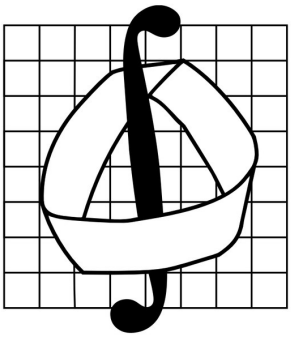
Московский государственный университет имени М. В. Ломоносова
механико-математический факультет
кафедра математической статистики и случайных процессов



Курсовая работа
студента 503 группы
Купрякова Василия Юрьевича

Непараметрическая вейвлет-оценка плотности мультипликативно зашумленных данных

Научный руководитель:
с.н.с., к.ф.-м.н.
Шкляев Александр Викторович

Москва, 2021

Оглавление

§1. Введение

В работе мы изучим задачу, которая возникает при исследовании коллоидных примесей в жидкости.

Примеси в исследуемой жидкости — это движущиеся частицы с размерами порядка 10^{-8} м. Для исследования таких примесей используется анализ траекторий наночастиц.

Жидкость просвечивают лазером, когда луч попадает на чатицу, она рассеивает свет. К микроскопу присоединена камера, которая фиксирует рассеянный свет.

Получается последовательность изображений. Для каждой частицы эта последовательность является последовательностью проекций частиц на площадь камеры. Мы можем построить векторы перемещений частиц в проекции на плоскость камеры по этим снимкам. Для отдельной частицы такие перемещения образуют броуновское движение с нулевым сносом и дисперсией $\sigma^2 = c/d$, где c — некоторая константа, а d — размер частицы.

Проблема в том, что размер частицы не связан напрямую с размером ее изображения. Наша задача — оценить распределение истинных размеров частиц по размерам на снимках.

Будем изучать равносильную задачу: оценить распределение σ^2. Рассмотрим n случайно выбранных частиц $E_1, \ldots, E_n$. Обозначим дисперсии для их движения как $\sigma_1^2, \ldots, \sigma_n^2$. Для i-й частицы у нас есть два $k(i)$-мерных вектора перемещений: $A_i^1, \ldots, A_i^{k(i)}$ по оси x и $A_i^{k(i)+1}, \ldots, A_i^{2k(i)}$ по оси y. Мы будем рассматривать только частный случай, когда все $k(i)$ равны k, а σ_i^2 непрерывна.

$A_i^1, \ldots, A_i^{2k}$ условно независимы при условии σ_i^2 и имеют условное распределение $\mathcal{N}(0, \sigma_i^2)$. Дальше вместо выборки $A_i^1, \ldots, A_i^{2k}$ будем рассматривать достаточную статистику $Z_i = \sum\limits_{j=1}^{2k} \left(A_i^j\right)^2$. Заметим далее, что $Z_i = \sigma_i^2 Y_i$, где $Y_i \sim \chi_{2k}^2$. При этом, Y_i независимы и не зависят от дисперсии σ_i^2.

Обозначим $X_i = \sigma_i^2$. Тогда задачу можно сформулировать так: $X_1, \ldots, X_n$ — независимые одинаково распределенные непрерывные случайные величины с неизвестным распределением и положительным носителем; $Y_1, \ldots, Y_n$ — независимые от них н.о.р. с.в. с распределением χ_{2k}^2; $Z_1, \ldots, Z_n = X_1 Y_1, \ldots, X_n Y_n$ — наблюдаемые случайные величины; а сама задача — по наблюдениям $Z_1, ..., Z_n$ оценить распределение X_1.

§2. Сведение задачи к вычислению обратного преобразования Лапласа

Есть случайные величины X, Y, Z. Мы не знаем распределение X, знаем распределение Y и наблюдаем Z. Кроме того, известно, что $Z = XY$, и что все величины непрерывны. Нужно оценить распределение X.

Мы будем использовать вейвлет «Mexican hat», потому что он прост и непрерывен. Его формула:

$$\psi(t) = \frac{2}{\sqrt{3}\pi^{1/4}}(1 - t^2)e^{-t^2/2}.$$

Определим элементы фрейма:

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}}\psi\left(\frac{t}{2^m} - n\right) = \frac{1}{\sqrt{2^m}}\frac{2}{\sqrt{3}\pi^{1/4}}\left(1 - \left(\frac{t}{2^m} - n\right)^2\right)e^{-\left(\frac{t}{2^m} - n\right)^2/2}.$$

Рассмотрим случай $Y \sim \chi^2_{2k}$; $X > \delta > 0$, абсолютно непрерывен. Плотность Y:

$$\chi^2_{2k} \sim \frac{1}{2^k}\frac{1}{\Gamma(k)}x^{k-1}e^{-x/2}.$$

Будем строить функции $g_{m,n}$ такие, что $\mathrm{E}\, g_{m,n}(Z) = \mathrm{E}\, \psi_{m,n}(X)$. Заметим, что достаточно выполнения:

$$\forall x \in \operatorname{Im} X \quad \mathrm{E}\, g_{m,n}(xY) = \psi_{m,n}(x).$$

Заменим мат. ожидание преобразованием Лапласа и раскроем $\psi_{m,n}$:

$$\left(\frac{1}{2x}\right)^k \frac{1}{\Gamma(k)}L_z\left[g_{m,n}(z)z^{k-1}\right]\left(\frac{1}{2x}\right) = \left(\frac{1}{\sqrt{2}}\right)^m \psi_{m,n}\left(\frac{x}{2^m} - n\right).$$

Сделаем замену $u = \frac{1}{2x}$:

$$u^k \frac{1}{\Gamma(k)}L_z\left[g_{m,n}(z)z^{k-1}\right](u) = \left(\frac{1}{\sqrt{2}}\right)^m \psi_{m,n}\left(\frac{1}{2^{m+1}u} - n\right).$$

Используя обратное преобразование Лапласа, найдем $g_{m,n}(t)$:

$$u^k \frac{1}{\Gamma(k)} L_z \left[g_{m,n}(z) z^{k-1} \right](u) = \left(\frac{1}{\sqrt{2}} \right)^m \psi_{m,n} \left(\frac{1}{2^{m+1} u} - n \right)$$

$$L_z \left[g_{m,n}(z) z^{k-1} \right](u) = \frac{\Gamma(k)}{u^k} \left(\frac{1}{\sqrt{2}} \right)^m \psi_{m,n} \left(\frac{1}{2^{m+1} u} - n \right)$$

$$g_{m,n}(t) t^{k-1} = L_u^{-1} \left[\frac{\Gamma(k)}{u^k} \left(\frac{1}{\sqrt{2}} \right)^m \psi_{m,n} \left(\frac{1}{2^{m+1} u} - n \right) \right](t)$$

$$g_{m,n}(t) = \frac{1}{t^{k-1}} \frac{\Gamma(k)}{\sqrt{2^m}} L_u^{-1} \left[\frac{1}{u^k} \psi_{m,n} \left(\frac{1}{2^{m+1} u} - n \right) \right](t).$$

Таким образом мы получили выражение для $g_{m,n}(t)$. Далее мы выразим его через ряды, используя формулу Меллина и основную теорему о вычетах.

§3. Альтернативный подход к задаче

Вспомним наше изначальное интегральное уравнение:

$$\psi_{m,n}(x) = \int\limits_0^\infty g(xy) f_Y(y) dy.$$

Преобразуем интеграл, чтобы интегрирование было по xy:

$$\psi_{m,n}(x) = \int\limits_0^\infty g(xy) f_Y(\frac{xy}{x}) d\frac{xy}{x} = \int\limits_0^\infty \int\limits_0^\infty g(z) \frac{1}{x} f_Y(\frac{z}{x}) dz.$$

Таким образом мы получили интегральное уравнение Фредгольма первого рода:

$$\psi_{m,n}(x) = \int\limits_0^\infty K(x,z) g(z) dz.$$

Дальше мы будем использовать равномерную сетку $\left[\frac{1}{n_x}, \ldots, \frac{l_x n_x}{n_x}\right]$ для x, $\left[\frac{1}{n_z}, \ldots, \frac{l_z n_z}{n_z}\right]$ для z и дискретизируем наше уравнение. Получаем:

$$\psi_{m,n}[x] = \int\limits_0^\infty K[x,z] g[z] dz = \frac{1}{n_z} \sum_{p=1}^{l_z n_z} g\left(\frac{p}{n_z}\right) K\left[x, \frac{p}{n_z}\right].$$

Таким образом, мы получили систему линейных уравнений. Запишем их в матричном виде:

$$\boldsymbol{\psi}_{m,n} = \frac{1}{n_z} \boldsymbol{K} \boldsymbol{g}.$$

Увеличим матрицу K, чтобы добавить регуляризацию.

$$\tilde{\boldsymbol{K}} = \begin{pmatrix} \boldsymbol{K} \\ \alpha \boldsymbol{E} \end{pmatrix}.$$

И соответствующий $\tilde{\boldsymbol{f}}$:

$$\tilde{\boldsymbol{f}} = \begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{0} \end{pmatrix}.$$

И будем использовать МНК-оптимизацию. Получаем:

$$\boldsymbol{g_*} = \arg\min_{\boldsymbol{g}} \|\tilde{\boldsymbol{K}}\boldsymbol{g} - \boldsymbol{f}\|.$$

3.1 Градиентный спуск

Вместо процедур для решения МНК-задачи мы можем использовать метод градиентного спуска. Будем использовать матричное представление

$$\boldsymbol{\psi}_{m,n} = \frac{1}{n_z}\boldsymbol{K}\boldsymbol{g}.$$

Тогда можно ввести функцию потери $L(\boldsymbol{\psi_{m,n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{m,n}})$, где $\hat{\boldsymbol{\psi}}_{\boldsymbol{m,n}} = \boldsymbol{K}\hat{\boldsymbol{g}}_{\boldsymbol{m,n}}$, а $\hat{\boldsymbol{g}}_{\boldsymbol{m,n}}$ — оценка для $\boldsymbol{g_{m,n}}$.

В частности, будем рассматривать следующие функции потерь:

- $l1$-потеря: $L(\boldsymbol{x}, \boldsymbol{y}) = \|x - y\|_1$;
- $l2$-потеря: $L(\boldsymbol{x}, \boldsymbol{y}) = \|x - y\|_2$;
- функция потери Хьюбера:

$$L(x, y) = \begin{cases} \frac{1}{2}(x-y)^2, \text{ при } |x-y| \leqslant 1 \\ |x-y| - \frac{1}{2}, \text{ при } |x-y| > 1 \end{cases}.$$

$$L(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{k} L(x_i, y_i)$$

Для каждой из них будем использовать L_1- или L_2-регуляризацию:

$$\tilde{L}(\boldsymbol{\psi_{m,n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{m,n}}) = L(\boldsymbol{\psi_{m,n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{m,n}}) + \|\boldsymbol{g_{m,n}} - \hat{\boldsymbol{g}}_{\boldsymbol{m,n}}\|$$

3.2 Итеративные методы

В статье [1] рассматриваются итеративные методы решения задачи Фредгольма первого рода: аддитивный и мультипликативный.

В приложении к задаче аддитивный метод использует следующие итерации:

$$g_{m,n;k}(z) = g_{m,n;k-1}(z) = \int_0^\infty K(x,z)(\psi_{m,n;k}(x) - \psi_{m,n}(x))dx,$$

$$\psi_{m,n;k}(x) = \int_0^\infty K(x,z)g_{m,n;k}(z)dz.$$

Для мультипликативного метода используются такие итерации:

$$g_{m,n;k}(z) = \frac{g_{m,n;k-1}(z)}{\int_0^\infty K(x,z)dx} \int\limits_0^\infty \frac{K(x,z)\psi_{m,n}(x)}{\psi_{m,n;k}(x)}dx,$$

$$\psi_{m,n;k}(x) = \int\limits_0^\infty K(x,z)g_{m,n;k}(z)dz.$$

Так как ψ и g могут принимать отрицательные значения, производится следующее преобразование: выбирается параметр t, $\psi_{m,n}$ заменяется на $\tilde{\psi}_{m,n} = \psi_{m,n} + t$, $f_{m,n;0}$ заменятся на $\tilde{f}_{m,n;0} = f_{m,n;0} + t$.

3.3 Поправка для оценок

Будем также использовать поправку, предложенную в статье [2] В ней рассматриваются два случая: когда интеграл

$$\int \max(\hat{f}(x), 0)dx$$

больше 1, и когда меньше единицы.

В первом случае оценка $\hat{f}$ заменяется на $\tilde{f}(x) = \max(0, \hat{f}(x) - \xi)$, где ξ выбирается так, чтобы выполнялось

$$\int \tilde{f}(x)dx = 1.$$

Во втором случае используется оценка

$$\tilde{f}(x) = \tilde{f}(x; M) = \begin{cases} \max(0, \hat{f}(x)) + \eta_M, \text{ для } |x| \leqslant M, \\ \max(0, \hat{f}(x)), \text{ для } |x| > M, \end{cases}$$

где

$$\eta_M = \frac{1}{2M}\left(1 - \int \max(0, \hat{f}(x))dx\right).$$

§4. Эксперименты

Для аналитического способа.

Функция	Способ вычисления	Машинная точность (размер мантиссы),	Значение
$g_{0,0}(1)$	численно, интеграл, контур $[1-100i, 1+100i]$	100 десятичных знаков	0.864
	численно, ряд	256 двоичных знаков	0.864
$g_{0,0}(10)$	численно, интеграл контур $[1-100i, 1+100i]$	100 десятичных знаков	0.591
	численно, ряд	256 двоичных знаков	0.591
$g_{0,0}(100)$	численно, интеграл контур $[1-10i, 1+10i]$	100 десятичных знаков	-2×10^{19}
	численно, ряд	256 двоичных знаков	-0.440

Для численного вычисления интеграла. Мы использовали шаг 0,1 и $\alpha = 0{,}1$. И использовали $m = \{-5, \ldots, 5\}$, $n = \{-5, \ldots, 5\}$.

Рис. 4.1: $X \sim \mathcal{N}(0, 1)$

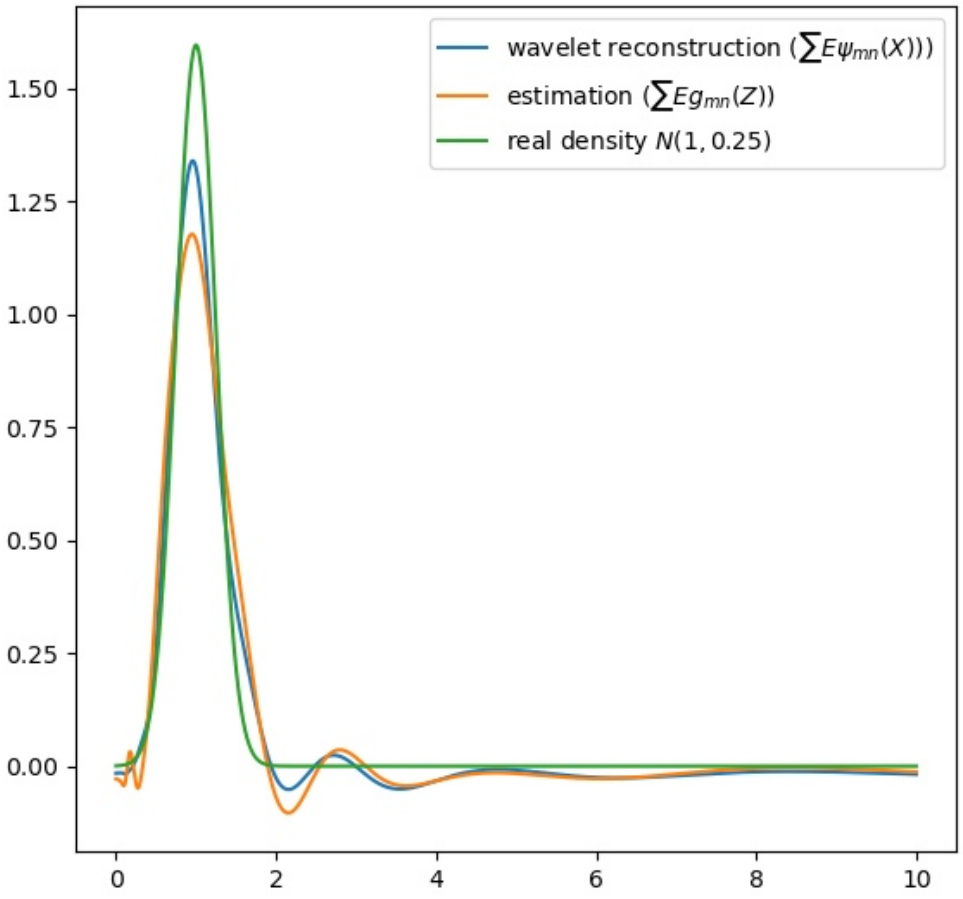


Рис. 4.2: $X \sim \exp(1)$

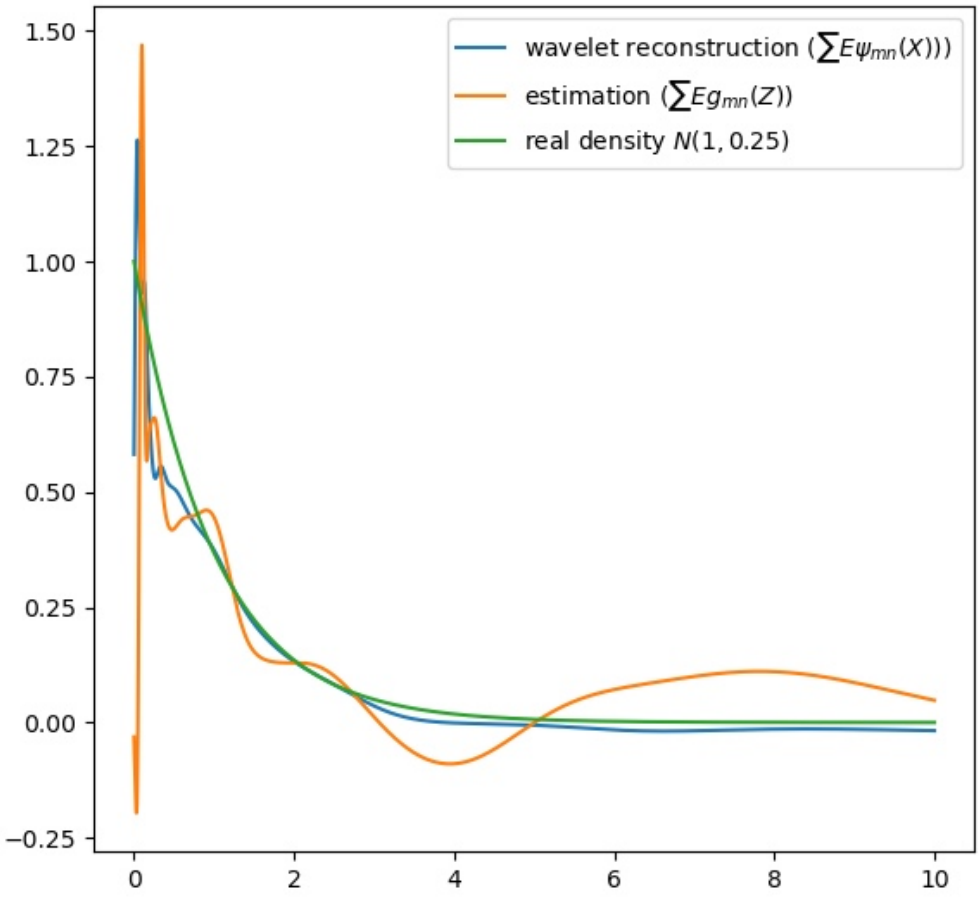


Рис. 4.3: $X \sim \chi^2_5$

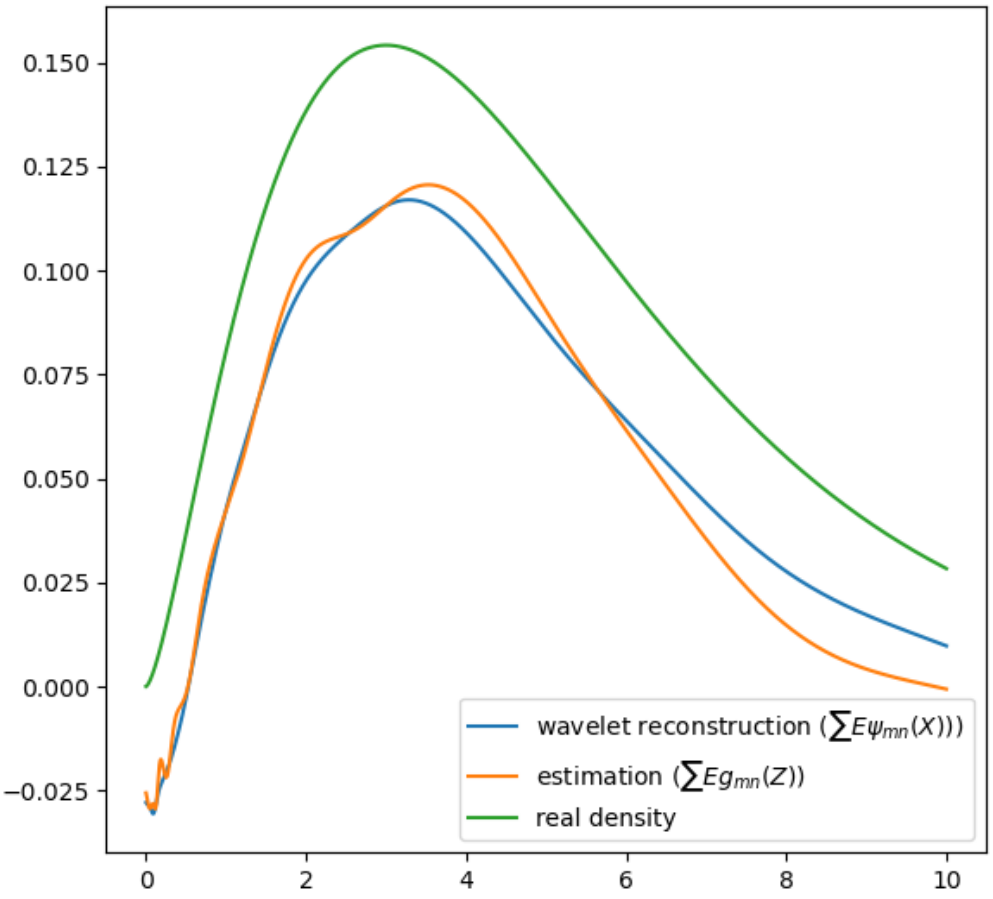


Рис. 4.4: Сравнение функций ошибок для метода градиентного спуска

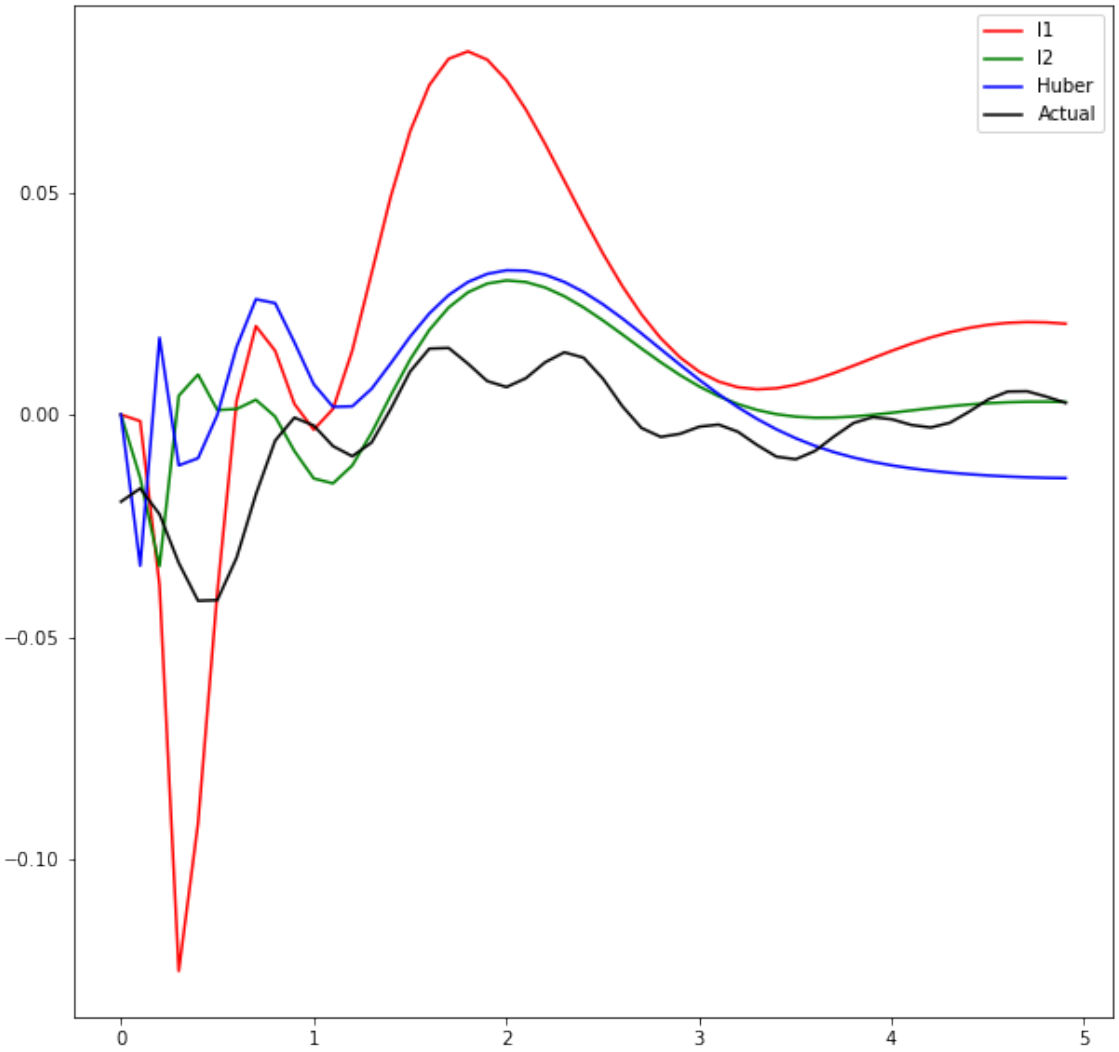


Рис. 4.5: Сравнение методов градиентного спуска, итеративного и МНК-оценки

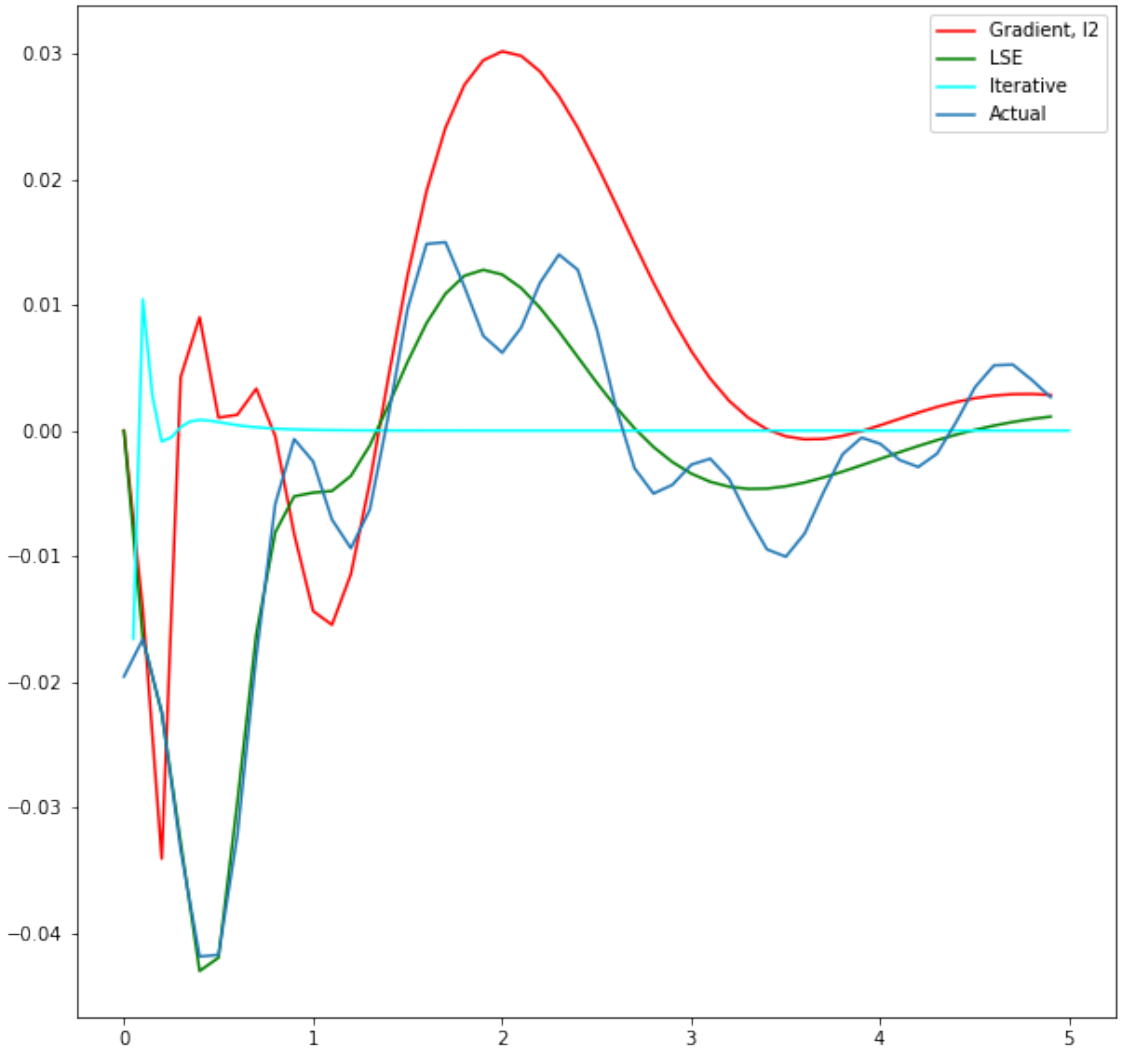


Рис. 4.6: МНК-оценка для смеси нормальных распределений

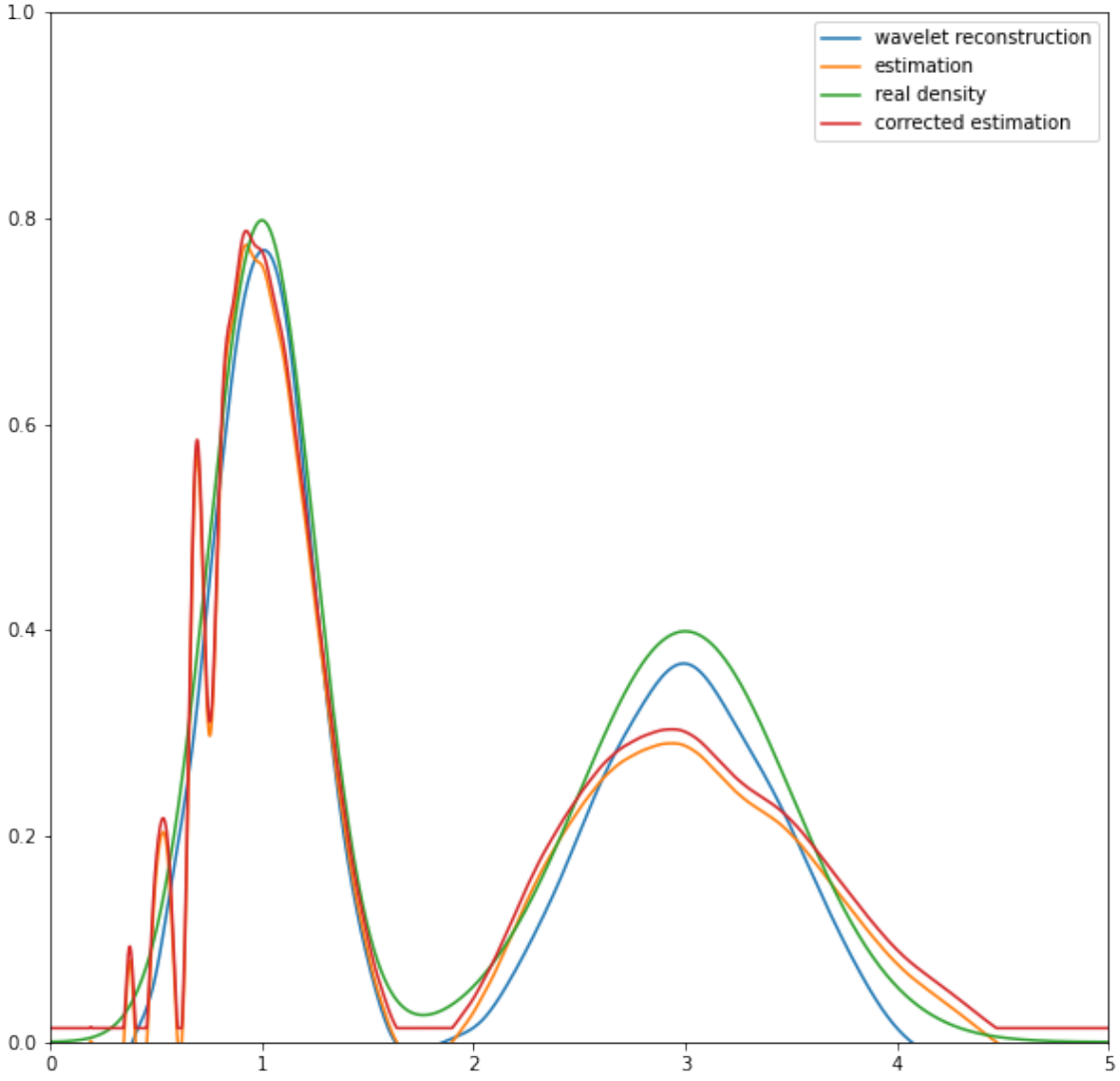


Рис. 4.7: Оценка методом градиентного спуска для смеси нормальных распределений

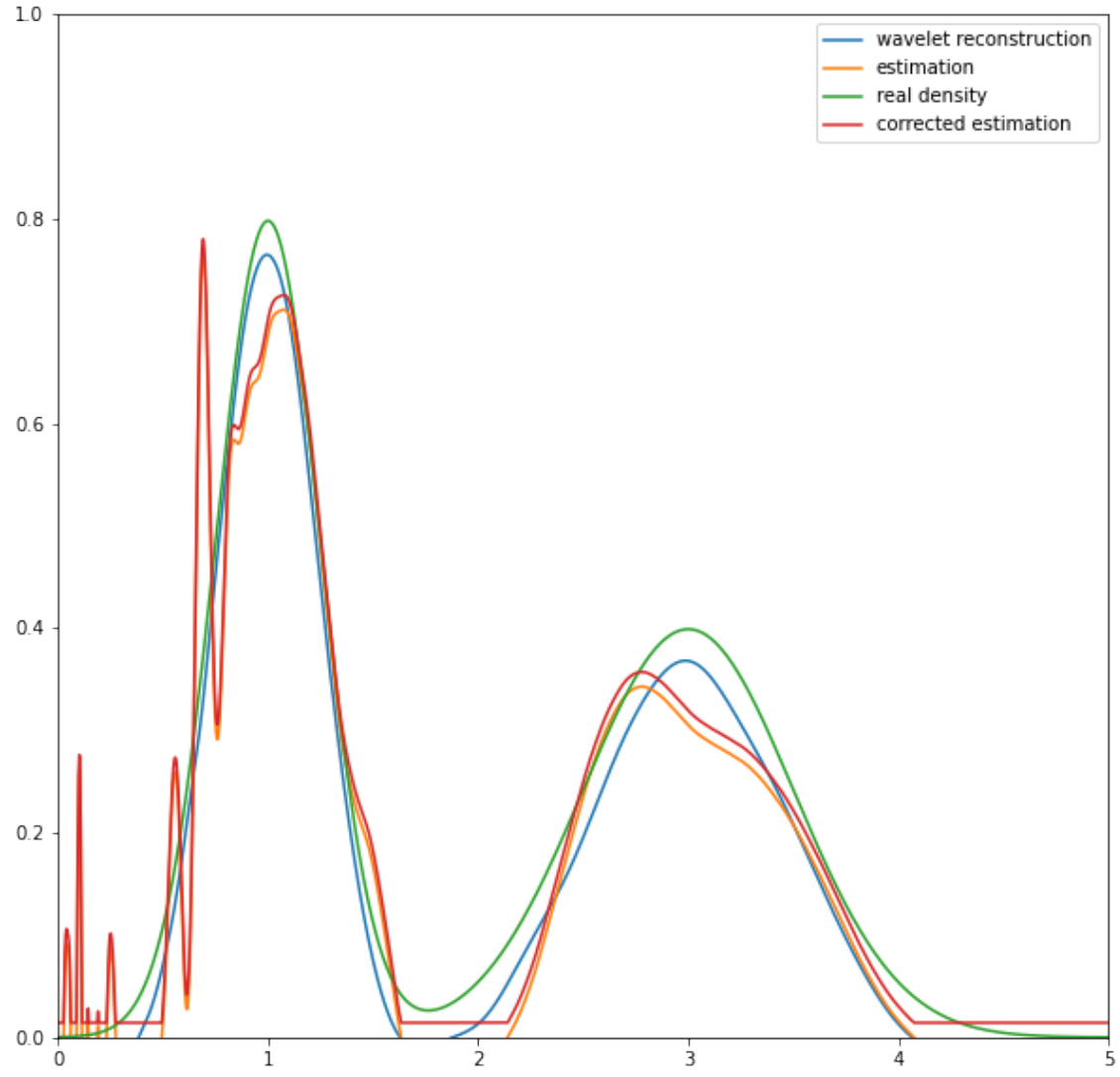
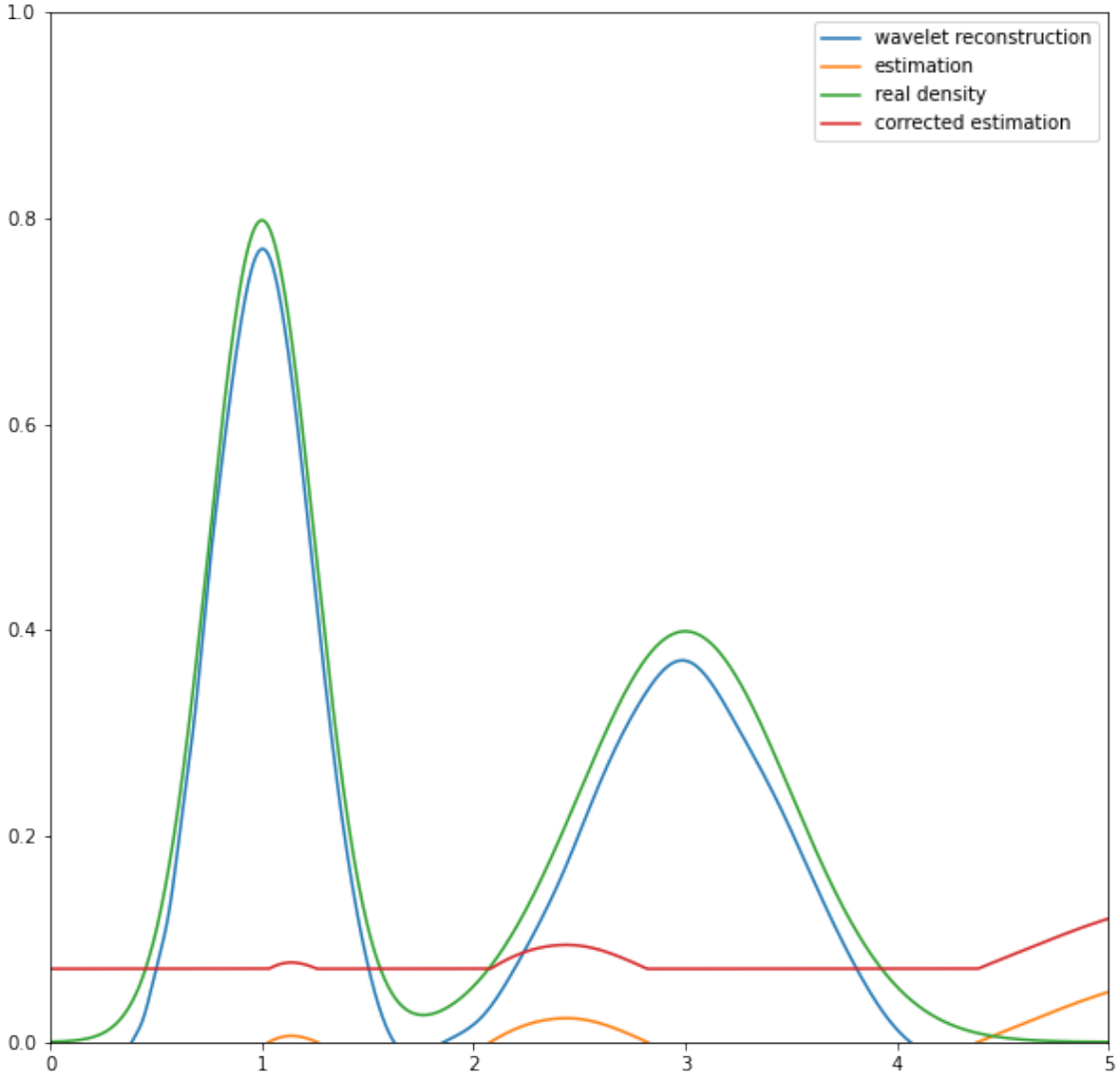


Рис. 4.8: Оценка итеративным методом для смеси нормальных распределений



§5. Обобщение на случай разных длин траекторий

Мы строили функции вида:

$$\mathbf{E}g_{m,n}(XY) = \mathbf{E}g_{m,n}(X) = c_{m,n}$$

и находили оценку плотности как

$$f_X(x) = c_{m,n}\psi_{m,n}(x).$$

Теперь рассмотрим случай, когда длины траекторий могут различаться. Для каждой длины k построим функции $g_{m,n,k}$ как описано выше и построим оценку $f_{X,k}(x)$

Пусть для длины траектории k у нас есть s_k наблюдений. И всего S наблюдений Тогда оценкой $f_X(x)$ будет

$$\sum_{k=1}^{K} \frac{s_k f_{X,k}(x)}{S}.$$

Докажем это. Разложим f_X в ряд по вейвлету:

$$f_X(x) = \sum_{m,n} c_{m,n}\psi_{m,n}(x).$$

Раскроем вейвлет-коэффициенты:

$$f_X(x) = \sum_{m,n} \mathbf{E}\psi_{m,n}(XY)\psi_{m,n}(x).$$

Представим математическое ожидание в виде математического ожидания условного математического ожидания при условии длины траектории:

$$f_X(x) = \sum_{m,n} \mathbf{E}_k \left(\mathbf{E}\left(\psi_{m,n}(XY)|k\right)\right) \psi_{m,n}(x).$$

По линейности математического ожидания, можем внести сумму внутрь:

$$f_X(x) = \mathbf{E}_k \left(\sum_{m,n} \mathbf{E}\left(\psi_{m,n}(XY)|k\right) \psi_{m,n}(x) \right).$$

Вычислим вейвлет-коэффициенты:

$$f_X(x) = \mathbf{E}_k \left(\sum_{m,n} c_{m,n,k} \psi_{m,n}(x) \right).$$

Заменим вейвлет-разложение на оригинальную функцию:

$$f_X(x) = \mathbf{E}_k f_{X,k}(x).$$

Получаем оценку:

$$f_X(x) = \sum_{k=1}^{K} \frac{s_k f_{X,k}(x)}{S}.$$

§6. Вывод

Лучшие результаты показывает МНК-оценка.

Оценка методом градиентного спуска более шумная, но позволяет использовать существенно более точный шаг дискретизации, так как возможно пожертвовать производительностью и не вычислять матрицу K заранее, что существенно снижает требования к количеству видеопамяти.

Итеративная оценка показывает неудовлетворительные результаты и сходится крайне медленно: разница между 1000 итераций и 10000 итераций несущественна.

Поправка для оценок плотностей несильно улучшает оценку.

Список литературы

[1] Minwoo Chae, Ryan Martin и Stephen G. Walker. “On an algorithm for solving Fredholm integrals of the first kind”. В: *Statistics and Computing* 29.4 (июль 2019), с. 645—654. ISSN: 1573-1375. DOI: 10.1007/s11222-018-9829-z. URL: https://doi.org/10.1007/s11222-018-9829-z.

[2] Ingrid K. Glad, Nils Lid Hjort и Nikolai G. Ushakov. “Correction of Density Estimators That Are Not Densities”. В: *Scandinavian Journal of Statistics* 30.2 (2003), с. 415—427. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4616772.