УДК 520.88

ДВУХТОЧЕЧНАЯ КОРРЕЛЯЦИОННАЯ ФУНКЦИЯ И ОЦЕНКА ЕЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

В. С. Лебедев, И. А. Лебедева

Описана методика вычисления доверительных интервалов для численных оценок значений корреляционных функций. Приведены алгоритмы и примеры их использования.

Method of calculation of confidence intervals for quantitative evaluation of correlation function values is described. Calculation algorithms and examples of their application are presented.

Двухточечная пространственная корреляционная функция $\xi(r)$ является одним из методов изучения распределения объектов в пространстве (Пиблс, 1983). Через нее выражается вероятность найти два объекта в элементарных объемах dv_1 и dv_2, разделенных расстоянием, попадающим в интервал от r до r+dr:

$$dP = n^2*(1 + \xi(r))*dv_1*dv_2 \ , \qquad (1)$$

где n – пространственная плотность объектов.

Рабочей формулой, по которой проводятся оценки значений корреляционной функции, служит выражение:

$$\xi(r) = N(r)/Np(r) - 1 \ , \qquad (2)$$

в котором N(r) – число пар объектов, расстояние между которыми попадает в интервал (r, r+dr), Np(r) – аналогичные величины для каталога случайных чисел, имеющего такое же число объектов K, что и в реальном каталоге. Организация вычислений по этой формуле не представляет трудностей. Требуется только иметь возможность генерировать случайное распределение точек в том же объеме пространства, что занимает изучаемая выборка реальных объектов. При этом желательно привнести в случайную выборку все эффекты наблюдательной селекции, которые имеются в реальной выборке. Таковыми могут быть ограничения по наблюдаемой звездной величине, угловым размерам или поверхностной яркости объектов, приводящие к градиенту средней пространственной плотности объектов с расстоянием от наблюдателя, или поглощение в Галактике, приводящее к тренду поверхностной плотности с галактической широтой. Число пар взаимных расстояний пропорционально квадрату числа объектов в выборке K и при больших объемах выборок могут потребоваться значительные затраты машинного времени. Некоторая трудность, связанная с появлением в случайном каталоге пустых интервалов, в которые не попадает ни одно расстояние (т. е. $Np(r) = 0$), обходится путем увеличения шага dr или использованием вместо $Np(r)$ его среднего значения, полученного по M случайным реализациям:

$$\langle Np(r)\rangle = \left(\sum_{m=1}^{M} Np_m(r)\right)/M. \tag{3}$$

При этом выбором dr или M фактически добиваемся того, чтобы Np стало отличным от нуля на всех интервалах r.

Если в нашем распоряжении имеется несколько независимых оценок корреляционных функций L, то можно найти их среднее значение и его среднеквадратичное отклонение:

$$\langle\xi(r)\rangle = \left(\sum_{\ell=1}^{L} \xi_\ell(r)\right)/L, \tag{4}$$

$$\delta_\xi(r) = \left(\sum_{\ell=1}^{L} (\xi_\ell(R) - \langle\xi(R)\rangle)^2/L/(L-1)\right)^{1/2}. \tag{5}$$

Поиск доверительных интервалов для исходных оценок корреляционных функций затруднен тем обстоятельством, что в отличие от случайного каталога выборка реальных объектов, как правило, единственна. Некоторое представление о величине доверительного интервала для $\xi(r)$ можно получить из разброса подсчетов $Np(r)$ для множества случайных каталогов. При этом порядок величины разброса для $N(r)$, по-видимому, не сильно отличается от такового для случайного каталога. Тогда нижняя и верхняя границы доверительного интервала для корреляционной функции приблизительно выразятся соотношениями:

$$\xi_1(r) = (N(r) - u_\gamma*\sigma(r))/(Np(r) + u_\gamma*\sigma_p(r)) - 1, \tag{6}$$

$$\xi_2(r) = (N(r) + u_\gamma*\sigma(r))/(Np(r) - u_\gamma*\sigma_p(r)) - 1. \tag{7}$$

При больших N и Np можно использовать простые соотношения

$$\begin{aligned} \sigma(r) &= (N(r))^{1/2}, \\ \sigma_p(r) &= (Np(r))^{1/2}. \end{aligned} \tag{8}$$

Для множества случайных каталогов в качестве оценок $\sigma_p(r)$ можно взять

$$\sigma_p(r) = \left(\sum_{m=1}^{M} (Np_m(r) - \langle Np(r)\rangle)^2/M/(M-1)\right)^{1/2}. \tag{9}$$

В этих соотношениях u_γ – квантили нормального (гауссова) распределения, а γ – доверительная вероятность.

При программной реализации описанных в работе алгоритмов мы использовали процедуру NDTI-пакета научных подпрограмм из "Сборника научных программ на ФОРТРАНе" (1974), позволяющую вычислять квантили нормального распределения в широком диапазоне вероятностей.

Из структуры выражения для корреляционной функции (2) видно, что доверительный интервал для нее можно вычислить из доверительного интервала для отношения

$$p(r) = N(r)/Np(r). \qquad (IO)$$

Выражение (IO) можно переписать иначе:

$$p(r) = \frac{N(r)/N_t}{Np(r)/N_t} = q(r)/qp(r). \qquad (IO\ )$$

где $N_t = K(K-1)/2$ – число пар взаимных расстояний в выборке из K- объектов.

Величины q(r) и qp(r) являются по смыслу параметрами биномиального распределения, поэтому приведенные ниже алгоритмы определения доверительных границ для этого параметра применимы и к этим величинам. Если же найдены доверительные границы (q1, q2 и qp1, qp2) для величин q и qp с доверительной вероятностью γ, то в предположении независимости величин q и qp легко найти доверительные границы для их отношения P с доверительной вероятностью $\gamma^{\star}=1-(1-\gamma)^2$:

$$p1 = q1/qp2; \quad p2 = q2/qp1.$$

Это наиболее последовательный метод оценки доверительных интервалов для корреляционной функции. Но, в первом приближении, можно и саму величину p(r), определенную из соотношения (IO), при N(r) < Np(r) считать параметром некоторого гипотетического биномиального распределения. Если же N(r) > Np(r), то в качестве такого параметра следует взять отношение

$$p(r) = Np(r)/N(r). \qquad (II)$$

Правила же определения доверительных границ для параметра биномиального распределения хорошо известны. В биномиальной схеме при общем числе испытаний n и числе испытаний, в которых появляется событие A – n1, оценка параметра p и доверительных границ для него p1 и p2 проводится следующим образом:

$$p = n1/n. \qquad (I2)$$

При n1 = O, в качестве доверительных границ берут величины

$$p1 = 0, \quad p2 = 1 - (1 - \gamma)^{1/n}. \qquad (I3)$$

При n1 = n,

$$p1 = (1 - \gamma)^{1/n}, \quad p2 = 1. \qquad (I4)$$

Наконец, при n1 $\neq$ O и n1 $\neq$ n доверительные пределы находятся из общих выражений

$$\sum_{x=0}^{n1-1} C_n^x * p1^x * (1-p1)^{n-x} = \gamma 1, \qquad (I5)$$

$$\sum_{x=0}^{n1} C_n^x * p2^x * (1-p2)^{n-x} = 1-\gamma 2, \qquad (I6)$$

где C_n^x – число сочетаний из n по x.
Если задаться двусторонней доверительной вероятностью $\gamma_\star$ и предположением о ее симметричном разделении, то значения односторонних доверительных вероятностей находятся из соотношения

$$\gamma = \gamma 1 = \gamma 2 = (\gamma_\star + 1)/2. \qquad (17)$$

Вычисления по формулам (I5) и (I6) можно организовать методом итераций. Для того, чтобы избежать переполнений и исчезновения порядков при вычислении составляющих сомножителей в суммах, вычисления следует проводить для логарифмов этих сомножителей, а при вычислении факториалов в биномиальных коэффициентах использовать процедуру определения логарифма гамма-функции.

Вместо точных вычислений по (I5) и (I6) часто используют приближенные алгоритмы. В частности, ГОСТом 11.010-81 (I986) рекомендован следующий алгоритм. При $1 < n1 < n2 = n - n1$, для доверительных границ имеем

$$p1 = 2\star z1/(2\star t1+z1-(2\star(n1^2-1)+(n1-1)\star z1-z1^2)/(6\star t1)), \qquad (I8)$$

$$p2 = 2\star z2/(2\star t2+z2-(2\star(n1^2+2\star n1)+n1\star z2-z2^2)/(6\star t2)), \qquad (I9)$$

где

$$t1 = 2\star n - n1 + 1,$$
$$t2 = 2\star n - n1, \qquad (20)$$

а z1 и z2 – квантили распределения хи-квадрат $z1 = \chi^2_{1-\gamma}(2\star n1)$, $z2 = \chi^2_\gamma(2\star n1+2)$, значения которых приведены в статистических таблицах, например, в книге Большева и Смирнова (I983). При $2\star n1 > 30$, значения z1 и z2 можно приближенно вычислить через квантили нормального распределения.

$$z1 = 2\star n1\star(1-1/(9\star n1)-u_\gamma/(9\star n1)^{1/2})^3,$$
$$z2 = 2\star(n1+1)\star(1-1/(9\star(n1+1))+u_\gamma/(9*(n1+1))^{1/2})^3. \qquad (2I)$$

При $n1 > n2$

$$p1 = 1 - \bar{p}2,$$
$$p2 = 1 - \bar{p}1, \qquad (22)$$

где p1 и p2 получены, как описано выше, с заменой n1 на n2, а n2 на n1.

В работе Гехрелс (I986) приведена таблица значений p2 для 10 значений доверительной вероятности и ряда значений n1 и n2 (0–10, 100). Точное выражение

$$p2 = \begin{cases} \gamma^{1/(n1+n2)} & \text{при } n2 = 1, \\ 1-(1-\gamma)^{1/n2} & \text{при } n1 = 0. \end{cases} \tag{23}$$

И приближенное с точностью 4 %

$$p2 = ((n1+1)*\exp(2*w)+e*n2)/((n1+1)*\exp(2*w)+n2), \tag{24}$$

где

$$w = u*(h+t)^{1/2}/h+(1/(2*n2-1)-1/(2*n1+1))*(t+5/6-2/3/h), \tag{25}$$

$$h = 2/(1/(2*n2-1)+1/(2*n1+1)), \tag{26}$$

$$t = (u_\gamma^2-3)/6, \tag{27}$$

$$e = 0.64*(1-u_\gamma)*\exp(-n2). \tag{28}$$

Для случая, когда значения n1 и n2 больше 50, с ошибкой, не превосходящей 1 %, имеем более простое выражение:

$$p2 = n1/(n1+n2)+u_\gamma *(n1*n2)^{1/2}/(n1+n2)^{1/2}. \tag{29}$$

Значения нижнего доверительного предела находятся с использованием соотношения

$$p1 = 1 - \overline{p}2 , \tag{30}$$

где $\overline{p}2$ – верхний предел, найденный по приведенным выше формулам для случая, когда n1 и n2 переставлены местами.

Из доверительных интервалов p1 и p2 для параметра биномиального распределения легко получить доверительные интервалы ξ1 и ξ2 для корреляционной функции. При N(r)<Np(r) оценивались доверительные интервалы для p=N(r)/Np(r), поэтому

$$\xi 1 = p1 - 1, \tag{31}$$
$$\xi 2 = p2 - 1.$$

А при N(r)>Np(r) доверительные интервалы оценивались для параметра p=Np(r)/N(r) и тогда

$$\xi 1 = 1/p2 - 1, \tag{32}$$
$$\xi 2 = 1/p1 - 1.$$

По трем описанным методикам мы для иллюстрации рассчитали 99 %-ные доверительные интервалы для значений N = 10, 100, 1000, 10000 и для Np = 5, 50, 500, 5000. При этих значениях ожидаемая величина корреляционной функции равна 1. Значения доверительных интервалов для корреляционной функции приведены в таблице.

Доверительные интервалы, вычисленные с использованием методик (18-19) и (24-29), получились близкими друг другу, в то время как при использовании простых соображений (6) и (7) эти интервалы оказались завышенными, чего и следовало ожидать.

Таблица. 99 %-ные доверительные интервалы

N	Np	ξ1 (6-7)	ξ2	ξ1 (18-19)	ξ2	ξ1 (24-29)	ξ2
10	5	–	–	0.143	6.900	0.147	6.680
100	50	0.088	2.957	0.584	1.714	0.584	1.711
1000	500	0.647	1.445	0.848	1.180	0.849	1.177
10000	5000	0.880	1.129	0.950	1.053	0.950	1.053

Значения доверительных границ, рассчитанные по приведенным методикам, относятся только к статистическим соображениям и ничего не говорят о причинах, приводящих к тем или иным значениям корреляционной функции. Среди таких причин, кроме действительного скучивания объектов в пространстве, могут быть и различные селекционные эффекты, приводящие к крупномасштабным неоднородностям, которые также проявляются в значимых отклонениях корреляционной функции от нуля. Для решения вопроса о происхождении даже статистически значимых отклонений корреляционной функции требуется выходить за пределы чисто статистического рассмотрения.

Наблюдательные данные для внегалактических объектов имеют две громадные неоднородности, связанные с особенностями наблюдений Вселенной с поверхности земного шара, т. е. из точки, находящейся внутри Галактики. Одна из этих особенностей заключается в существовании зоны избегания для внегалактических объектов, связанной с поглощением оптического излучения пылевой составляющей, концентрирующейся к плоскости Галактики. Другая особенность связана с ограничением доступной наблюдениям области неба при наблюдениях на одном инструменте. Сочетание этих двух эффектов приводит к тому, что область неба, доступная для исследований, становится двусвязанной и имеет несимметричные границы. В случае, когда наблюдениями из нескольких пунктов удается избавиться от неоднородности второго типа, двусвязанная область становится симметричной и совпадает с областями северного и южного галактических полушарий ($|b|>b_o$). Анализ и интерпретация результатов пространственного распределения объектов упрощаются для случая односвязанной области, в частности, заданной объемом V = (a1, a2)*(d1,d2)*(r1,r2).

Для того, чтобы получить пуассоновский случайный каталог объектов в этом объеме, поступают одним из двух способов:

1) находят прямоугольный параллелепипед, включающий в себя объем V;

генерируют декартовы координаты точек, случайно разбросанных в этом параллелепипеде; от декартовых координат переходят к сферическим и проверяют условия принадлежности точек объему V; при таком подходе коэффициент использования датчика случайных чисел равен отношению объема области V к объему описывающего параллелепипеда;

2) величины α, равномерно распределенные на отрезке (0, 1), преобразуют таким образом, чтобы получить равномерное распределение точек в трехмерном объеме, представляющем собой часть сферы; это достигается при следующих преобразованиях:

$$a - (a2 - a1)\star\alpha + a1, \tag{33}$$

$$d = \arcsin((\sin(d2) - \sin(d1))\star\alpha + \sin(d1)), \tag{34}$$

$$r = ((r2^3 - r1^3)\star\alpha + r1^3)^{1/3}. \tag{35}$$

При втором подходе коэффициент использования случайных величин равен 1, но требуются некоторые вычисления, объем которых, однако, не превосходит тех, что необходимы в первом способе для пересчета координат из декартовой системы в сферическую.

Еще одним источником глобальной наблюдательной неоднородности являются селекции данных по звездным величинам объектов, их угловым размерам и поверхностным яркостям. Для того, чтобы освободить оценки корреляционных функций от влияния этих наблюдательных селекций, требуется и случайные каталоги генерировать с учетом этих селекций. В случае, когда анализируется выборка объектов с предельной наблюдаемой звездной величиной m_{lim}, учет влияния этого обстоятельства можно провести следующим образом. Предположим, что объекты исследуемой выборки имеют некоторую известную функцию светимости $\varphi(L)$. Тогда процедуру генерации случайного каталога следует дополнить пунктом, содержащим розыгрыш случайной величины, распределенной на отрезке (L1, L2) по закону, с точностью до нормировки совпадающему с дифференциальной функцией светимости. А имея уже и расстояние до объекта случайного каталога, найдем его звездную величину m:

$$m = M_{\odot} + 5\star\lg(r) - 2.5\star\lg(L) + 25 , \tag{36}$$

где $M_{\odot}$ – абсолютная звездная величина Солнца в фотометрической системе, в которой выделены объекты до предельной звездной величины. И если объект оказывается слабее m_{lim}, то он не включается в случайный каталог. Аналогичную процедуру следует применять и при учете эффекта селекции из-за ограничений на угловые размеры, только вместо распределения по светимости нужно задаться распределением объектов по абсолютным линейным размерам. Для более детального рассмотрения методики освобождения оценок корреляционной функции от избирательности данных по угловым размерам необходимо учитывать и распределение яркости по диску галактики, и зависимость его от морфологического типа галактики, и ориентацию галактик в пространстве. Аналогично, но несколько более громоздко, можно освободиться от влияния селекции по поверхностной яркости.

Для функции светимости, интегральное (комулятивное) представление которой легко разрешается относительно светимости L, случайную величину, распределенную по закону $\varphi(L)$, можно генерировать методом обратных функций (Соболь, I973). Но часто используемое представление $\varphi(L)$ в виде функции Шехтера (I976) не разрешается относительно L. В этом случае следует использовать более общие методы генерирования случайных величин, например, метод Неймана (отказов, отбора, элиминации). Функция Шехтера имеет вид

$$\varphi(L) = A*L^{-g}*\exp(-L/L_o), \tag{37}$$

где A, g, L_o– численные параметры.
Для нее в качестве мажорирующей функции естественно взять степенную часть

$$\Phi(L) = A*\exp(-L1/L_o)*L^{-g}. \tag{38}$$

Процедура получения случайной величины L включает следующие пункты:
1) разыгрываем случайную величину α, равномерно распределенную на отрезке (0, 1);
2) вычисляем $L = (L2-L1)*\alpha + L1$;
3) вычисляем значения $\varphi(L)$ и $\Phi(L)$;
4)разыгрываем еще одно значение равномерно распределенной случайной величины α;
5) вычисляем значение $y = \Phi(L)*\alpha$;
6) если выполняется условие $y<\varphi(L)$, то найденное в пункте 2 значение L есть искомая величина, распределенная по закону $\varphi(L)$, иначе – снова начинаем выполнение с пункта 1.

Мы пробовали и другой метод учета эффектов селекции, приводящих к тренду средней плотности объектов с расстоянием. Он заключается в том, что поведение средней плотности с расстоянием аппроксимируется простой функцией, например,

$$n(r) = n_o*\exp(-r/r_o). \tag{39}$$

И случайный каталог генерируется с таким расчетом, чтобы поведение средней плотности его объектов также описывалось выражением (39).
Для этого достаточно сконструировать генератор случайных чисел, распределенных по закону (39).

Для получения случайных каталогов требуется датчик псевдослучайных чисел α, равномерно распределенных на отрезке (0, 1). Теория и практические советы по конструированию и тестированию таких датчиков изложены во многих монографиях и учебниках по численным методам и методам статистических испытаний (Иванова, I984; Кнут, I977; Поллард, I982; Полляк, I97I; Поляченко, I987; Харин и Степанова, I987).

В "Сборнике научных программ на ФОРТРАНе"(I974) и в работах Агеева, и др.(I978); Виталиева и др. (I980), Гуснина и др. (I98I); Кармалиты (I986), Серебрякова и Шульмейстера (I987); Форсайта и др.(I980) приведены вычислительные программы, реализующие некоторые типы датчиков. Относительно любого конкретного программного датчика случайных чисел, по-видимому, можно утверждать, что найдется область применения, для которой он не пригоден. Например, известный и широко распространенный датчик RANDU (Сборник научных программ на Фортране, I974) в системе IBM/360 и ОС ЕС совершенно непригоден для оценки пространственной корреляционной функции, поскольку последовательные тройки чисел, которые он генерирует, связаны между собой функциональной зависимостью, приводящей к тому, что в трехмерном пространстве точки располагаются на нескольких поверхностях. Количественно функциональная зависимость выражается в виде:

$$\alpha_{i+2} = 6\star\alpha_{i+1} - 9\star\alpha_i (\mathrm{mod}\ m), \qquad (40)$$

где (mod m) означает, что вычисления проводятся с отбрасыванием старших разрядов и оставление только m младших. Значение m является параметром алгоритма и выбирается в зависимости от разрядности используемой ЭВМ.

Мы рекомендуем использовать датчики, алгоритмы и программы которых на языке ПЛ/1 приведены в работе Виталиева и др. (I980). Но перед использованием любого датчика целесообразно провести его проверку. Проверка качества генератора случайных чисел проводилась нами по некоторым методикам,описанным в книге Кнута (I977) (критерий хи-квадрат для распределений точек на отрезке, квадрате и кубе, критерии серий, разностей, Колмогорова-Смирнова, на основе вычислений автокорреляционной функции). Еще одна проверка качества случайных чисел и всего алгоритма определения корреляционных функций проводилась путем вычисления пространственной корреляционной функции для той же самой выборки богатых эйбелловских скоплений галактик в северном конусе Метагалактики, для которой в работе Копылова и др. (I987) приводится корреляционная функция. Оказалось, что различия корреляционных функций, вычисленных по двум разным программам, связаны только с неопределенностью попадания взаимных расстояний в соседние интервалы из-за ошибок вычислений и округлений, связанных с различной разрядностью используемых вычислительных машин. По описанной методике проводились вычисления пространственной корреляционной функции и ее доверительных интервалов для сверхскоплений галактик (Лебедев и Лебедева, I988) и других подвыборок объектов различной природы (галактик, скоплений галактик, квазаров).

Таким образом, в настоящей работе рассмотрены практические вопросы, возникающие при вычислении корреляционной функции и доверительных интервалов для нее. Приведены алгоритмы программ для обеспечения этих вычислений.

ЛИТЕРАТУРА

Агеев М. И., Алик В. П., Марков Ю. И.: I978, Библиотека алгоритмов 101б–150б. Вып. 3. М.: Советское радио, I28 с.

Большев Л. Н., Смирнов Н. В.: I983, Таблицы математической статистики. М.: Наука, 4I6 с.

Виталиев Г. В., Жуков А. В., Чугунов А. П.: I980, Вопросы радиоэлектроники. Электронная вычислительная техника. Вып. I. С. IOI-II2.

Гехрелс (Gehrels N.): I986, Astrophys. J., V. 303. P. 336-346.

ГОСТ 11.010-81 (СТ СЭВ 5313-85). Правила определения оценок и доверительных

границ для параметров биномиального и отрицательного биномиального распределений: I986, М.: Издательство стандартов, 24 с.
Гуснин С. Ю., Омельянов Г. А., Резников Г. В., Сироткин В. С.: I98I, Минимизация в инженерных расчетах на ЭВМ. М.: Машиностроение, I20 с.
Иванова В. М.: I984, Случайные числа и их применение. М.: Финансы и статистика, II2 с.
Кармалита В. А.: I986, Цифровая обработка случайных колебаний, М.: Машиностроение, 80 с.
Копылов А. И., Кузнецов Д. Ю., Фетисова Т. С., Шварцман В. Ф.: I987, Активные ядра и звездная космогония. М.: МГУ С. I49–I57.
Кнут Д.: I977, Искусство программирования для ЭВМ. 2. Получисленные алгоритмы. М.: Мир, 724 с.
Лебедев В. С., Лебедева И. А.: I988, Письма в Астрон. ж., I4, С. I8–22.
Пиблс Ф. Дж. Э.: I983, Структура Вселенной в больших масштабах. М.: Мир, 408 с.
Поллард Дж.: I982, Справочник по вычислительным методам статистики. М.: Финансы и статистика, 344 с.
Полляк Ю. Г.: I97I, Вероятностное моделирование на электронных вычислительных машинах. М.: Советское радио, 400 с.
Поляченко А. Л.: I987, Численные методы в ядерной геофизике. М.: Энергоатомиздат, I52 с.
Сборник научных программ на ФОРТРАНе. Вып. 1. Статистика: I974, М.: Статистика, 3I6 с.
Серебряков А. С., Шульмейстер В. М.: I987, Программирование, № 7, С. I03–I04.
Соболь И. М.:I973, Численные методы Монте-Карло. М.: Наука, 3I2 С.
Форсайт Дж., Малькольм М., Моулер К.: I980, Машинные методы математических вычислений. М.: Мир, 280 с.
Харин Ю. С., Степанова М. Д.: I987, Практикум на ЭВМ по математической статистике. Минск: Университетское, 304 с.
Шехтер (Schechter P.): I976, Astrophys. J., V. 203. P. 297–306.

Поступила в редакцию
I марта I989 г.