xi)$ и $G(\xi)$ при $t = 1$, $x = 3{,}5$ и при $t = 1$, $x = 0{,}5$

Обозначив для краткости $\varepsilon = 1/\log(1 + 1/h)$, с помощью этого неравенства продолжим (24) следующим образом:

$$\mathbf{E} \sup_{t\in\Omega_h} \big[|W(t)| - V_h(t)\big]_+ \leqslant K \sum_{k=1}^{\infty} \frac{\sqrt{\omega_k}}{v_h(\omega_k)} \exp\left[-\frac{v_h(\omega_k)}{2}\right] \leqslant$$
$$\leqslant K \sum_{k=1}^{\infty} \frac{1}{[\log(\omega_k)]^{1+\varepsilon}} \leqslant K \sum_{k=1}^{\infty} \frac{1}{(hk)^{1+\varepsilon}} \leqslant \frac{K}{\varepsilon(h)^{1+\varepsilon}} \leqslant \frac{K \log(1 + 1/h)}{h}. \quad \blacktriangle \tag{25}$$

Доказательство леммы 2. Рассмотрим следующие функции:

$$F(\xi) = \varphi(x + \xi; t),$$
$$g(\xi) = \varphi(x; 2t) - \varphi(\xi; t),$$
$$G(\xi) = \varphi(x; 0) + \varphi(\xi; t).$$

Проще всего проверить, что $g(\xi) \leqslant F(\xi) \leqslant G(\xi)$, посмотрев на графики этих функций на рис. 4. ▲

Доказательство леммы 3 практически аналогично доказательству леммы 1 и приводится здесь только для полноты изложения. Пусть, как и ранее, $\varepsilon = \log^{-1}(1 + h^{-1})$. Тогда аналогично (25) получим

$$\mathbf{E} \sup_{\omega\in\Omega_h} \sup_{t>\omega} \varphi\big[W(t) - W(\omega); \sqrt{(t-\omega)v_h(t-\omega)}\big] \leqslant$$
$$\leqslant \mathbf{E} \sum_{s=1}^{\infty} \sum_{k>s} \varphi\big[W(\omega_k) - W(\omega_s); \sqrt{(\omega_k - \omega_s)v_h(\omega_k - \omega_s)}\big]_+ =$$
$$= C \sum_{s=1}^{\infty} \sum_{k=s+1}^{\infty} \frac{\sqrt{\omega_k - \omega_s}}{v_h(\omega_k - \omega_s)} \exp\left[-\frac{v_h(\omega_k - \omega_s)}{2}\right] \leqslant$$
$$\leqslant C \sum_{s=1}^{\infty} \sum_{k=s+1}^{\infty} \frac{1}{[\log(\omega_k - \omega_s)]^{1+\varepsilon}} \leqslant C \sum_{s=1}^{\infty} \sum_{k=s+1}^{\infty} \frac{1}{\log^{1+\varepsilon}[(1+h)^k - (1+h)^s)]} \leqslant$$
$$\leqslant C \left[\sum_{s=1}^{\infty} \frac{1}{(hs)^{1+\varepsilon}}\right]^2 \leqslant \frac{C}{\varepsilon^2 h^{2+2\varepsilon}} = CK_h^2. \quad \blacktriangle$$

В заключение автор хотел бы поблагодарить рецензента за сделанные замечания, способствовавшие улучшению статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Ибрагимов И.А., Хасьминский Р.З.* О непараметрическом оценивании значения линейного функционала в гауссовском белом шуме // Теория вероятн. и ее примен. 1984. Т. 29. № 1. С. 19–32.
2. *Akaike H.* Information Theory and an Extension of the Maximum Likelihood Principle // Proc. 2nd Int. Symp. on Information Theory. Tsaghkadsor, Armenia, USSR. Sept. 2–8, 1971. Budapest: Akad. Kiadó, 1973. P. 267–281.
3. *Kneip A.* Ordered Linear Smoothers // Ann. Statist. 1994. V. 22. № 2. P. 835–866.
4. *Лепский О.В.* Об одной задаче адаптивного оценивания в гауссовском белом шуме // Теория вероятн. и ее примен. 1990. Т. 35. № 3. С. 459–470.
5. *Goldenshluger A., Lepski O.* Universal Pointwise Selection Rule in Multivariate Function Estimation // Bernoulli. 2008. V. 14. № 4. P. 1150–1190.
6. *Laurent B., Ludeña C., Prieur C.* Adaptive Estimation of Linear Functionals by Model Selection // Electron. J. Stat. 2008. V. 2. P. 993–1020.
7. *Golubev Yu., Levit B.* An Oracle Approach to Adaptive Estimation of Linear Functionals in a Gaussian Model // Math. Methods Statist. 2004. V. 13. № 4. P. 392–408.
8. *Lacour C., Massart P.* Minimal Penalty for Goldenshluger–Lepski Method // Stochastic Process. Appl. 2016. V. 126. № 12. P. 3774–3789.
9. *Голубев Г.К.* Концентрации рисков выпуклых комбинаций линейных оценок // Пробл. передачи информ. 2016. Т. 52. № 4. С. 31–48.

Голубев Георгий Ксенофонтович
Институт проблем передачи информации
им. А.А. Харкевича РАН
golubev.yuri@gmail.com

Поступила в редакцию
14.02.2020
После доработки
25.02.2020
Принята к публикации
28.02.2020