объединенный
институт
ядерных
исследований
дубна

Р10-88-94

В.Б.Злоказов

ЗАДАЧА ОПТИМАЛЬНОГО ГИСТОГРАММИРОВАНИЯ СПЕКТРАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

Направлено в журнал "NIM"

1988

Злоказов В.Б. P10-88-94

Задача оптимального гистограммирования спектральных распределений

Описан подход к оптимизации выбора параметров ячейки гистограммы широкого класса регрессии — спектральных распределений. Показано, что для регрессии $s(x)$ минимальные статистические погрешности дают ячейки равного веса: такие интервалы Xk, что $\int_{X_k} s(x)\,dx = \int_{X_i} s(x)\,dx$ для любого $k \neq i$, а минимальные погрешности представления гистограммой функции получаются, если приписывать абсциссу ординаты гистограммы точке, в которой функция равна своему интегральному среднему (при использовании квадратичной погрешности (или амплитудному среднему) в случае равномерной метрики). В обоих случаях эта абсцисса не обязательно совпадает с серединой ячейки. Описанные методы оптимизации применены к наиболее часто встречающимся моделям спектральных функций — экспоненте, функциям Гаусса и Лоренца.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1988

Перевод автора

Zlokazov V.B. P10-88-94

On Optimal Binning for Spectral Distributions

The paper describes A method of the optimation of the selection of parameters of a histogram bin for a very large class of regressions — those of spectral distributions is described. It is shown that for a regression $s(x)$ the minimal statistical errors are given by bins of equal weights: they are such intervals Xk that the integrals of $s(x)$ over them are equal for any pair of indexes (k, j), $j \neq k$; whereas the minimal errors of the representation of a function by a histogram are obtained, when the abscissa of any histogram ordinate is assigned to the value, where the function is equal to its integral mean (while using the quadratic metrics) or to the amplitude mean (while using the uniform metrics). In both cases this abscissa may differ from the middle of the bin. The described method of the optimation has been applied to the most frequently used models of spectral functions — exponential, Gaussian and Lorentzian.

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

Preprint of the Joint Institute for Nuclear Research. Dubna 1988

Под гистограммированием понимается сопоставление функции непрерывного аргумента $s(x)$ ее гистограммы - функции дискретного аргумента $s(k)$, которая является представителем функции $s(x)$.

Аргумент k, называемый ячейкой гистограммы, соответствует некоторой подобласти значений непрерывного аргумента x.

Основная задача гистограммирования - это задача оптимального биннинга. Биннинг (bin) - транслитерированный английский термин для обозначения процедуры выбора характеристик ячейки гистограммы.

Прямая задача биннинга состоит в том, чтобы по имеющейся информации /более или менее полной/ о функции $s(x)$ выбрать характеристики ячейки гистограммы, представляющей $s(x)$, оптимальным в том или ином смысле способом.

Данная задача играет большую роль в прикладной математике, но относится к наименее исследованным. Имеется ряд работ по описанию оптимальной ячейки в некоторых частных случаях для функций и плотностей распределения/1/, но вопрос о регрессиях остается открытым. Наша работа обязана восполнить этот пробел для очень важного класса регрессий - спектральных распределений ядерной физики.

Наиболее общим толкованием значения ординаты дискретного спектра $s(k)$ в ячейке k будет оцененный с определенной погрешностью интеграл с функции $s(x)$ по области Xk.

Таким образом, когда мы ставим в соответствие функции $s(x)$ ее гистограмму $s(k)$, то последняя представляет первую с погрешностью, в которой мы можем выделить две компоненты, далее называемые погрешностями биннинга 1-го и 2-го рода:

1/ погрешность 1-го рода из-за неточного представления ординатами $s(k)$ интегралов $\int\limits_{Xk} s(x)\,dx$;

2/ погрешность 2-го рода из-за неточного представления ординатами $s(k)$ значений функции $s(xk)$.

Пусть задано отображение области X на множество пар $\langle Xk, xk \rangle$, где Xk - замкнутые интервалы, на которые разбита область X, а xk - выделенные точки в этих интервалах.

Мы ставим в соответствие совокупности значений функции $\langle s(x) \rangle$, $x \in X$, множество $\langle s(k) \rangle$, где k принадлежит области K, и считаем, что $s(k) \sim s(xk)$.

Можем записать

$$s(k) = \int_{Xk} s(x)dx + epsk =$$

$$= \int_{Xk} (s(xk) + (s(x) - s(xk)))\,dx + epsk = s(xk)\cdot Lk + delk + epsk.$$

Здесь $epsk$ - погрешность биннинга 1-го рода, $delk$ - погрешность биннинга 2-го рода, а Lk - длина ячейки Xk.

Приступим теперь к формальному описанию процедуры оптимального гистограммирования. Предварительно мы должны указать факторы, влияющие на ошибки биннинга $delk$ и $epsk$. Поскольку $s(k)$ - случайная величина, факторами, определяющими ее погрешность, будут дисперсия и смещение, т.е. разность $MEXP\ s(k) - \int s(x)dx$, а также специфически статистический фактор: корреляция между ординатами $s(k)$ и $s(k+j)$, $k = 0, \ldots, m$. Факторами, определяющими погрешность $delk$, будут разбиение X на интервалы Xk и выбор точек xk в последних.

Рассмотрим детальнее влияние факторов, определяющих погрешность $epsk$. Мы считаем, что непрерывная линия спектра $f(x)$ есть математическое ожидание случайной величины $s(x)$ - функции аргумента x. Математическое ожидание величины $s(k)$ будет равно, очевидно,

$$Mk = \int_{v} f(x)dx.$$

С другой стороны, мы можем воспользоваться моделью последовательных двоичных независимых испытаний и считать, что $s(x)$ есть сумма независимых событий, каждое из которых происходит с вероятностью pk. Если разыгрывалось n событий, то математическое ожидание и дисперсия суммы событий, попавших в Xk, будут равны соответственно

$$n\cdot pk = Mk \qquad n\cdot pk\cdot(1-pk) = Mk - Mk^2/n = Dk.$$

Пусть $Pk(z)$ - функция распределения случайной величины $s(k)$ - ординаты спектра на интервале Xk. На основании неравенства Чебышева для заданной вероятности Pk можно записать

$$ABS(s(k) - Mk) < SQRT(Dk/Pk).$$

Это неравенство выполняется с вероятностью $1 - Pk$, и оно определяет стохастическую точность, с которой $s(k)$ равняется n-кратному интегралу от $f(x)$ на Xk. Выражение справа может быть записано как

$$epsk = SQRT((Mk - Mk^2/n)/Pk). \qquad /1/$$

Если число событий n задано /статистика эксперимента фиксирована/, то при фиксированном $P = Pk$ это выражение является функцией величины Mk. Задачу оптимизации гистограммирования в данном случае можно сформулировать так: построить оператор гистограммирования для функции $f(x)$ и области X такой, чтобы сумма квадратов относительных погрешностей оценок интегралов Mk была минимальной, т.е. чтобы было минимальным выражение

$$\sum_k (epsk/Mk)^2.$$

Положим число интервалов равным m. Подставив /1/ вместо $epsk$ и убрав постоянную величину P, перепишем оптимизирующий критерий следующим образом:

$$\sum_{k=1}^{m} (Mk - Mk^2/n)/Mk^2 \qquad /2/$$

при условии

$$\sum_{k=1}^{m} Mk = const, \quad 0 < Mk.$$

Константа в этом условии равняется интегралу от $f(x)$ по всей области X. Определим оптимальные Mk из условия минимума /2/. Функция Лагранжа имеет вид

$$\sum_{k=1}^{m} (1/Mk - 1/n) + \lambda \sum_{k=1}^{m} Mk.$$

Ее стационарная точка удовлетворяет уравнениям

$$1/Mk^2 = \lambda, \ \sum_{k=1}^{m} Mk = const.$$

Отсюда находим $M1 = M2 = \ldots = Mm$, и, следовательно, $Mk = const/m$. Матрица вторых производных будет диагональной и положительно определенной и, следовательно, данное решение действительно соответствует минимуму. Таким образом, оптимальную в смысле критерия /2/ гистограмму мы получим, разделив область изменения аргумента X на интервалы равного веса. Суммарное наилучшее квадратичное отклонение будет, очевидно, равно

$$eps = (1/P)\cdot\sum_{k=1}^{m} (1/Mk - 1/n) = m\cdot(m/const - 1/n)/P.$$

Если величина m фиксирована, оптимальная относительная погрешность 1-го рода eps будет функцией SQRT(1/n). Если $m = n^a$, то для улучшения относительной погрешности 1-го рода при росте n к ∞, надо потребовать, чтобы a было строго меньше 0,5; поскольку

$$eps = const \cdot (n^{(2a-1)} - n^{(a-1)}).$$

Очевидно, что данный критерий пока ничего не говорит о выборе узлов ячеек.

Если отказаться от требования сохранения числа ячеек, то результат будет таким: оптимальную в смысле /2/ гистограмму дает разделение области X на минимально допустимое число ячеек равного веса. Это утверждение легко следует из такого соотношения:

$$MIN \sum_{k=1}^{t} (1/Mk - 1/n) < MIN \sum_{k=1}^{t+r} (1/Mk - 1/n) \quad 0 < Mk \le 1, \ t, r \ge 1$$

при условиях $\sum_{k=1}^{t} Mk = 1$ и $\sum_{k=1}^{t+r} Mk = 1$ соответственно, которые получаются после нормировки Mk на const.

Для доказательства этого соотношения заметим, что первый минимум дается решением $Mk = 1/t$, второй - $Mk = 1/(t+r)$. Сравнив

$$\sum_{k=1}^{t} (t-1) = t^2 - t \quad и \quad \sum_{k=1}^{t+r} (t+r-1) = (t+r)^2 - (t+r)$$

и учитывая, что $t, r \ge 1$, получим наше утверждение.

Вывод: чем шире ячейки, тем гистограмма лучше в смысле критерия /2/.

ПРИМЕР 1. Чистая экспонента.

Пусть $f(x) = a \cdot \exp(-a \cdot t)$, $t \in (0, \infty)$. Пусть $<Tk>$, k = 1, ..., m - оптимальное разбиение оси $(0, \infty)$ на m ячеек равного веса, равного 1/m. Попытаемся вывести явные формулы для концов этих интервалов. Пусть $t(i-1)$ и $t(i)$ - концы i-го интервала. Интеграл от f(x) на i+1-ом интервале равен

$$\int f(x)\,dx = \exp(-a \cdot t(i)) - \exp(-a \cdot t(i+1)) =$$

$$= \exp(-a \cdot t(i)) \cdot (1 - \exp(-a \cdot d(i, i+1))) = 1/m,$$

где

$$d(i, i+1) = t(i+1) - t(i).$$

Имеем

$$\exp(-a \cdot d(i, i+1)) = 1 - \exp(a \cdot t(i))/m,$$

откуда

$$d(i, i+1) = (-1/a) \cdot \ln(1 - \exp(a \cdot t(i))/m), \quad t(0) = 0.$$

Вместе с формулой

$$t(i) = \sum_{k=0}^{i-1} d(k, k+1)$$

это дает рекуррентный алгоритм для вычисления концов интервалов оптимального разбиения.

Последний m-ый интервал будет бесконечным, т.е. $d(m-1, m) = \infty$. Если, как обычно в спектрометрии, $a = \ln(2)/T12$, где T12 - период полураспада, то легко вывести явные выражения для t(i) и d(i, i+1):

$$d(0, 1) = T12 \cdot \ln(m/(m-1))/\ln(2)$$

$$d(1, 2) = T12 \cdot \ln((m-1)/(m-2))/\ln(2) \qquad /3/$$

. . . .

$$d(i, i+1) = T12 \cdot \ln((m-i)/(m-i-1))/\ln(2)$$

. . . .

Перейдем теперь к погрешностям 2-го рода.

Будем предполагать, что оптимальное разбиение области X на интервалы $<Xk>$, k = 1, ..., m уже произведено, и теперь надо выбрать те точки xk внутри каждого интервала, в которых ордината yk, равная приближенно Mk/Lk, будет представлять значение непрерывной функции s(xk) в оптимальном смысле. Погрешность такого представления будет функцией расстояния s(x) от yk на интервале Xk, и задачу оптимизации 2-го рода можно сформулировать, очевидно так: выбрать xk такое, чтобы расстояние от s(x) до константы s(xk) на интервале Xk было минимальным.

Зададимся, как и ранее, квадратичной метрикой. Критерием оптимальности 2-го рода можно будет записать так:

$$xk = \underset{xk}{ARG\,MIN} \int_{Xk} (s(x) - s(xk))^2 dx.$$

Это вариационная задача, которая в классе непрерывных функций s(x) для конечного Xk имеет просто вычисляемое решение. Им будет аргумент

$$s(xk) = (1/Lk) \cdot \int_{Xk} s(x)dx, \qquad /4/$$

т.е. в качестве оптимального в смысле квадратичной метрики узла ячейки следует выбирать точку, в которой s(x) равна своему интегральному среднему на интервале /это будет не обязательно середина интервала, как обычно принято считать для гистограмм/.

Если среди интервалов Xk есть бесконечные, то квадратичная метрика не подходит. Возьмем для определенности метрику C. Нетрудно убедиться, что решением задачи

$$xk = \underset{Xk}{ARG\,INF}\ \underset{x}{SUP}\ ABS(s(x) - s(xk))$$

будет аргумент

$$s(xk) = (\underset{x}{SUP}\ s(x) - \underset{x}{INF}\ s(x))/2, \qquad /5/$$

т.е. в качестве оптимального в смысле равномерной метрики узла ячейки следует выбирать аргумент амплитудного среднего значения функции на интервале /тоже в общем случае не совпадающий с его серединой/. В обоих случаях таких аргументов может быть несколько. Выбор одного из них возможен только на основе дополнительных условий задачи, о чем речь пойдет ниже. Заметим сразу, что оптимизация гистограммирования в смысле погрешности 2-го рода заинтересована в общем случае в максимально возможном сужении интервалов.

ПРИМЕР 1. Случай чистой экспоненты /продолжение/.

Определим теперь узлы tk оптимальных интервалов Tk. Будем полагать, что a = ln(2)/T12. На основании /3/ и /4/ имеем

$$\ln(2)\cdot \exp(-a\cdot t(i-1))\cdot(1-\exp(-\ln((m-i+1)/(m-i)))/(T12\cdot\ln((m-i+1)/(m-i)) =$$

$$= \ln(2)\cdot\exp(-a\cdot t(i-1))/(m-i+1)\cdot\ln((m-i+1)/(m-i))\cdot T12 =$$

$$= \exp(-a\cdot(t(i-1)+h)).$$

Отсюда находим

$$h = (T12/\ln(2))\cdot\ln(T12\cdot\ln((m-i+1)/)m-i))\cdot(m-i+1)/\ln(2)) \qquad /6/$$

при $i \neq m$. Для $i = m$, имеем, на основании /5/,

$$a\cdot\exp(-a\cdot t(m-1))/2 = a\cdot\exp(-a\cdot(t(m-1)+h)).$$

Отсюда получаем

$$h = T12. \qquad /7/$$

ПРИМЕР 2. Гауссова функция.

Гауссиан

$$s(x) = (1/SQRT(2\cdot pi)\cdot w)\cdot\exp(-(x-p)^2/w^2)$$

является немонотонной функцией; кроме того, первообразная в явном виде от него не берется. Ввиду этого построить явные формулы для вычисления концов m интервалов равного веса не удается. Здесь можно указать следующий алгоритм. Пусть $m = 2\cdot k + 1$. Решаем численно или с помощью таблиц уравнение

$$\int_{p-h}^{p+h} s(x)dx = (1/m)$$

относительно h и полагаем $t(1) = p - h$, $t(2) = p + h$; затем процесс решения уравнений

$$\int_{t(i-1)}^{t(i)} s(x) = (1/m)$$

продолжаем по обе стороны от точки и получаем последовательно пары интервалов $<t(i-1), t(i)>$, $<t(j-1), t(j)>$ и т.д.

Точки t(i), t(j) при $i = j = k$ зададут концы двух бесконечных интервалов. Если $m = 2\cdot k$, то процесс следует начать, положив $t(1) = p$ и вычисляя пары интервалов $<t(i-1), t(i)>$, $<t(j-1), t(j)>$ и т.д.

Такой алгоритм переносит на ординаты конечных ячеек гистограммы важное свойство функции - симметрию относительно точки p.

Аналогично численно или с помощью таблиц определяются узлы ячеек. При этом узлы бесконечных ячеек определяются явно

$$h = SQRT(L^2 + 2\cdot w^2\cdot\ln(2)) - q,$$

где q равно одному из (t(k) - p).

Если $m = 2\cdot k + 1$, то возможных узлов ячейки $<t(1), t(2)>$ будет два: слева от p, и справа от p. Если выбрать любой, симметрия гистограммы нарушится. Можно взять в качестве узла их полусумму, т.е. точку p, но и тогда ордината гистограммы будет представлять систематически заниженное значение $s(x)$ в точке максимума, т.е. возникнет систематическая ошибка представления.

ПРИМЕР 3. Лоренцева функция.

Примеру 2 аналогичен случай функции $s(x) = 1/(w \cdot 2 \cdot pi \cdot (1+z^2))$, где $z = (x-p)/w$. Но здесь задача построения интервалов равного веса решается легче, так как первообразная от $s(x)$ существует в явном виде. Например, если $m = 2 \cdot k + 1$, то, применяя вышеизложенный алгоритм, получим

$$\int_{p-h}^{p+h} s(x)\,dx = (1/2 \cdot pi) \cdot arctg(h/w) = 1/m.$$

Отсюда $h = w \cdot tg(1/m)$. Далее для $i = 2,\ldots, m-1$ имеем

$$\int_{t(i-1)}^{t(i)} s(x)\,dx = (arctg(t(i)/w) - arctg(t(i-1)/w))/(2 \cdot pi) = 1/m.$$

Отсюда последовательно находим

$$t(i) = w \cdot tg(2 \cdot pi/m + arctg(t(i-1)/w)).$$

Аналогично находятся интервалы слева от точки p.

Узлы ячеек $x(k)$ можно найти по формулам

$$x(k) = p + w \cdot SQRT((t(k)-t(k-1))/(arctg(t(k)/w) - arctg(t(k-1)/w)) - 1).$$

Узел бесконечной ячейки можно определить так:

$$x(m) = p + w \cdot SQRT(2 \cdot (1+z^2)), \qquad z = (t(m-1)-p)/w.$$

Здесь так же, как и в случае гауссовой функции, особо решается вопрос об узле центральной ячейки. Если приоритет отдается минимуму погрешности, то решением будет один из двух асимметричных узлов. Если же приоритет отдается симметрии представления, то узлом будет точка p.

Последние примеры продемонстрировали действие дополнительных условий, предъявляемых к гистограммам, таких, как симметрия и несмещенность в точке максимума. Другими такими условиями могут также быть и несмещенность в точке минимума, сохранение площади /интеграла/ функции и другие.

На практике при измерении спектров описанные методы гистограммирования могут быть применены на основании априорной информации об измеряемой функции $s(x)$ следующим образом:

1/ иногда настройкой измерительной аппаратуры;

2/ чаще при построении интегральных распределений из отдельных событий;

3/ и, наконец, при преобразованиях одних координат спектров к другим.

Оптимальное гистограммирование может существенно повысить качество анализа данных с малой статистикой.

В массовых измерениях многоканальных спектров ячейки имеют, как правило, одинаковую ширину, и здесь задача оптимального гистограммирования является задачей оптимального представления гистограммой функции, заданной интегралами на фиксированных интервалах с фиксированными погрешностями 1-го рода.

В заключение остается рассмотреть еще вопрос о погрешности гистограммирования в тех случаях, когда по техническим причинам или вследствие отсутствия априорной информации ячейки гистограммы и узлы этих ячеек фиксированы. Итак, пусть Xk - к-ая ячейка гистограммы $s(k)$ функции $s(x)$ и xk - фиксированный узел этой ячейки. Требуется оценить норму разности $s(xk) - s(k)$. Будем считать, что измерение дает нам приближенное значение интеграла $\int s(x)\,dx$ по интервалу Xk.

Пусть это будет величина yk.

С другой стороны, такой же оценкой является величина $s(k) \cdot L$, где L = длина ячейки. Имеем

$$r = s(k) - s(xk) = (1/L) \cdot \int (s(x) - s(xk))\,dx.$$

Имеем следующую оценку

$$ABS(r) \le (1/L) \cdot \int ABS(s'(x) \cdot (xk - x))\,dx = MAX \ ABS(s'(x)) \cdot L/2 \quad x \in Xk.$$

Это оценка погрешности 2-го рода. Полная погрешность будет включать в себя еще и член 1-го рода.

ПРИМЕР. Рассмотренные ранее функции.

Пусть

$$f(t) = a \cdot exp(-a \cdot t), \quad f'(t) = -a^2 \cdot exp(-a \cdot t).$$

Максимум модуля 1-й производной этой функции достигается при $t = 0$ и обратно пропорционален квадрату периода полураспада.

Далее возьмем

$$f(x) = c \cdot exp(-(x-p)^2/(2 \cdot w^2)),$$

где $c = 1/(SQRT(2 \cdot pi) \cdot w)$. Для этой функции максимум модуля 1-й производной достигается в точках

$$x = p +- w/SQRT(2).$$

И наконец, $f(x) = c/(1+(x-p)^2/w^2)$, где $c = 1/(2 \cdot \pi \cdot w)$.

Для этой функции максимум модуля 1-й производной достигается в точках

x = p + – w/SQRT(3).

Качественное заключение изложенного материала можно сформулировать следующим образом.

Математические соображения, лежащие в основе выбора ширины ячеек спектральной гистограммы, отвечают двум противоположным требованиям:

1/ сделать размер ячейки достаточно широким для обеспечения хороших статистических свойств будущей гистограммы;

2/ сделать размер ячейки по возможности узким для того, чтобы изображаемая гистограммой функция описывалась ею достаточно точно внутри каждой ячейки.

ЛИТЕРАТУРА

1. Ченцов Н.Н. Статистические решающие правила и оптимальные выводы. М.: Наука, 1972.

Рукопись поступила в издательский отдел
23 марта 1988 года.