Моделирование копулы зависимости длин интервалов между последовательными эпизодами поведения индивида в гамма-пуассоновской модели

В. Ф. Столярова
Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук
valerie.stoliarova@gmail.com

***Аннотация*. Поведение индивида играет важную роль в ряде задач эпидемиологии, охраны общественного здоровья, безопасности социотехнических систем. В подобных системах в качестве оценки степени влияния поведения на тот или иной аспект работы системы может выступать интенсивность. Наиболее доступными и экономически обоснованными являются анкетные данные, в том числе данные о последних эпизодах. Для использования подобных данных для оценки интенсивности поведения была предложена гамма-пуассоновская модель поведения. В статье на модельных данных построена копула зависимости длин интервалов между последовательными эпизодами поведения индивида. Значения параметров копулы могут использоваться для построения оценки интенсивности поведения индивида.**

Ключевые слова: последние эпизоды; копула; неточные данные

I. Введение

Задача оценки риска в сферах охраны общественного здоровья, эпидемиологии, кибербезопасности включает в себя численную оценку параметров поведения индивида, которое связано с риском [1]. Человек и его поступки играют важную роль в этих областях. Если поведение индивида представляет собой последовательность эпизодов, то основным параметром, характеризующим его, является *интенсивность*. Существуют две группы методов измерения интенсивности поведения: прямые и непрямые [1]. Прямое измерение интенсивности предполагает регистрацию каждого эпизода тем или иным способом и является дорогостоящим и затратным по времени. Непрямые оценки опираются на информацию, полученную в результате интервью. Получение такой информации не требует значительных денежных и временных затрат, но данные, полученные с использованием интервью, могут быть неточными [18].

Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2018-0001 и гранта РФФИ №18-01-00626

A. *Поведение, связанное с риском*

К примеру, пищевое поведение беременных женщин может оказывать влияние на развитие плода [9, 11]. При наблюдении беременности специалисту может требоваться *быстрая* оценка частоты употребления тех или иных продуктов питания с тем, чтобы назначить дополнительные обследования или дать диетические рекомендации. Важной особенностью в этом случае является сложность припоминания обыденных действий [19].

Другими примерами поведения, связанного с риском, являются употребление алкоголя и иных наркотических веществ (в том числе внутривенное употребление наркотических веществ), незащищенный секс со случайным партнером [1]. Такое поведение само по себе является объектом исследования в областях эпидемиологии и охраны общественного здоровья, так как связано с повышенными рисками смертности и нетрудоспособности и приводит к распространению неизлечимых заболеваний. Особенностью является стремление индивида дать социально-ожидаемый ответ на вопросы об этом поведении [13].

B. *Последние эпизоды для оценки интенсивности*

В качестве альтернативы имеющимся подходам к непрямой оценке интенсивности поведения, были предложены вопросы о последних эпизодах [1]. Этот метод состоит из двух частей: опросника о последних эпизодах поведения индивида и математической модели поведения.

В рамках этого подхода респонденту задаются вопросы о последних эпизодах его поведения, к примеру, «Когда в последний раз Вы употребляли алкоголь?», а также рекордных интервалов между эпизодами. Вопросы такого типа позволяют минимизировать смещение припоминания (recall bias) при регистрации ответов. Кроме того, существенным отличием от других непрямых методов измерения интенсивности поведения является то, что ответы на такие вопросы являются количественной

характеристикой поведения. Их особенностью является то, что такие ответы являются неточными, так как даются на естественном языке (лингвистические переменные).

Для построения оценки интенсивности поведения по данным о последних эпизодах используется математическая модель поведения: пуассоновский (или смешанный пуассоновский) процесс [7]. В практических приложениях используются байесовские сети доверия [5] для построения оценки и принятия решений, однако аналитические способы для гамма-пуассоновской модели так и не были разработаны [3].

Целью данной статьи является построение копулы для двух интервалов между эпизодами поведения в гамма-пуассоновской модели. Оценка параметров построенной копулы может быть использована для оценки параметров распределения интенсивности в популяции. Вычисления сопровождаются численным примером на модельных данных.

II. Предположения и математический аппарат

A. *Пуассоновская и гамма–пуассоновская модель поведения*

Пусть поведение индивида представляет собой последовательность эпизодов на временной оси, и пусть эта последовательность эпизодов представляет собой однородный процесс Пуассона. В этом случае интенсивность поведения представляется интенсивностью процесса λ. По свойствам пуассоновского процесса [6], интервалы между последовательными эпизодами процесса независимы и одинаково экспоненциально распределены с параметром λ. По вопросам о последних эпизодах поведения исследователь получает сверхкороткую выборку таких интервалов, которая может служить для построения искомой оценки [3]. В случае пуассоновской модели предполагается, что у каждого индивида в генеральной совокупности одинаковая интенсивность поведения.

Гамма–пуассоновская модель поведения [18] возникает в предположении, что каждый индивид обладает свойственной ему интенсивностью поведения. При отсутствии данных о виде распределения λ, в качестве априорного предположения предлагается выбрать двухпараметрическое гамма-распределение с параметрами [6]: формы $a > 0$ и масштаба $s > 0$ и плотностью вероятности

$$f(x) = \frac{1}{s^a \Gamma(a)} x^{a-1} e^{-(x/s)} \quad (1)$$

Таким образом, в случае гамма-пуассоновской модели по сверхкороткой выборке, сформированной в результате опроса, необходимо оценить параметры гамма–распределения вероятности

В рамках данной статьи рассматривается случай двух интервалов между последовательным эпизодами поведения τ_1 и τ_2. Согласно гамма-пуассоновской модели поведения, τ_1 и τ_2 имеют экспоненциальное распределение вероятности $F(x) = P[\tau_1 < x] = 1 - e^{-\lambda x}$ с параметром $\lambda > 0$, причем λ имеет гамма-распределение вероятности.

B. *Копула*

n-копула [2, 12] представляет собой ограниченное на единичный куб совместное распределение вероятности n равномерно распределенных на отрезке [0;1] случайных величин.

Теорема Склара [2, 12] устанавливает факт существования для n случайных величин $X_1, X_2, \ldots, X_n$ с функциями распределения $F_1(x), F_2(x), \ldots, F_n(x)$ и совместной функцией распределения H такой копулы C, что

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)) \quad (2)$$

В случае непрерывных маргинальных распределений, такая функция C единственна. Верно и обратное утверждение: если C – копула и $F_1(x), F_2(x), \ldots, F_n(x)$ – функции распределения, то H – совместная функция распределения случайных величин $X_1, X_2, \ldots, X_n$.

Копулы широко используются для моделирования сложных зависимостей в различных приложениях [8, 12]. Ниже перечислены особенности копул.

- Копулы позволяют разделить задачи оценки параметров зависимости и маргинальных распределений [8].
- Копулы служат для того, чтобы генерировать новые классы распределений вероятности [8].
- Копула является исчерпывающим описанием формы зависимости переменных, причем при монотонном преобразовании случайных величин, копула не изменяется вовсе (изменяется предсказуемым образом) [12].
- Архимедовы копулы могут объединяться попарно, что используется для многомерного моделирования. В общем случае, задача построения многомерных копул достаточно трудоемкая [8].
- Аппарат копул дал начало новой вероятностной графической модели с непрерывными переменными в узлах – модели лозы [10].

III. Копула зависимости длин двух интервалов между последовательными эпизодами в гамма-пуассоновской модели

A. *Теоретический вывод*

Итак, обозначим

$$\hat{C}_{12}(u, v) = P[F(\tau_1) > u, F(\tau_2) > v] \quad (3)$$

копулу, описывающую взаимосвязь между τ_1 и τ_2. Следуя выводу в [4] получаем:

$$\bar{H}(x,y) = P(\tau_1 > x, \tau_2 > y) = \frac{1}{(sx+sy+1)^a} \ .$$

Маргинальные распределения совместного распределения $\bar{H}$:

$$\bar{G}_1(x) = P(\tau_1 > x) = \bar{H}(x,0) = (1+sx)^{-a} \ .$$

Квазиобратная функция к этому распределению вероятности:

$$\bar{G}_1^{(-1)}(u) = \frac{u^{-\frac{1}{a}} - 1}{s} \ .$$

Чтобы получить вид копулы $\hat{C}_{12}$ в (3), вычислим

$$\hat{C}_{12}(u,v) = \bar{H}(\bar{G}_1^{-1}(u), \bar{G}_2^{-1}(v)) = (u^{-\frac{1}{a}} + v^{-\frac{1}{a}} - 1)^{-a} \qquad (4)$$

Полученная копула является копулой Клейтона (survival Clayton copula). Переменные, связные копулой Клейтона, обладают свойством зависимости в первом квадранте (Positive Quadrant Dependence) [8, 12], что означает, что выполнено:

$$P(\tau_1 > x, \tau_2 > y) > P(\tau_1 > x)P(\tau_2 > y) \ .$$

Это свойство говорит о том, что в гамма–пуассоновской модели поведения интервалы *кластеризованы*, то есть длинные (короткие) интервалы имеют тенденцию реализовываться вместе.

Обратим внимание, что параметр копулы соответствует параметру формы распределения вероятности интенсивности (1), а параметр масштаба не участвует в формуле копулы, так как вид копулы не изменяется при монотонном преобразовании [12].

Таким образом, оценка задача оценки параметров распределения интенсивности поведения может быть представлена как задача оценки параметра копулы в случае, если наблюдается два интервала между последовательными эпизодами.

B. Моделирование копулы

Пусть имеется n респондентов, каждый из которых имеет свою собственную интенсивность некоторого поведения. Так же предположим, что в целом в популяции интенсивность поведения имеет гамма-распределение вероятности (1). Для иллюстрации, была сгенерирована выборка объема $n = 10000$ из гамма-распределения с параметрами $a = 1.1$ и $s = 1$. Затем для каждого значения λ^b, $b = 1{:}10000$ были сгенерированы значения длин двух интервалов τ_1^b, τ_2^b, которые следуют экспоненциальному распределению вероятности с параметром λ^b.

Затем наблюдения τ_1^b, τ_2^b были приведены к равномерному распределению вероятности при помощи

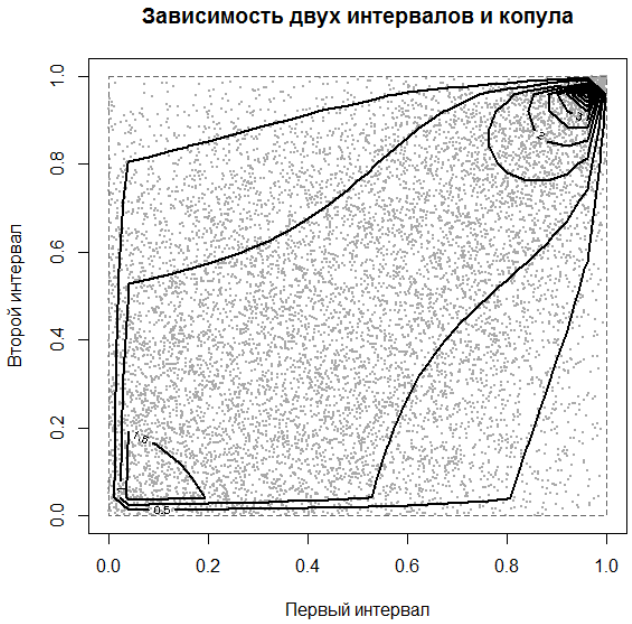

Рис. 1. Псевдонаблюдения и контурная диаграмма копулы, которая их связывает

преобразования $u_{ij} = \frac{r_{ij}}{(n+1)}$, $i, j = 1 \ldots n$, r_{ij} представляет собой ранг наблюдения в ряду всех $x_{kj}, k = 1 \ldots n$

На последнем этапе при помощи метода максимума правдоподобия по псевдонаблюдениям был оценен параметр копулы: $a = 1.23$ (95% доверительный интервал $(1.18; 1.28)$). На рис. 1 представлена диаграмма рассеяния псевдонаблюдений и контуры построенной копулы [12].

IV. Заключение

Итак, в рамках статьи задача оценки параметров распределения интенсивности поведения по данным о длинах двух интервалов между эпизодами поведения была сформулирована в терминах копул. Предположения, налагаемые гамма-пуассоновской моделью поведения позволяют определить вид копулы, характеризующей зависимость интервалов – копула Клейтона. Задача оценки параметров гамма-распределения становится задачей оценки параметра копулы.

Предложенный способ является новым. Применение было проиллюстрировано на модельных данных.

Список литературы

[1] Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С., Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей, 2008, 140 с.

[2] Благовещенский Ю.Н. Основные элементы теории копул // Прикладная эконометрика, № 2(26). 2012. Стр.113–130.

[3] Степанов Д.В., Мусина В.Ф., Суворова А.В., Тулупьев А.Л., Сироткин А.В., Тулупьева Т.В. Функция правдоподобия с гетерогенными аргументами в идентификации пуассоновской модели рискованного поведения в случае информационного дефицита // Труды СПИИРАН, 4(23). 2012. С. 157–184.

[4] Столярова В.Ф. Виды и свойства попарной зависимости длин интервалов между последовательными эпизодами в пуассоновской и гамма–пуассоновской моделях поведения // IV Международная летняя школа–семинар по искусственному интеллекту для студентов, аспирантов, молодых ученых и специалистов интеллектуальные системы и технологии: современное состояние и перспективы–2017 (ISYT–2017) г. Санкт-Петербург, 30 июня – 3 июля, 2017 г. Сборник научных трудов. С. 160-167.

[5] Суворова А.В., Тулупьев А.Л., Сироткин А.В. Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления, Т.9, № 2. 2014. С. 115–129.

[6] Феллер В. Введение в теорию вероятностей и её приложения. Том 1. М.: Мир, 1984. 528 с.

[7] Daley D.J., Vere–Jones D. An introduction to the theory of point processes: volume I.Springer Science & Business Media, 2007.

[8] Balakrishnan N., Lai C. D. Continuous bivariate distributions. Springer Science & Business Media, 2009. 684 p.

[9] Kuczkowski K.M. Caffeine in pregnancy // Archives of gynecology and obstetrics, 280(5), 2009, pp.695--698

[10] Kurowicka D., Joe H. (eds.) Dependence modeling: vine copula handbook. World Scientific Publishing Co, 2011. 370 p.

[11] Mulligan M.L., Felton S.K., Riek A.E., & Bernal-Mizrachi, C., Implications of vitamin D deficiency in pregnancy and lactation // American Journal of Obstetrics & Gynecology, 202(5), 2010, 429-e1.

[12] Nelsen R.B. An introduction to copulas. Springer Science & Business Media, 2007. 270 p.

[13] Davis C.G., Thake J., & Vilhena N., Social desirability biases in self-reported alcohol consumption and harms // Addictive behaviors, 35(4), 2010. 302–311.

[14] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[15] Marius Hofert, Ivan Kojadinovic, Martin Maechler, Jun Yan. copula: Multivariate Dependence with Copulas. R package version 0.999–17 URL: https://CRAN.R-project.org/package=copula

[16] Jun Yan. Enjoy the Joy of Copulas: With a Package copula // Journal of Statistical Software, 21(4), 2007. С. 1–21. URL: http://www.jstatsoft.org/v21/i04/

[17] Ivan Kojadinovic, Jun Yan. Modeling Multivariate Distributions with Continuous Margins Using the copula R Package // Journal of Statistical Software, 34(9), 2010. С. 1–20. URL: http://www.jstatsoft.org/v34/i09/

[18] Зельтерман Д., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Гро Л.Е., Хаймер Р. Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пауссоновской модели поведения // Труды СПИИРАН. 2011. Вып. 16. С. 160–185.

[19] Shim J.S., Oh K., Kim H.C. Dietary assessment methods in epidemiologic studies // Epidemiology and health. 2014. Т. 36. e2014009