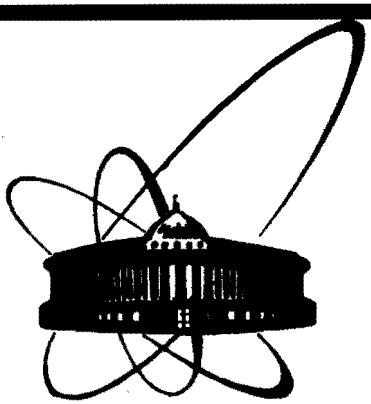


сообщения
объединенного
института
ядерных
исследований
дубна

Р10-86-764

В.Б.Злоказов

ROBUS - ПРОГРАММА ДЛЯ РОБАСТНОГО СГЛАЖИВАНИЯ ДИСКРЕТНЫХ ДАННЫХ С ТРЕНДАМИ РЕЗОНАНСНОГО ТИПА

1986

Данная работа представляет собой распространение идей и метода [I] на случай помехоустойчивого сглаживания.

Итак, пусть дискретные числовые данные $f(t)$, $t=t_0,\ldots,t_m$ представимы в виде:

$$f(t)=b(t)+n(t) \quad ,$$

где $b(t)$ – полезная компонента (информационная), а $n(t)$ – бесполезная, которая может иметь не только случайное, но и детерминированное происхождение. Требуется по $f(t)$ построить дискретную функцию $g(t)$, которая была бы как можно гладкой оценкой функции $b(t)$. Поскольку предполагаются сколь угодно сильные частотные перекрытия между функциями $b(t)$ и $a(t)$, аппарат линейных оптимальных частотных фильтров оказывается неприменимым. Так же нежелательно использование параметрических фильтров. В работе [I] искомый фильтр строился следующим образом:

I) вводилось понятие меры осцилляций;

2) оценка $g(t)$ искалась из условия одновременной минимизации ее меры осцилляций и расстояния до $f(t)$ в квадратичной метрике.

В настоящей работе предлагается некоторое обобщение меры осцилляций, а квадратичную метрику предлагается заменить на робастную [2]. Полученный непараметрический фильтр оказывается амплитудно-частотным, помехоустойчивым и не портящим существенно высокочастотные компоненты функции $b(t)$.

§ I. Выбор меры осцилляций

Для учета не только частотных свойств функции, но и амплитудных, в работе [I] в качестве меры осцилляций функции $f(x)$ была предложена не просто величина, пропорциональная норме 2-й производной этой функции, но выражение

$$\mu_f(x)=\frac{f''(x)}{\sqrt{1+f'^2(x)}} \quad , \quad \mu_f(a,b)=\int_a^b \mu^2(x)\,dx \quad , \qquad (I.I)$$

Здесь в знаменателе стоит дифференциал дуги функции.

Выражение (I.I) допускает следующие обобщения:

I) для комплекснозначной функции

$$\mu_f(a,b)=\int_a^b \frac{f''(x)f''^{*}(x)}{1+f'(x)f'^{*}(x)}\,dx \quad , \qquad (I.2)$$

где f'^{*} и f''^{*} – функции, комплексно-сопряженные к f' и f'' ;

2) произвольную степень знаменателя

$$\mu_f(x)=\frac{f''(x)}{\sqrt{(1+f'^2(x))^{\beta}}}\ ,\quad \mu_f(a,b)=\int_a^b \mu^2(x)dx. \qquad (I.3)$$

Рассмотрим совокупность гармонических функций $\{\exp(-ikx)\}, h=0,\dots,\infty$ и определим с помощью (I.3) их интегральные меры осцилляций на участке $[0,\ 2\pi]$:

$$\mu_{e^{-ikx}}(0,2\pi)=\int_0^{2\pi}\frac{k^4}{(1+k^2)^{\beta}}\,dx=\frac{2\pi k^4}{(1+k^2)^{\beta}}\sim ck^{4-2\beta}\ , \qquad (I.4)$$

т.е. мера осцилляций базисных гармонических функций $\exp(-ikx)$ при $\beta = I$ растет пропорционально квадрату частоты k . Этот закон можно менять, используя в качестве меры осцилляций выражение (I.3) с β , отличным от единицы. В частности, кривизна $f''/(1+f'^2)^{3/2}$ является линейной функцией гармонической частоты.

Рассмотрим, далее, семейство функций, зависящих от параметров следующим образом:

$$f(x,A,P,W)=Am(\tfrac{x-P}{W}),$$

Сюда относятся, например, функции экспоненциального типа. Для них интегральная мера осцилляций будет равна

$$\frac{\frac{A^2}{W^4}\,m''^2_{xx}}{(1+\frac{A^2}{W^2}m'^2_x)^{\beta}}\sim\frac{A^{2\beta-2}}{W^{2\beta-4}}. \qquad (I.5)$$

Отсюда видно, что при $\beta = I$, мера осцилляций функции $f(x)$ инвариантна относительно "высоты" фигуры, график которой изображается функцией $f(x)$, а при $\beta = 2$ – относительно ее "ширины".

Конкретный выбор значения β при сглаживании данных, моделью которых служат финитные функции, на которые наложены осциллирующие гармонические шумы, определяется анализом на основе (I.4), (I.5) частотного состава полезной и шумовой компоненты этой модели. Качественные соображения при этом будут следующими: если спады частотных характеристик у полезной компоненты и шума существенно различны, β можно брать поменьше, иначе более подходящими окажутся большие β .

§ 2. Построение фильтра

Искомый фильтр для сглаживания функции $f(t)$ на отрезке (t_1,t_2) представляет собой следующую задачу минимизации: построить по данной функции $f(t)$ функцию $g(t)$, имеющую минимальные меру осцилляций и

расстояние до $f(t)$. Математически это выглядит так: найти минимум при заданном $\lambda > 0$:

$$\int_{t_1}^{t_2}\frac{g''(t)^2}{(1+g'(t)^2)^{\beta}}+\lambda\rho(g(t),f(t),t_1,t_2) \qquad (2.I)$$

при граничных условиях

$$g(t_1)=f(t_1),\ g(t_2)=f(t_2)\ ,\ g''(t_1)=0,\ g''(t_2)=0\ . \qquad (2.2)$$

Метрика ρ может быть произвольной, в частности, квадратической. Однако больший интерес представляют робастные метрики.

Рассмотрим прежде всего дискретный аналог задачи. Итак, для дискретной функции $f(k),k=1,\dots,m$ мы должны выбрать дискретные аналоги I-й и 2-й производной. Вопрос о 2-й производной решается просто: вместо нее берем 2-ю разность. Сложнее выглядит выбор дискретного аналога I-й производной. Мы имеем такие аппроксимации I-й производной:

$$f'(i)\sim f(i)-f(i-1) \qquad (2.3)$$

$$f'(i)\sim f(i+1)-f(i) \qquad (2.4)$$

$$f'(i)\sim f(i+1)-f(i-1))/2\ . \qquad (2.5)$$

Поскольку (2.5) есть среднее (2.3) и (2.4), и тем самым более сглажено, чем они, (2.3) и (2.4) больше подходят для таких участков данных, как вершины пиков, а (2.5) – для более низких по амплитуде, но осциллирующих данных.

Далее, было произведено следующее упрощение задачи (2.I). Введено понятие априорной сглаженной оценки $g_o(i)$. Ее можно строить, например, так : если тройка чисел $f(i-1),f(i),f(i+1)$ разнятся между собой в пределах $\sigma(i)$, то $g_o(i)$ равно их среднему (т.е. м.н.к.-оценке), иначе, $g_o(i)$ равна их медиане (т.е. робастной оценке). С помощью априорной оценки $g_o(i)$ была упрощена запись меры осцилляций, а в качестве робастной метрики в (2.I) была использована взвешенная квадратическая. Окончательно дискретный аналог (2.I), (2.2) выглядел так: найти минимум

$$\sum_{i=2}^{m-1}\frac{g''(i)^2}{(1+g_o'(i)^2)^{\beta}}+\lambda R\sum_{i=1}^{m}W(i)(g(i)-f(i))^2 \qquad (2.6)$$

при граничных условиях

$$g(1)=f(1),\ g(m)=f(m),\ g''(1)=0,\ g''(m)=0. \qquad (2.7)$$

Веса w_i строились следующим образом:

$$w_i=\begin{cases} I, \text{ если } |g_o(i)-f(i)|\le c\sigma(i) \\ \dfrac{1}{|g_o(i)-f(i)|}, \text{ иначе}. \end{cases} \qquad (2.8)$$

фактор R введен для нормировки. Он равен $\sum_{i=2}^{n-1}(1+g'_o(i)^2)^{\beta}/\sum_{i=1}^{n} w_i$.

Величины $\sigma(i)$ вычислялись как корни квадратные из $f(i)$, т.е. программа ROBUS ориентирована на сглаживание данных пуассоновского типа, но, разумеется, сам метод не зависит от выбора погрешностей данных.

Величина λ считается заданной. Она регулирует степень сглаживания и определяется с помощью тестов. Благодаря возможности сделать меру осцилляций инвариантной относительно различных преобразований данных и нормировке через R, λ можно выбрать так, чтобы оно годилось для очень большого круга данных.

Квадратическая метрика с весами (2.8) близка по свойствам к метрике Хьюбера /2/ и обладает сходными робастными свойствами.

§ 3. Вычислительная схема

Уравнение для нахождения минимума (2.6) при ограниченных (2.7) имеет вид:

$$e(i+1)g''(i+1)-2e(i)g''(i)+e(i-1)g''(i-1)+\bar{\lambda}w(i)g(i)=\bar{\lambda}w(i)f(i), \quad (3.1)$$

где $e(i)=1/(1+g'_o(i)^2)^{\beta}$, $\bar{\lambda}=\lambda/R$.

Сравнивая (3.1) с соответствующим уравнением в работе /I/, мы видим, что оно отличается только множеством $w(i)$ при $\bar{\lambda}$. Отсюда легко переписать формулы из /I/ таким образом. Для коэффициентов матричной прогонки

$$\bar{P}_i=A_i\bar{P}_{i+1}+B_i, \quad (3.2)$$

где $\bar{P}_i$ есть вектор $(g(i),g''(i))$, A_i – матрица размером 2 x 2, B_i – вектор размером 2, имеем

$$a_i^{11}=d_i^{11},\ a_i^{12}=d_i^{12},\ a_i^{21}=d_i^{21},\ a_i^{22}=d_i^{22},\ a^{11}=a_i^{12}=a_i^{21}=a_i^{22}=0$$
$$b_i^1=d_i^{11}b_{i-1}^1-d_i^{12}(\bar{\lambda}w(i)f(i)-e(i-1)b_{i-1}^2)/e(i+1), \quad b_1^1=f_o \quad (3.3)$$
$$b_1^2=d_i^{21}b_{i-1}^1-d_i^{22}(\bar{\lambda}w(i)f(i)-e(i-1)b_{i-1}^2)/e(i+1), \quad b_1^2=0,$$

где d_i^{kj} элементы матрицы, обратной к

$$\begin{pmatrix} 2-a_{i-1}^{11} & 1-a_{i-1}^{12} \\ -(e(i-1)a_{i-1}^{21}+\bar{\lambda}w(i))/e(i+1) & (2e(i)-e(i-1)a_{i-1}^{22})/e_{i+1} \end{pmatrix}.$$

Сначала вычисляются прогоночные коэффициенты по (3.3), затем по (3.2) получается решение $(g(i),g''(i))$.

§ 4. Локальный робастный фильтр

Локальный фильтр описанного типа проще всего строится по рекурсивной схеме для 5 точек. Система уравнений для определения $g(i+1),g(i),g(i-1)$ при найденном $g(i-2)$ и $g(i+2)=f(i+2)$ имеет вид:

$$(4e(i+1)+e(i)+\bar{\lambda}w(i+1))g(i+1)-2(e(i+1)+e(i))g(i)+e(i)g(i-1)=\bar{\lambda}w(i+1)f(i+1)+$$
$$+2e(i+1)f(i+2);$$
$$-2(e(i)+e(i+1))g(i+1)+z_i g(i)-2(e(i)+e(i-1)g(i-1)=\bar{\lambda}w(i)f(i)-e(i+1)f(i+2)-$$
$$-e(i-1)g(i-2);$$
$$e(i)g(i+1)-2(e(i)+e(i-1)g(i)+(e(i)+4e(i-1)+\bar{\lambda}w(i-1))g(i-1)=\bar{\lambda}w(i-1)f(i-1)+$$
$$+2e(i-1)g(i-2), \text{ где } \quad z_i=(e(i+1)+4e(i)+e(i-1)+\bar{\lambda}).$$

Эта система проще всего решается обращением матрицы 3-го порядка. Затем i сдвигается на I (если только $g(i-1)$ добавляется к решению), на 2 (если добавляется $g(i-1)$ и $g(i)$), на 3 (если добавляются все три числа).

§ 5. Фильтрация больших массивов данных

Так как описанный фильтр требует в общем случае 8 массивов для хранения исходных данных, величин $e(i)$ и прогоночных данных, каждый размером m чисел, то при больших m возникают слишком большие требования к объему памяти ЭВМ.

Целесообразно разбить такие данные на малые группы и обрабатывать каждую группу отдельно. Как показали тестовые испытания, при этом качество фильтрации при фиксированном λ может ухудшиться лишь за счет нарушения масштабной (или ширинной) инвариантности фильтра. Во избежание этого рекомендуется вычислять фактор R для всех данных и сглаживать данные с таким общим для всех групп R.

§ 6. Многомерные обобщения

Фильтр может быть обобщен на многомерный случай следующим образом: Если $f(\bar{k})$, $\bar{k}$ – вектор индексов размерности n, и $g_o(\bar{k})$ – априорная оценка сглаженных данных, то мера осцилляций функции n аргументов в области A строится так

$$\mu_f(A)=\sum_{\{\bar{k}\}\in A}\frac{f''^2_{k_1k_1}(\bar{k})+f''^2_{k_2k_2}(\bar{k})\ldots+f''^2_{k_nk_n}(\bar{k})}{(1+f'_{k_1}(x)^2+f'_{k_2}(x)^2+\ldots+f'_{k_n}(x)^2)^{\beta}} \quad (6.1)$$

Здесь k_i – i –ая компонента вектора k.

Метрика в (2.6) остается неизменной, лишь аргумент i заменяется на многомерный k.

Многомерный фильтр требует очень большую вспомогательную память для хранения промежуточных результатов. Компромиссными вариантами будут следующие:

I) сглаживание с помощью фильтра (2.6) по каждой координате одномерных сечений функции $f(\bar{k})$ последовательно или по какому-либо другому закону;

2) переход к локальным фильтрам.

Приведем пример локального фильтра для дискретной функции двух аргументов $f(i,j)$.

Мера осцилляций в двумерном случае выглядит так:

$$\mu_f(A)=\sum_{i,j\in A}(f''^2_{ii}(i,j)+f''^2_{jj}(i,j))/((1+f'^2_{io}(i,j)+f'^2_{jo}(i,j))^\beta$$

при наличии априорной оценки $f_o(i,j)$ сглаженной функции. Приведем пример рекурсивного 5-точечного локального фильтра. Пусть вычислен, как и в одномерном случае, глобальный фактор R для всей функции $f(i,j)$. Первый шаг - получение сглаженных граничных при $i=1$ и $j=1$ значений либо из теоретических соображений, либо с помощью одномерных глобальных фильтров (2.6), (2.7). Затем по вычисленным $g(i-1,j)$ и $g(i,j-1)$ и $f(i+1,j)$ и $f(i,j+1)$ вычисляется сглаженное значение $g(i,j)$ по формуле:

$$g(i,j)=\frac{aW(i,j)f(i,j)+g(i-1,j)+g(i,j-1)+f(i,j+1)+f(i+1,j)}{4+aW(i,j)},$$

где

$$a=\frac{\lambda(1+(g_o(i+1,j)-g_o(i-1,j))^2/4+(g_o(i,j+1)-g_o(i,j-1))^2/4}{2R},$$

$$W(i,j)=\begin{cases}1, & \text{если } |g_o(i,j)-f(i,j)|<\sigma(i.j)\\ \dfrac{1}{|g_o(i,j)-f(i,j)|}, & \text{иначе.}\end{cases}$$

В качестве $g_o(i,j)$ можно брать, например, медиану из чисел $g(i-1,j)$, $g(i,j-1)$, $f(i,j)$, $f(i,j+1)$, $f(i+1,j)$ или приближенную медиану, определяемую для чисел $a_i, i=1,\dots,n$ так:

$$M\{a_i\}=(\sum_{i=1}^{n}a_i-\max_i\{a_i\}-\min\{a_i\})/(n-2).$$

§ 7. Описание программы

Обращение к программе имеет вид:

CALL ROBUS(SA,M,SB,SD,RR,RMODE,BET,UR) , где:

SA(2M) – массив исходных данных длиной M чисел;

SB(2M) – рабочий массив;

SD(4M) – рабочий массив, после фильтрации SD содержит результат сглаживания M чисел ;

RR – λ

RMODE =
- 0 – строится априорная функция $g_o(t)$ так: сначала берется скользящая медиана, затем она сглаживается с помощью фильтра 4-й разности /4/:
  $g_o(t)=(12(g(t+1)+g(t-1)-3(g(t+2)+g(t-2))+17g(t))/35$
  Затем по $g_o(t)$ строится мера осцилляций и далее фильтр работает по описанной в работе схеме.
- 1 – В качестве результата выдается $g_o(t)$
- 2 – Для построения меры осцилляций берется заданная в массиве SD функция $g(t)$.

BET = β

UR = c (в формуле 2.8)

§ Тестирование алгоритма

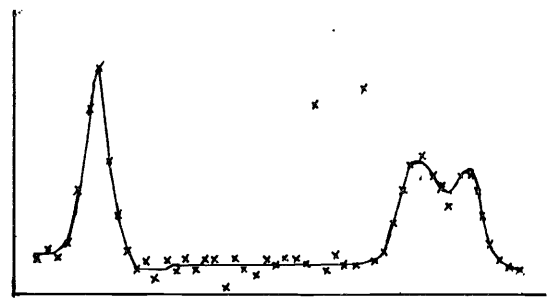
Поскольку преимущества описанного фильтра перед обычными частотными были показаны еще в работе /1/, целью тестов было следующее:

1) показать на данных пуассоновского типа, содержащих "выбросы", работу фильтра;

2) проверить некоторые статистические характеристики данного фильтра.

Результаты тестирования приведены в таблице

Работу программы иллюстрирует рисунок.



Крестиками обозначены исходные данные, непрерывной линией - сглаженные. Видно, что фильтр не реагирует на грубые выбросы, почти не исправляет пики и в основном сглаживает плавный участок данных.

Таблица

В таблице приведены сначала для данных без выбросов, а затем для тех же данных с выбросами:

C1,C2,C3 – средние значения на трех последовательных участках модуля дискретного преобразования Фурье;

SM – величина, характеризующая гладкость данных – средняя кривизна;

DE – среднеквадратичное отклонение от исходных данных;

VAN – нормированное среднеквадратичное отклонение от линии, являющейся математическим ожиданием данных;

VAUN – ненормированное такое же отклонение;

C – среднее значение нормированного отклонения от истинной линии.

Эти величины приведены для исходных данных, для отфильтрованных с помощью фильтра ROBUS и фильтра 4 разности (локально-регрессионного).

	c_1	c_2	c_3	SM	DE	VAN	VAUN	C
исх.	.77	.I6	.07	72	0	II.4	5954	.27
ROBUS	.80	.I6	.04	0	I382	5.6	4I38	.30
4 разн.	.83	.I4	.02	8I	I672	4.9	2596	.34

	c_1	c_2	c_3	SM	DE	VAN	VAUN	C
исх.	.54	.22	.24	4877	0	60I.7	I24974	4.67
ROBUS	.79	.I6	.05	0	II7847	5.8	4623	.2
4 разн.	.69	.22	.08	I6	37672	295.0	6II75	4.74

Комментарий к таблице

Видно, что выбросы сильно смещают спектр Фурье в сторону верхних частот: от (.77, .I6, .7) к (.54, .22, .24), нарушают резко гладкость, т.е. среднюю кривизну: от 72 к 4677, резко увеличивают средние дисперсию и отклонение.

При этом ROBUS обеспечивает оптимальную гладкость: меньше единицы в среднем, сохраняет спектр Фурье почти нечувствительным к выбросам и обеспечивает минимальное нормированное отклонение. Меньшая средняя дисперсия у фильтра 4-й разности объясняется тем, что ROBUS сохраняет почти неизменными пики, где дисперсия максимальна. Но зато ROBUS сглаживает плавную часть данных значительно лучше, чем фильтр 4-й разности. Результаты сравнения с другими фильтрами: Фурье, частотным и 2-й производной не приведены, так как характеристики этих фильтров намного хуже, чем характеристики фильтра 4-й разности.

Литература

I. Zlokazov V.B., Comp. Phys. Comm., 21(1981), 373-383, ОИЯИ, PI0-80-5I0, Дубна, I980.

2. Хьюбер П.Дж. Робастность в статистике. Мир, М., I984.

3. Wilks S.S., Mathematical Statistics (John Wiley, New York, London, 1962).

4. Lanczos C., Applied Analysis. Prentice Hall, inc, Englewood Cliffs, N.J., 1956.

Рукопись поступила в издательский отдел
26 ноября I986 года.

Злоказов В.Б.

P10-86-764

ROBUS - программа для робастного сглаживания дискретных данных с трендами резонансного типа

Описан сглаживающий фильтр сплайнового типа, сохраняющий пики и нечувствительный к отдельным выбросам. Для дискретной функции g(k) ее сглаженный вариант s(k) можно найти как решение задачи: минимизировать

$$\sum_{i=2}^{m-1} e_i s''(i)^2 + \lambda \rho\,(s(i), g(i)),\quad i = 1,\ldots,m,$$

где e_i - мера осцилляций функции g(k), $e_i = 1/(1 + g'(i)^2)$, s'', g' - вторая и первая разности функций s и g, ρ - робастная метрика.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна 1986

Перевод О.С.Виноградовой

Zlokazov V.B.

P10-86-764

ROBUS - A Program for Robust Smoothing of Discrete Data Containing Resonance-Like Trends

A smoothing filter of the spline type is described, which saves peaks and is not sensitive to the single outliers. s(k) - a smoothed match to a discrete function g(k) is searched for as a solution of the problem:

$$\text{minimize} \sum_{i=2}^{m-1} e_i s''(i)^2 + \lambda \rho\,(s(i), g(i)),\quad i = 1,\ldots,m,$$

where e_i is an oscillation measure of the function g(k), $e_i = 1/(1 + g'(i)^2)$, s'', g' are the second and the first differences of functions s and g, and ρ is a robust metrics.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna 1986