РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 5 2001

УДК 519.816

С. Н. Моисеев
(Воронеж)

ХАРАКТЕРИСТИКИ ЧАСТНОЙ И МНОЖЕСТВЕННОЙ СВЯЗИ СЛУЧАЙНЫХ ВЕЛИЧИН ИЗ РАСПРЕДЕЛЕНИЙ С ТЯЖЕЛЫМИ ХВОСТАМИ

На основе информационного расстояния Кульбака – Лейблера предложены информационные коэффициенты частной и множественной связи случайных величин из абсолютно непрерывных распределений, в том числе с тяжелыми хвостами. С помощью введенных характеристик проведен анализ структуры взаимосвязи значений симметричного устойчивого процесса авторегрессии p-го порядка.

Введение. В радиофизических приложениях широкое распространение получили вероятностные модели случайных процессов в виде стохастических дифференциальных уравнений. Связано это с простотой и естественностью физической интерпретации, легкостью прогнозирования и моделирования в рамках таких моделей. Способы анализа моделей, заданных линейными стохастическими дифференциальными уравнениями, хорошо разработаны. В частности, удобными инструментами изучения стохастической связи между отсчетами процесса в разнесенные моменты времени являются коэффициенты частной и множественной корреляций.

Коэффициент множественной корреляции описывает линейную связь случайной величины X_0 с совокупностью случайных величин $\mathbf{X}_p = \| X_{k+1}, X_{k+2}, \ldots, X_{k+p} \|$, $k \geq 0$, и определяется следующим образом [1]:

$$r^2(X_0, \mathbf{X}_p) = 1 - \frac{\det \mathbf{K}}{|\mathbf{K}|_{00}}, \qquad (1)$$

где $\mathbf{K} = \| r_{ij} \|$, $i, j = 0, k+1, k+2, \ldots, k+p$, – нормированная корреляционная матрица случайных величин $X_0, X_{k+1}, X_{k+2}, \ldots, X_{k+p}$; $|\mathbf{K}|_{ij}$ – алгебраическое дополнение элемента r_{ij} в определителе матрицы $\mathbf{K}$. Он максимизирует по множеству весов обычный коэффициент корреляции между X_0 и линейно взвешенной суммой $X_{k+1}, \ldots, X_{k+p}$. Эта характеристика обычно используется для расчета максимального интервала линейной зависимости между значениями случайного процесса.

Коэффициент частной корреляции описывает линейную парную связь случайных величин X_0 и X_{k+1} при фиксированных $\mathbf{X}_k = \|X_1, X_2, ..., X_k\|$ и определяется следующим образом [1]:

$$r(X_0, X_{k+1} \mid \mathbf{X}_k) = -\frac{|\mathbf{K}|_{0k+1}}{\sqrt{|\mathbf{K}|_{00}\,|\mathbf{K}|_{k+1k+1}}}. \qquad (2)$$

С его помощью обычно оценивают неизвестный порядок линейного стохастического дифференциального уравнения, которому подчиняются отсчеты процесса $X_0, ..., X_{k+1}$.

В случае нелинейных стохастических дифференциальных уравнений для этих же целей применяют частные и множественные корреляционные отношения [2, 3], которые кроме линейных связей способны выделять и некоторые виды нелинейных. Они широко используются в метеорологии при построении нелинейных динамических моделей и прогнозировании [4].

Однако, все перечисленные здесь характеристики существуют не всегда, так как требуют наличия конечных вторых моментов у исследуемого процесса. Так, максимальная электронная концентрация спорадического слоя E ионосферы описывается нелинейным стохастическим дифференциальным уравнением с порождающим процессом из вероятностного распределения Коши [5], у которого, как известно, не существует моментов. Подобные модели, вероятностная структура которых описывается распределениями с тяжелыми хвостами, встречаются при анализе телекоммуникационных трафиков, эконометрических данных и т. д. [6, 7].

В данной работе на основе сформулированных требований введем характеристики частной и множественной связи между случайными величинами, в том числе из распределений с тяжелыми хвостами, способные выявлять не только линейные, но и нелинейные связи. С помощью введенных характеристик проведем анализ одной вероятностной модели, имеющей важное теоретическое и практическое значение.

Характеристики частной множественной связи. Рассмотрим характеристику частной множественной связи $R(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$ между случайной величиной X_0 и вектором $\mathbf{X}_p$ при фиксированном векторе $\mathbf{X}_k$, включающую в себя как частные случаи характеристики множественной $R(X_0, \mathbf{X}_p)$ при $k = 0$ и частной парной $R(X_0, X_{k+1} \mid \mathbf{X}_k)$ при $p = 1$ связей. Сформулируем основные требования к свойствам характеристики $R(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$ для произвольных случайных величин $X_0, X_1, ..., X_{k+p}$ из абсолютно непрерывных распределений.

1. $0 \le R(X_0, \mathbf{X}_p \mid \mathbf{X}_k) \le 1$.
2. $R(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = R(\mathbf{X}_p, X_0 \mid \mathbf{X}_k)$.
3. $R(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = 0$, если X_0 и $\mathbf{X}_p$ независимы при фиксированных $\mathbf{X}_k$.
4. Если $R(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = 0$, то X_0 и $\mathbf{X}_p$ независимы при фиксированных $\mathbf{X}_k$.
5. Для функциональной связи $X_0 = f(\mathbf{X}_p)$ при фиксированных $\mathbf{X}_k$, где $f(\mathbf{x})$ – произвольная неслучайная функция p переменных, $R(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = 1$.
6. Для случайных величин $Y_i = f_i(X_i)$, $i = 0, k+p$, где $f_i(x)$ – взаимно однозначные функции, $R(Y_0, \mathbf{Y}_p \mid \mathbf{Y}_k) = R(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$.
7. Для совместно гауссовских случайных величин $X_0, ..., X_{k+p}$

$$R(X_0, \mathbf{X}_p) = r(X_0, \mathbf{X}_p), \qquad R(X_0, X_{k+1} \mid \mathbf{X}_k) = |r(X_0, X_{k+1} \mid \mathbf{X}_k)|.$$

Перечисленные требования являются достаточно жесткими, поэтому для синтеза характеристик требуется весьма мощный аналитический аппарат. Такой аппарат предоставляет информационное расстояние Кульбака – Лейблера, которое лежит в основе многих оптимальных процедур математической статистики [8] 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:

$$\rho[p(\mathbf{x}), q(\mathbf{x})] = \mathbf{M}[\ln p(\mathbf{x}) - \ln q(\mathbf{x})] = \int_{S_p} p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} \mu(d\mathbf{x}), \tag{3}$$

где $\mathbf{M}[\cdot]$ – оператор математического ожидания плотности $p(\mathbf{x})$; $S_p = \{\mathbf{x}: p(\mathbf{x}) > 0\}$ – носитель распределения $p(\mathbf{x})$; μ – мера Лебега. $\rho[p(\mathbf{x}), q(\mathbf{x})] \geq 0$, причем равенство достигается тогда и только тогда, когда $p(\mathbf{x}) \equiv q(\mathbf{x})$. Полагая в (3) $p(\mathbf{x}) = p(x_0, \mathbf{x}_k, \mathbf{x}_p) p(\mathbf{x}_k) = p(x_0, \mathbf{x}_p \mid \mathbf{x}_k) p^2(\mathbf{x}_k)$, $q(\mathbf{x}) = p(x_0, \mathbf{x}_k) p(\mathbf{x}_k, \mathbf{x}_p) = p(x_0 \mid \mathbf{x}_k) p(\mathbf{x}_p \mid \mathbf{x}_k) p^2(\mathbf{x}_k)$ (или, наоборот, $p(\mathbf{x}) = p(x_0 \mid \mathbf{x}_k) p(\mathbf{x}_p \mid \mathbf{x}_k) p^2(\mathbf{x}_k)$, $q(\mathbf{x}) = p(x_0, \mathbf{x}_p \mid \mathbf{x}_k) p^2(\mathbf{x}_k)$), где $p(\mathbf{x})$ – безусловные, а $p(\mathbf{x} \mid \mathbf{y})$ – условные (при фиксированном векторе $\mathbf{y}$) маргинальные плотности вероятностей случайного вектора $\mathbf{X}$, видим, что расстояние $\rho[\cdot]$ естественным образом характеризует связь случайной величины X_0 с величинами $\mathbf{X}_p$ при фиксированных $\mathbf{X}_k$. На основе расстояния Кульбака – Лейблера (3) можно построить много характеристик частной множественной связи, удовлетворяющих требованию 1, из которых были отобраны следующие наиболее простые характеристики:

$$R_1^2(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = 1 - \exp(-2\rho_1), \tag{4}$$

$$R_2^2(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = \frac{2\rho_2}{1 + 2\rho_2}, \tag{5}$$

$$R_3^2(X_0, \mathbf{X}_p \mid \mathbf{X}_k) = \frac{\rho_1 + \rho_2}{1 + \rho_1 + \rho_2}, \tag{6}$$

где

$$\rho_1 = \rho[p(x_0, \mathbf{x}_k, \mathbf{x}_p) p(\mathbf{x}_k), p(x_0, \mathbf{x}_k) p(\mathbf{x}_k, \mathbf{x}_p)] =$$

$$= \int_S \rho[p(x_0, \mathbf{x}_p \mid \mathbf{x}_k), p(x_0 \mid \mathbf{x}_k) p(\mathbf{x}_p \mid \mathbf{x}_k)] p(\mathbf{x}_k) \mu(d\mathbf{x}_k),$$

$$\rho_2 = \rho[p(x_0, \mathbf{x}_k) p(\mathbf{x}_k, \mathbf{x}_p), p(x_0, \mathbf{x}_k, \mathbf{x}_p) p(\mathbf{x}_k)] =$$

$$= \int_S \rho[p(x_0 \mid \mathbf{x}_k) p(\mathbf{x}_p \mid \mathbf{x}_k), p(x_0, \mathbf{x}_p \mid \mathbf{x}_k)] p(\mathbf{x}_k) \mu(d\mathbf{x}_k),$$

$S = \{\mathbf{x}_k : p(\mathbf{x}_k) > 0\}$. В терминах теории информации расстояние Кульбака – Лейблера ρ_1 совпадает со средней информацией, содержащейся в векторе $\mathbf{X}_p$ относительно X_0 при фиксированных $\mathbf{X}_k$; ρ_2 – с отрицательной информа-

цией, содержащейся в векторе $\mathbf{X}_p$ относительно X_0 при фиксированных $\mathbf{X}_k$ и усредненной в предположении, что она равна нулю.

Характеристики (4)–(6) удовлетворяют первым шести перечисленным выше требованиям. Доказательства свойств 1–4 достаточно очевидны и здесь не приводятся. При функциональной связи $X_0 = f(\mathbf{X}_p)$ совместная плотность вероятностей случайных величин X_0 и $\mathbf{X}_p$ имеет вид $p(x_0, \mathbf{x}_p) = \delta(x_0 - f(\mathbf{x}_p))p(\mathbf{x}_p)$, где $\delta(x)$ – дельта-функция. Подставляя $p(x_0, \mathbf{x}_p)$ в формулы (3)–(6), получаем доказательство пятого свойства для характеристик (4)–(6). Шестое свойство легко доказывается простой заменой переменных $Y_i = f_i(X_i)$, $i = \overline{0, k+p}$, в интеграле (3). Седьмому свойству удовлетворяют в полной мере только $R_1(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$ и $R_3(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$. Для характеристики связи $R_2(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$ седьмое свойство выполняется асимптотически:

$$R_2(X_0, \mathbf{X}_p) \to r(X_0, \mathbf{X}_p), \qquad R_2(X_0, X_{k+1} \mid \mathbf{X}_k) \to |r(X_0, X_{k+1} \mid \mathbf{X}_k)|$$

при $r_{ij} \to 0$, $i \neq j$. При $k = 0$, $p = 1$ максимальное относительное отклонение $R_2(X_0, X_1)$ от $|r_{01}|$ достигается в точке $|r_{01}| = 0{,}700\ldots$ и не превышает 6,5 %. При $k = 0$, $p = 1$ характеристика парной связи $R_1(X_0, X_1)$ уже встречалась в литературе под названием информационного коэффициента корреляции [9]. По аналогии назовем характеристики $R_1(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$, $R_2(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$, $R_3(X_0, \mathbf{X}_p \mid \mathbf{X}_k)$ информационными коэффициентами частной множественной корреляции 1–3-го рода соответственно.

Анализ зависимостей симметричного устойчивого процесса авторегрессии *p*-го порядка. Рассмотрим теперь применение введенных характеристик частной множественной связи для анализа структуры зависимостей одной важной как с теоретической, так и с практической точки зрения вероятностной модели. Симметричные устойчивые линейные процессы являются естественным обобщением гауссовских процессов, являющихся их частным случаем, и определяются аналогично: произвольный линейный функционал от такого процесса при фиксированном характеристическом показателе α есть симметричная устойчивая случайная величина $Y(a, b, \alpha)$ с характеристической функцией $\theta_Y(u) = \exp(iua - b|u|^\alpha)$, $b > 0$, $\alpha \in (0, 2]$. 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 процессы, обладающие свойством марковости p-го порядка, которые в дискретном времени описываются уравнением авторегрессии p-го порядка:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = a_t, \tag{7}$$

где $a_t \overset{d}{=} Y(0, b, \alpha)$ – независимые симметричные устойчивые случайные величины с медианой $a = 0$; ϕ_i, $i = \overline{1, p}$, – параметры авторегрессии. Необходимым условием стационарности процесса (7) является следующее ограничение на авторегрессионные коэффициенты: корни уравнения $1 - \phi_1 x - \ldots - \phi_p x^p = 0$ должны лежать вне единичного круга на комплексной плоскости.

Процесс X_t представим в виде суммы независимых случайных величин: $X_t = \sum_{i=0}^{\infty} \psi_0(i) a_{t-i}$. Отсюда, используя аппарат характеристических функций, в стационарном режиме, когда переходные процессы линейной системы, формирующей процесс X_t, уже не сказываются, получаем одномерное распределение процесса (7) симметричным устойчивым:

$$X_{t+l} \stackrel{d}{=} Y\left(0, b\sum_{i=0}^{\infty}|\psi_0(i)|^{\alpha}, \alpha\right), \quad l = 1,2,\ldots,$$

где

$$\psi_k(i) = \begin{cases} 0 & \text{при } i<1,\ i \neq k, \\ 1 & \text{при } i = -k, \\ \sum_{j=1}^{p} \phi_j \psi_k(i-j) & \text{при } i \geq 1. \end{cases}$$

Распределение случайной величины X_{t+l} при фиксированном векторе $\mathbf{X}_p = \|X_t, X_{t-1}, \ldots, X_{t-p+1}\|$ также будет симметричным устойчивым:

$$X_{t+l}(\mathbf{X}_p) \stackrel{d}{=} Y\left(\sum_{i=0}^{p-1}\psi_i(l)X_{t-i},\ b\sum_{i=1}^{l}|\psi_0(l-i)|^{\alpha},\ \alpha\right).$$

Марковское свойство p-го порядка процесса (7) заключается в том, что для любых t и l справедливо $p(x_{t+l} \mid x_t, \ldots, x_{t-m+1}) = p(x_{t+l} \mid \mathbf{x}_p)$, $m \geq p$.

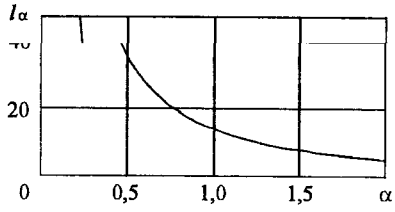
Характеризовать связь между значением процесса X_{t+l} и значениями $\mathbf{X}_p$ будем информационным коэффициентом множественной корреляции $R_1(X_{t+l}, \mathbf{X}_p)$, так как он с учетом перечисленных выше свойств наиболее просто вычисляется для модели (7) по формуле (4):

$$R_1^2(X_{t+l}, \mathbf{X}_p) = 1 - \left(\frac{\sum_{i=0}^{l-1}|\psi_0(i)|^{\alpha}}{\sum_{i=0}^{\infty}|\psi_0(i)|^{\alpha}}\right)^{2/\alpha}.$$

В силу свойства марковости p-го порядка $R_1(X_{t+l}, \mathbf{X}_m) = R_1(X_{t+l}, \mathbf{X}_p)$ при $m \geq p$; $\mathbf{X}_m = \|X_t, X_{t-1}, \ldots, X_{t-m+1}\|$. Поэтому с помощью $R_1(X_{t+l}, \mathbf{X}_p)$ легко рассчитать интервал детерминации (зависимости) процесса X_t аналогично определению интервала корреляции

$$l_{\alpha} = \sum_{l=1}^{\infty} R_1(X_{t+l}, \mathbf{X}_p).$$

Если рассматривать t как настоящий момент времени, а $t+l$ – как момент времени, на который дается прогноз процесса X_t, то l_{α} показывает максималь-



ное упреждение, при котором еще возможен прогноз, значимо отличающийся от тривиального прогноза в виде медианы процесса X_t.

На рисунке изображено поведение интервала детерминации l_α в зависимости от α при следующих заданных значениях параметров уравнения (7): $p=2, \phi_1 = 0{,}6, \phi_2 = 0{,}2$. Эти параметры характерны для флуктуаций максимальной электронной концентрации спорадического слоя E ионосферы, которые описываются моделью (7). Из рисунка хорошо видно, что утяжеление хвостов распределения, которое происходит при уменьшении α, приводит к росту интервала детерминации процесса X_t. Это полностью согласуется с известным положением о том, что тяжелые хвосты распределений способствуют усилению взаимосвязи.

Рассчитаем для модели (7) информационный коэффициент частной корреляции 1-го рода $R_1(X_t, X_{t-k-1} \mid \mathbf{X}_k)$ между значениями процесса X_t и X_{t-k-1} при фиксированных промежуточных отсчетах $\mathbf{X}_k = \| X_{t-1}, X_{t-2}, \ldots, X_{t-k} \|$. Учитывая, что условное распределение случайной величины X_t при фиксированных $\mathbf{X}_k$, $k \le p$, устойчиво:

$$X_t(\mathbf{X}_k) \stackrel{d}{=} Y\left(\sum_{i=1}^{k} \phi_i X_{t+i-k-1}, b\left[1 + \sum_{i=0}^{\infty} \left| \sum_{j=k+1}^{p} \phi_j \psi_0(i+k+1-j) \right|^{\alpha} \right], \alpha \right),$$

получаем по формуле (4)

$$R_1^2(X_t, X_{t-k-1} \mid \mathbf{X}_k) = \begin{cases} 1 - \left(\dfrac{1 + \sum\limits_{i=0}^{\infty} \left| \sum\limits_{j=k+2}^{p} \phi_j \psi_0(i+k+2-j) \right|^{\alpha}}{1 + \sum\limits_{i=0}^{\infty} \left| \sum\limits_{j=k+1}^{p} \phi_j \psi_0(i+k+1-j) \right|^{\alpha}} \right)^{2/\alpha}, & 0 \le k < p, \\ 0, & k \ge p. \end{cases} \quad (8)$$

Пороговое в зависимости от k поведение информационного коэффициента частной корреляции (8), являющееся следствием марковости p-го порядка, позволяет использовать его подобно частному коэффициенту корреляции (2)

в гауссовских моделях для оценки неизвестного порядка p авторегрессионного уравнения (7). В качестве оценки p берется такое минимальное k, что для всех $k' \geq k$ выполняется $R_1(X_t, X_{t-k'-1} \mid \mathbf{X}_{k'}) = 0$, или иначе $p-1 = \max\{k: R_1(X_t, X_{t-k-1} \mid \mathbf{X}_k) > 0\}$.

Заключение. Рассмотренные в этой работе информационные характеристики частной и множественной связи случайных величин из абсолютно непрерывных распределений обладают необходимой полнотой свойств для проведения глубокого анализа структуры как линейных, так и нелинейных зависимостей широкого класса процессов, в том числе имеющих распределения с тяжелыми хвостами.

СПИСОК ЛИТЕРАТУРЫ

1. **Айвазян С. А., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985.
2. **Ивченко Г. И., Медведев Ю. И.** Математическая статистика. М.: Высш. шк., 1984.
3. **Моисеев С. Н.** Прогноз нелинейных временных рядов через взвешенную сумму одномерных регрессий // Радиотехника и электроника. 1999. **44**, № 6. С. 715.
4. **Кравцов Ю. И.** Случайность, детерминированность, предсказуемость // УФН. 1989. **158**, вып. 1. С. 93.
5. **Моисеев С. Н.** Механизм образования и вероятностное распределение максимальной электронной концентрации слоя E_s // Геомагнетизм и аэрономия. 1997. **37**, № 3. С. 107.
6. **Resnick S. I.** Heavy tail modeling and teletraffic data. N. Y., 1995. (Prepr. /School of ORIE, Cornell University, Ithaca).
7. **Mandelbrot B. B.** The Pareto – Levy law and the distribution of income // Internat. Econom. Rev. 1960. N 1. P. 79.
8. **Боровков А. А.** Математическая статистика. М.: Наука, 1984.
9. **Губарев В. В.** Вероятностные модели. Новосибирск: НГУ, 1992.

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

Поступила в редакцию
5 марта 2000 г.