РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 4 2001

УДК 519.816

С. Н. Моисеев
(Воронеж)

ХАРАКТЕРИСТИКИ ВЗАИМОСВЯЗИ СЛУЧАЙНЫХ ВЕЛИЧИН ИЗ РАСПРЕДЕЛЕНИЙ С ТЯЖЕЛЫМИ ХВОСТАМИ

Предложены информационные коэффициенты корреляции, построенные на основе информационного расстояния Кульбака – Лейблера, для анализа взаимосвязи случайных величин из абсолютно непрерывных распределений, в том числе с тяжелыми хвостами. С помощью предложенных характеристик проведен анализ структуры взаимосвязи случайных величин, описываемых некоторыми известными вероятностными моделями.

Введение. Максимальные значения электронной концентрации среднеширотного спорадического слоя нижней ионосферы распределены по вероятностному закону Коши, а их динамика во времени описывается нелинейным стохастическим дифференциальным уравнением [1]. Для прогноза максимальных значений электронной концентрации, с целью расчета ионосферных линий УКВ радиосвязи, необходим тщательный анализ взаимозависимости отсчетов в разнесенные моменты времени. Однако из-за тяжелых хвостов распределения Коши традиционно и широко используемые характеристики парной стохастической линейной связи – коэффициент корреляции и нелинейной связи – корреляционное отношение в данном случае не существует. Подобного рода затруднения возникают также при анализе телекоммуникационных трафиков, эконометрических данных и т. п. [2, 3].

В данной работе на основе сформулированных требований предложены новые и рассмотрены некоторые уже известные характеристики стохастической взаимосвязи между случайными величинами, в том числе из распределений с тяжелыми хвостами, способные выявлять не только линейные, но и нелинейные связи. С помощью рассмотренных характеристик будет проанализирован ряд известных вероятностных моделей, не всегда поддающихся анализу традиционными методами.

Характеристики взаимосвязи. Сформулируем основные требования к свойствам характеристики взаимосвязи $R(X, Y)$ между произвольными случайными величинами X и Y из абсолютно непрерывных распределений.

1. $0 \le R(X, Y) \le 1$.
2. $R(X, Y) = R(Y, X)$.
3. $R(X, Y) = 0$, если X и Y независимы.
4. Если $R(X, Y) = 0$, то X и Y независимы.

5. При функциональной связи между случайными величинами $Y = f(X)$, где $f(x)$ – произвольная неслучайная функция, $R(X, Y) = 1$.

6. Для случайных величин $U = f(X), V = g(Y)$, где $f(x)$ и $g(y)$ – взаимно однозначные функции, $R(U, V) = R(X, Y)$.

7. Для совместно гауссовских случайных величин X и Y с коэффициентом корреляции r $R(X, Y) = |r|$.

Перечисленные требования являются достаточно жесткими – им не удовлетворяет подавляющее большинство традиционно используемых характеристик взаимосвязи. Поэтому для синтеза характеристик требуется достаточно мощный аналитический аппарат. Такой аппарат предоставляет информационное расстояние Кульбака – Лейблера, которое лежит в основе многих оптимальных процедур математической статистики [4] и тесно связано с теорией информации.

Расстояние Кульбака – Лейблера между абсолютно непрерывными относительно меры Лебега распределениями с плотностями вероятностей $p(\mathbf{x})$ и $q(\mathbf{x})$ определяется следующим образом:

$$\rho[p(\mathbf{x}), q(\mathbf{x})] = \mathbf{M}[\ln p(\mathbf{x}) - \ln q(\mathbf{x})] = \int_{S_p} p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} \mu(d\mathbf{x}),$$

где $\mathbf{M}[\cdot]$ – математическое ожидание плотности $p(\mathbf{x})$, $S_p = \{\mathbf{x} : p(\mathbf{x}) > 0\}$ – носитель распределения $p(\mathbf{x})$, μ – мера Лебега. Расстояние Кульбака – Лейблера обладает свойством $\rho[p(\mathbf{x}), q(\mathbf{x})] \geq 0$, причем равенство достигается тогда и только тогда, когда $p(\mathbf{x}) \equiv q(\mathbf{x})$. Полагая $p(\mathbf{x}) = p(x, y)$, $q(\mathbf{x}) = p_X(x) p_Y(y)$ (или, наоборот, $p(\mathbf{x}) = p_X(x) p_Y(y)$, $q(\mathbf{x}) = p(x, y)$), где $p(x, y)$ – совместная двумерная плотность вероятностей случайных величин X и Y с соответствующими маргинальными одномерными плотностями $p_X(x)$ и $p_Y(y)$, видим, что расстояние $\rho[\cdot]$ естественным образом характеризует взаимосвязь случайных величин X и Y. С помощью расстояния Кульбака – Лейблера можно построить много характеристик парной взаимосвязи случайных величин, из которых были отобраны следующие наиболее простые:

$$R_1^2(X, Y) = 1 - \exp(-2\rho_1), \tag{1}$$

$$R_2^2(X, Y) = \frac{2\rho_2}{1 + 2\rho_2}, \tag{2}$$

$$R_3^2(X, Y) = \frac{\rho_1 + \rho_2}{1 + \rho_1 + \rho_2}, \tag{3}$$

где $\rho_1 = \rho[p(x, y), p_X(x) p_Y(y)]$, $\rho_2 = \rho[p_X(x) p_Y(y), p(x, y)]$. Расстояние Кульбака – Лейблера ρ_1 в теории информации называется средней взаимной информацией, содержащейся в случайных величинах X и Y относительно друг друга. Расстояние ρ_2 в этих терминах является отрицательной взаимной информацией, усредненной в предположении, что она равна нулю.

Характеристики взаимосвязи R_i, $i = 1, 2, 3$, удовлетворяют первым шести перечисленным выше требованиям, доказательства которых достаточно оче-

видны и здесь не приводятся. Рассмотрим более подробно седьмое требование. Для совместно гауссовских случайных величин с коэффициентом корреляции r легко получить

$$\rho_1 = -\frac{1}{2}\ln(1-r^2), \qquad \rho_2 = \frac{1}{2}\ln(1-r^2) + \frac{r^2}{1-r^2}.$$

Отсюда следует, что седьмое требование в полной мере удовлетворено только для R_1 и R_3. Для R_2, учитывая, что $\ln(1-r^2) = -\frac{r^2}{1-r^2} + O(r^4)$ и, следовательно, $\rho_2 = \frac{r^2}{2(1-r^2)} + O(r^4)$, справедливо асимптотическое соответствие $R_2 \to |r|$ при $r \to 0$. Максимальное относительное отклонение R_2 от $|r|$ достигается в точке $|r| = 0{,}700\ldots$ и не превышает 6,5 %.

Характеристика, подобная R_1, уже встречалась в литературе [5] под названием информационного коэффициента корреляции. Аналогично характеристики $R_1(\cdot)$, $R_2(\cdot)$, $R_3(\cdot)$ назовем информационными коэффициентами корреляции 1–3-го рода соответственно.

Наряду с введенными информационными коэффициентами корреляции для анализа распределений с тяжелыми хвостами, заслуживает внимания коэффициент конкорреляции, предложенный в [6]. Модуль коэффициента конкорреляции определяется следующим образом:

$$R_4(X, Y) = \frac{|K_{XY}|}{\sqrt{K_{XX} K_{YY}}}, \tag{4}$$

где $K_{XY} = \mathbf{M}\{[F_X(X) - \mathbf{M}F_X(X)][F_Y(Y) - \mathbf{M}F_Y(Y)]\}$, $F_X(x)$ и $F_Y(y)$ – интегральные функции распределения случайных величин X и Y. Характеристика (4) не удовлетворяет сформулированным требованиям 4 и 7, однако ее несомненным достоинством является простота оценивания по выборке наблюдаемых данных.

Используя теорему о среднем, для информационного коэффициента корреляции 1-го рода $R_1(X, Y)$ можно получить еще одно полезное представление:

$$R_1^2(X, Y) = 1 - \frac{\operatorname{med} V^2_{Y|X}(u)}{V^2_Y(\mathrm{v})}, \quad u, \mathrm{v} \in [0{,}5,\ 1], \tag{5}$$

где $V_{Y|X}(u) = Y_{(u)}(X) - Y_{(1-u)}(X)$ и $V_Y(\mathrm{v}) = Y_{(\mathrm{v})} - Y_{(1-\mathrm{v})}$ – интерквантильные широты порядка u, v условного при фиксированном X и безусловного распределения случайной величины Y соответственно, $Y_{(u)}(X)$ и $Y_{(\mathrm{v})}$ – условный при фиксированном X и безусловный квантили величины Y порядка u, v; med – операция взятия медианы.

Примеры использования. Рассмотрим теперь применение характеристик взаимосвязи (1)–(4) для анализа некоторых известных вероятностных моделей, описывающих зависимые случайные величины из распределений с тяжелыми хвостами.

Распределение Коши. Многомерное распределение Коши является частным случаем многомерного распределения Стьюдента (t-распределения) и имеет следующую n-мерную плотность вероятностей:

$$p(\mathbf{x}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi^n \det\mathbf{K}}}\left(1+\mathbf{x}\mathbf{K}^{-1}\mathbf{x}^{\mathrm{T}}\right)^{-\frac{n+1}{2}}, \tag{6}$$

где $\mathbf{x} = \|x_i\|$, $\mathbf{K} = \|r_{ij}\|$, $r_{ii} = 1$, $i, j = \overline{1,n}$; т – символ транспонирования, $\Gamma(\cdot)$ – гамма-функция. Положительно-определенная симметрическая невырожденная матрица $\mathbf{K}$ характеризует зависимость случайных величин $X_1, \ldots, X_n$. Случайные величины из этого распределения формируются по алгоритму: $X_i = \xi_i / |\eta|$, где ξ_i – стандартные совместно гауссовские величины с корреляционной матрицей $\mathbf{K}$, η – стандартная гауссовская величина некоррелированная с ξ_i, $i = \overline{1,n}$. Плотность (6) и все плотности в этой работе выписаны с точностью до несущественных параметров сдвига и масштаба, так как требование к свойству 6 справедливо для всех четырех характеристик (1)–(4). Моментов у распределения (6) не существует. Для анализа парной связи между случайными величинами X_1 и X_2 достаточно рассмотреть двумерную маргинальную плотность (6) $p(x_1, x_2)$ при $n = 2$, где $r_{12} = r$.

Информационный коэффициент корреляции 1-го рода $R_1(X_1, X_2)$, рассчитанный численно, в зависимости от r изображен на рис. 1 (кривая 1). Информационные коэффициенты корреляции 2-го и 3-го рода $R_2(X_1, X_2)$ и $R_3(X_1, X_2)$ практически совпадают в масштабе графика (кривые 2, 3); коэффициент конкорреляции $R_4(X_1, X_2)$ – кривая 4. Поскольку информационные коэффициенты корреляции $R_i(X_1, X_2)$, $i = 1,2,3$, нигде, даже при $r = 0$, не обращаются в нуль, то ни при каких значениях параметров многомерное распределение Коши не может описывать независимые, или даже слабо связан-

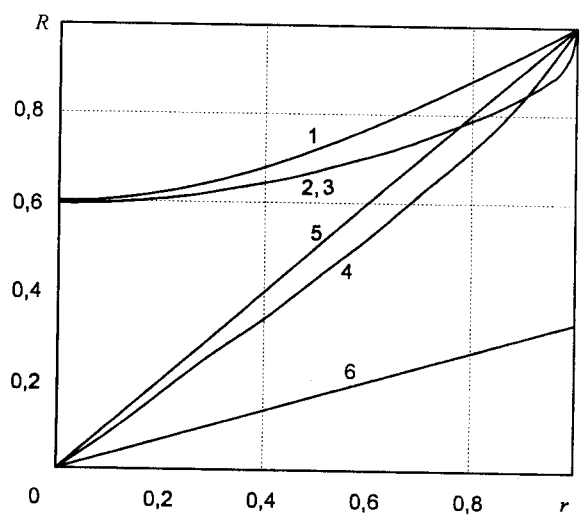

Рис. 1

ные, случайные величины. Коэффициент конкорреляции $R_4(X_1, X_2) = 0$ при $r = 0$, что является следствием невыполнимости для него требования к свойству 4, поэтому он не может быть использован для адекватного описания взаимосвязи случайных величин из многомерного распределения Коши.

Распределение Бурра. n-мерная плотность вероятностей многомерного распределения Бурра имеет вид [7]

$$p(\mathbf{x}) = \frac{\Gamma(1/r + n - 1)}{\Gamma(1/r - 1)} \left(1 + \sum_{i=1}^{n} x_i\right)^{1 - n - 1/r}, \quad r \in [0,1]. \tag{7}$$

Параметр r является параметром формы и одновременно параметром связи распределения. Тяжесть хвостов непосредственно зависит от r, распределение (7) не имеет моментов порядка выше $1/r - 1$. Расчеты показывают, что все четыре характеристики $R_i(X_1, X_2)$, $i = \overline{1,4}$, с ростом r ведут себя практически одинаково – возрастают по закону, близкому к линейному: $R_i \approx r$ (см. рис. 1, кривая 5). Тем самым для этой конкретной модели подтверждается известное общее положение о том, что тяжелые хвосты способствуют усилению взаимосвязи между случайными величинами.

Распределение Коши – Моргенштерна. Двумерное распределение Моргенштерна с плотностью вероятностей [5, 7]

$$p(x,y) = p_X(x)p_Y(y)\{1 + r[2F_X(x) - 1][2F_Y(y) - 1]\}, \quad r \in [0,1],$$

представляет собой функционал, определенный на множестве произвольных абсолютно непрерывных одномерных функций распределения $F_X(x)$ и $F_Y(y)$. Выбирая в качестве одномерного распределения закон Коши, получаем двумерное распределение Коши – Моргенштерна с плотностью

$$p(x,y) = \frac{\pi^2 + 4r\,\mathrm{arctg}(x)\,\mathrm{arctg}(y)}{\pi^4 (1 + x^2)(1 + y^2)}. \tag{8}$$

Моментов у плотности (8) не существует. Параметр r является параметром связи распределения. В зависимости от него были рассчитаны характеристики взаимосвязи $R_i(X_1, X_2)$, $i = \overline{1,4}$ (1)–(4). На рис. 1 кривой 6 показано поведение всех четырех характеристик, которые в масштабе графика практически сливаются. Максимальное различие между ними достигается при $r = 1$: $R_1(X, Y) = 0{,}336\ldots$, $R_2(X, Y) = 0{,}357\ldots$, $R_3(X, Y) = 0{,}342\ldots$, $R_4(X, Y) = 0{,}333\ldots$. Как следует из рисунка, распределение (8) ни при каких значениях параметров не может описывать функционально или даже сильно связанные случайные величины (ситуация в известном смысле противоположная ситуации с распределением Коши (6)).

Симметричные устойчивые процессы авторегрессии первого порядка. Симметричные устойчивые линейные процессы определяются так же, как и гауссовские, являющиеся их частным случаем: произвольный линейный функционал от такого процесса при фиксированном характеристическом показателе α есть симметричная устойчивая случайная величина $Y(a, b, \alpha)$ с

характеристической функцией $\theta_Y(u)=\exp(iua-b|u|^{\alpha})$, $a\in(-\infty,\infty)$, $b>0$, $\alpha\in(0,2]$. Они образуют полный класс процессов с двухпараметрическими (не считая α) одномерными распределениями, которые обладают подобным свойством. Гауссовские процессы получаются из них как частный случай при $\alpha=2$, линейные процессы Коши – при $\alpha=1$. При $\alpha<2$ случайная величина $Y(a,b,\alpha)$ не имеет моментов выше α. Важным подклассом симметричных устойчивых процессов являются марковские процессы, которые в дискретном времени описываются уравнением авторегрессии первого порядка

$$X_k - rX_{k-1} = Y_k, \quad r\in[-1,1], \tag{9}$$

где $Y_k \stackrel{d}{=} Y(0, 1-|r|^{\alpha}, \alpha)$ – независимые симметричные устойчивые случайные величины с нулевой медианой; r – параметр авторегрессии, характеризующий зависимость членов последовательности (9). Одномерное распределение процесса (9) в стационарном режиме будет устойчивым: $X_k \stackrel{d}{=} Y(0,1,\alpha)$. Распределение случайной величины X_k при фиксированной величине $X_{k-1}=x_{k-1}$ также устойчиво:

$$X_k(X_{k-1}) \stackrel{d}{=} Y(rx_{k-1}, 1-|r|^{\alpha}, \alpha).$$

Отсюда, учитывая марковское свойство, получаем n-мерную плотность вероятностей процесса (9)

$$p(\mathbf{x})=\frac{1}{\pi^n}\int_0^{\infty}\exp(-y^{\alpha})\cos(x_1 y)dy\prod_{i=2}^{n}\int_0^{\infty}\exp[-(1-|r|^{\alpha})y^{\alpha}]\cos[(x_i-rx_{i-1})y]dy. \tag{10}$$

Для исследования взаимосвязи случайных величин X_1, X_2, описываемых уравнением (9), наиболее подходящим является информационный коэффициент корреляции 1-го рода (1), который очень просто выражается через коэффициент авторегрессии r

$$R_1^2(X_1, X_2)=1-(1-|r|^{\alpha})^{2/\alpha}. \tag{11}$$

Безусловная и условная интерквантильные широты порядка $u\in[0,1]$ случайной величины X_2 равны $V_{X_2}(u)=2\lambda_u$, $V_{X_2|X_1}(u)=2\lambda_u(1-|r|^{\alpha})^{1/\alpha}$, где λ_u – корень уравнения $\frac{1}{\pi}\int_0^{\infty}\exp[-(x/\lambda_u)^{\alpha}]\frac{\sin x}{x}dx=|u-0{,}5|$. Отметим, что условная интерквантильная широта $V_{X_2|X_1}(u)$ не зависит от фиксируемого значения $X_1=x_1$. Представление $R_1(X_1, X_2)$ через отношение интерквантильных широт (5) в данном случае не зависит от u и принимает вид

$$R_1^2(X_1, X_2)=1-\frac{V_{X_2|X_1}^2(u)}{V_{X_2}^2(u)}, \quad \forall u.$$

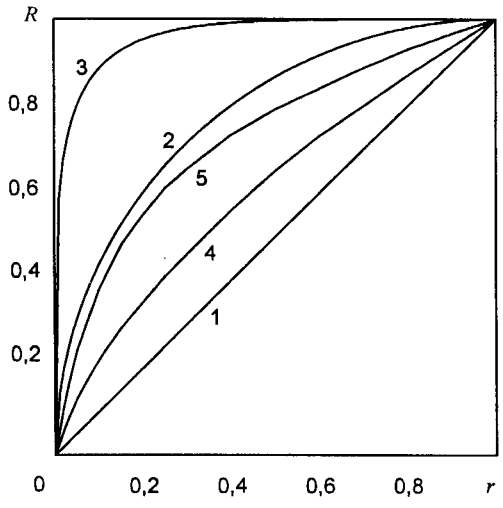

Рис. 2

Таким образом, взаимосвязь случайных величин модели (9) полностью характеризуется отношением параметров масштаба условного и безусловного распределений.

На рис. 2 кривыми 1–3 в зависимости от r изображены информационные коэффициенты корреляции (11) для $\alpha = 2, 1, 1/2$ соответственно. Кривыми 4, 5 обозначены коэффициенты конкорреляции для $\alpha = 1, 1/2$ соответственно. Для гауссовских случайных величин ($\alpha = 2$) коэффициент конкорреляции в масштабе графика практически совпадает с кривой 1. Все характеристики показывают увеличение взаимосвязи при утяжелении хвостов распределения (при уменьшении α). Скорость увеличения максимальна при $r = 0$. Такое поведение характеристик вызвано возрастанием при $\alpha \to 0$ с вероятностной точки зрения роли члена rX_{k-1}, которую он играет в уравнении (9):

$$\mathrm{med}|rX_{k-1}|/\mathrm{med}|Y_k| = r/(1-|r|^{\alpha})^{1/\alpha} \to \infty, \quad r \to 0,\ \alpha \to 0.$$

Это означает, что для распределений с тяжелыми хвостами даже небольшое отличие коэффициента авторегрессии r от нуля приводит к значительной зависимости между членами авторегрессионной последовательности (9).

Заключение. Предложенные в этой работе информационные характеристики взаимосвязи для произвольных случайных величин из абсолютно непрерывных распределений обладают необходимой полнотой свойств для выявления как линейных, так и нелинейных зависимостей и их анализа, что зачастую недостижимо в рамках традиционных методов исследования.

СПИСОК ЛИТЕРАТУРЫ

1. **Моисеев С. Н.** Механизм образования и вероятностное распределение максимальной электронной концентрации слоя E_s // Геомагнетизм и аэрономия. 1997. **37**, № 3. С. 107.
2. **Resnick S. I.** Heavy tail modeling and teletraffic data. N. Y., 1995. (Prepr. /School of ORIE, Cornell University, Ithaca).

3. **Mandelbrot B. B.** The Pareto – Levy law and the distribution of income // Internat. Econom. Rev. 1960. N 1. P. 79.

4. **Боровков А. А.** Математическая статистика. М.: Наука, 1984.

5. **Губарев В. В.** Вероятностные модели. Новосибирск, 1992. Ч. 1, 2.

6. **Губарев В. В.** Характеристики случайных элементов, инвариантные к взаимно однозначным безынерционным функциональным преобразованиям // Автометрия. 1984. № 6. С. 29.

7. **Johnson M. E., Wang C., Ramberg J. S.** Generation of continuous multivariate distributions for statistical applications // Am. Journ. Math. Manag. Sci. 1984. **4**, N 3–4. P. 225.

Воронежский государственный университет,
E-mail: mois@rf.main.vsu.ru

Поступила в редакцию
28 апреля 1999 г.