УДК 004.89 **DOI** 10.15622/ia.23.6.6

В.Ф. Столярова , Т.В. Тулупьева , А.А. Вяткин

ПОДХОДЫ К ОЦЕНИВАНИЮ КУМУЛЯТИВНЫХ ХАРАКТЕРИСТИК ПОВЕДЕНИЯ В ГРУППАХ РАЗНОРОДНЫХ ИНДИВИДОВ: ТОЧНОСТЬ И ПРИМЕНИМОСТЬ В УСЛОВИЯХ ОГРАНИЧЕННЫХ НАБЛЮДЕНИЙ

Столярова В.Ф., Тулупьева Т.В., Вяткин А.А. **Подходы к оцениванию кумулятивных характеристик поведения в группах разнородных индивидов: точность и применимость в условиях ограниченных наблюдений.**

Аннотация. В ряде социоориентированных областей знаний возникает задача оценки кумулятивных характеристик поведения индивидов, таких как частота, которые реализуются в группах индивидов, причем поступающие данные сопряжены с неопределенностью. Нередки ситуации ограниченных данных, когда для небольшого числа наблюдаемых объектов известны лишь несколько эпизодов. Существуют несколько подходов, позволяющих строить оценки искомой кумулятивной характеристики в условиях ограниченных ресурсов: классический подход регрессии Кокса, оценка параметра копулы, апостериорный вывод в байесовских сетях доверия, классических и гибридных, однако до сих пор не были проанализированы возможности применимости имеющихся методов. Целью работы является анализ особенностей применения существующих методов косвенного оценивания интенсивности рискованного поведения индивидов на основе ограниченных данных об эпизодах для определения рекомендаций по их применению: определение точности оценок, получаемых с помощью перечисленных подходов, на основе расстояния Канторовича-Рубинштейна от истинного распределения искомой частоты, а также выявление требований к данным, которые предъявляются для построения оценок. Было показано, что подход на основе копул дает самые точные оценки и обладает наименьшими требованиями к количеству наблюдаемых объектов, однако не может учитывать внешние факторы, которые могут оказывать влияние на реализацию эпизодов поведения. Среди моделей, позволяющих учитывать ковариванты процесса, наибольшей точностью обладают оценки, опирающиеся на апостериорный вывод в гибридных байесовских сетях доверия. Полученные результаты являются новыми, они нацелены на применение в автоматизированных системах обработки информации о поведении индивидов. Практической значимостью обладают рекомендации по применению имеющихся подходов в зависимости от имеющихся данных.

Ключевые слова: последние эпизоды, рискованное поведение, байесовские сети доверия, гибридные байесовские сети доверия, копула.

1. Введение. В ряде областей знаний возникают задачи анализа последовательностей эпизодов некоторого поведения, которые реализуются в группах различающихся индивидов. Основным источником таких задач выступают социоориентированные области знаний, где поведение может быть сопряжено с риском. Для оценки показателей такого риска требуется выявлять характеристики, отражающие последовательности риск-ассоциированных эпизодов. Эти характеристики также позволяют сравнивать группы индивидов

между собой в плане поведения или с некоторым установленным значением. Интенсивность является одной из таких характеристик и широко используется при анализе риска в сфере эпидемиологии при мониторинге неизлечимых заболеваний [1, 2] и угрозообразующего поведения [3, 4], а также при анализе затрат в системе здравоохранения [5]. В области кибербезопасности возникают схожие задачи, связаные с анализом цифровых следов пользователей онлайн социальных сетей, установлением взаимосвязи между активностью индивидов в онлайн среде и особенностями личности [6 – 8], анализом распространения социоинженерных атакующих воздействий [9].

Однако сбор информации о риск-ассоциированном поведении может быть сопряжен со значительной неопределенностью, так как некоторые виды поведения труднодоступны для наблюдения и могут характеризоваться лишь в рамках самоотчетов самих индивидов. Для косвенной оценки искомых кумулятивных характеристик в условиях значительной неопределенности поступающей информации можно прибегать к математическому моделированию. В этом случае накладываются некоторые предположения, которые лежат в основе формализации процесса обработки неточной информации и разработки подходов к оцениванию интенсивности поведения. Естественной математической моделью для последовательностей риск-ассоциированных эпизодов являются точечные случайные процессы и гамма-пуассоновский процесс в частности [10, 11], возникает гамма-пуассоновская модель поведения. В этом случае основной характеристикой, отражающей паттерн реализации эпизодов, выступает *функция интенсивности* и для численного описания паттернов необходимо определить ее вид. В работе [12] было показано, что для гамма-пуассоновского процесса, функция интенсивности задается смешивающим гамма-распределением. Существуют несколько подходов к ее восстановлению по данным с неопределенностью: параметрический подход оценки параметра копулы и приближенные подходы классических и гибридных байесовских сетей доверия (БСД). Также подгонка регрессии Кокса позволяет получить оценку искомого параметра. Актуальной является задача анализа свойств этих подходов и выявления особенностей их применения для достижения наибольшей эффективности вычисления интенсивности в практических приложениях.

Целью работы является анализ особенностей применения существующих методов косвенного оценивания интенсивности рискованного поведения индивидов на основе ограниченных данных об эпизодах для определения рекомендаций по их применению.

Рассматриваемые методы имеют вероятностный характер, и потому исследование опирается на методы теории вероятностей и математической статистики.

Для достижения цели исследования поставлены следующие задачи:

1. Провести анализ требований к данным для существующих подходов к косвенному оцениванию искомой характеристики – кумулятивной интенсивности поведения индивидов.

2. Определить при помощи вычислительного эксперимента точность существующих подходов к косвенному оцениванию интенсивности поведения. Так как существующие подходы к оцениванию характеризуют искомое распределение с разных точек зрения: одни позволяют получить оценку параметра, а другие – подгонку самого распределения, то в качестве показателя точности была использовано расстояние Канторовича-Рубинштейна (метрика Вассерштейна) между вероятностными распределениями. Для достижения цели исследования предлагается рассмотреть несколько значений параметров исходного смешивающего распределения, которые отвечают обыденному поведению индивидов.

3. На основе сравнения существующих подходов по результатам пунктов 1 и 2 сформулировать рекомендации для применения конкретных методов в рамках в зависимости от исходных данных.

Статья организована следующим образом. Во втором разделе представлено описание существующих вероятностных подходов оценивания интенсивности поведения. В основе рассматриваемых подходов лежит смешанный пуассоновский процесс и его свойства, которые описаны в разделе 2.1. Этот раздел служит также для введения терминов, которые используются в дальнейшем при описании подходов и сравнении их между собой. Существующие подходы косвенной оценки характеристик поведения можно разделить на две группы: опирающиеся на оценки параметра смешивающего гамма-распределения (раздел 2.2) и опирающиеся на методы приближения смешивающего распределения с использованием байесовских сетей доверия (раздел 2.3). В соответствующих подразделах приведены схемы получения оценки, особое внимание уделялось объему наблюдений, которые требуются для построения оценки, так как каждый из подходов может использовать различные методы, и, соответственно, допускает вариабельность в исходных предположениях. Раздел 3 содержит описание вычислительного эксперимента, которые служит для определения точности рассматриваемых методов. Для проведения

эксперимента был сгенерирован набор данных, который отражает структуру гамма-пуассоновской модели поведения для нескольких наборов параметров, отвечающих обыденному поведению индивидов. С использованием этого набора данных были определены косвенные оценки интенсивности поведения с использованием рассматриваемых подходов, затем эти оценки сравнивались с истинным значением параметра. Так как два блока подходов различаются по сути получаемой косвенной оценки (оценка параметра распределения и подгонка распределения), то в качестве меры сходства было выбрано расстояние Канторовича-Рубинштейна. Раздел 4 содержит обсуждение проведенного эксперимента и применимости рассматриваемых в исследовании методов. Практической значимостью обладают сформулированные в разделе рекомендации по применению методов в зависимости от характеристик ситуации, в которой требуется получить косвенную оценку интенсивности. Отметим, что ранее такая работа не проводилась. Научный вклад исследования состоит в систематизации существующих подходов к косвенному оцениванию интенсивности поведения, а также в представлении качественных и количественных характеристик их применимости в ситуации ограниченных наблюдений.

2. Существующие подходы к оцениванию интенсивности поведения на основе ограниченных данных. Как уже отмечалось, необходимость использование математических моделей для построения оценок характеристик поведения возникает в силу неопределенности доступных данных и знаний о поведении. Так, в рамках одномоментного среза (опроса), индивид может предоставить информацию лишь о самых недавних или же о самых запоминающихся эпизодах [3], при этом оценка совокупных (кумулятивных) характеристик таких последовательностей эпизодов в группах индивидов используется при мониторинге социально значимого поведения. Автоматизированные системы обработки такой информации о поведении опираются на математические модели поведения, которые позволяют моделировать возникающую неопределенность. С вероятностной точки зрения базовым математическим объектом для таких моделей является пуассоновский процесс и его вариации [11].

2.1. Гамма-пуассоновская модель поведения. Естественной математической моделью для последовательностей риск-ассоциированных эпизодов являются точечные случайные процессы и гамма-пуассоновский процесс в частности [10, 11]. Смешивающее распределение моделирует индивидуальные особенности при оценке кумулятивных характеристик поведения в группах индивидов.

Такая неоднородность (гетерогенность) возникает в силу некоторых ненаблюдаемых индивидуальных особенностей, которые носят название *склонности к событиям* или *склонности к риску* [10].

Пусть рассматриваются последовательности эпизодов некоторого интересующего нас поведения в группе из n индивидов. Предполагается, что рассмативаемые последовательности удовлетворяют предположениям гамма-пуассоновской модели поведения [12]: эпизоды для каждого индивида в отдельности могут быть описаны процессом Пуассона, а индивидуальные особенности – некоторой случайной величиной u_i, i=1. . . n, для наблюдаемых индивидов u_i – независимые и одинаково распределенные случайные величины. В этом случае последовательность эпизодов, реализующаяся в группе индивидов полностью характеризуется функцией интенсивности:

$$\lambda(t|u_i) = u_i \rho_i(t), i = 1 \dots n,$$

здесь u_i моделируют индивидуальную склонность к реализации эпизодов для наблюдаемых объектов, а $\rho_i(t)$ – возможное влияние времени и внешних ковариант на реализацию эпизодов процесса. Если нет иных предположений о природе таких индивидуальных особенностей, то удобно полагать, что u_i имеют гамма-распределение вероятности с одним параметром $\phi > 0$:

$$f(s;\phi) = \frac{s^{\phi^{-1}-1}\exp(-s/\phi)}{\phi^{\phi^{-1}}\Gamma(1/\phi)}, s > 0. \quad (1)$$

Таким образом, исследуемый паттерн реализации эпизодов в группе неоднородных индивидов, описывается гамма-распределением величины, моделирующей эту неоднородность, или просто параметрами этого распределения ϕ. Существуют несколько подходов к оценке искомого гамма-распределения вероятности, которую обозначим g_u^{est}:

– подход регрессии Кокса;

– подход на основе классических БСД, который требует дискретизации входных данных модели;

– подход на основе гибридных БСД, который позволяет отказаться от дискретизации;

– оценка параметра копулы (который является искомым параметром функции интенсивности).

Для построения оценок используется доступная информация об эпизодах рассматриваемого поведения: для каждого респондента

i известна информация о датах m последних последовательных эпизодов, т.е. определены длины $m-1$ интервалов между последовательными эпизодами поведения $W_j^i, j = 1 \dots (m-1)$.

2.2. Методы оценки параметра смешивающего гамма-распределения в гамма-пуассоновской модели поведения. **Регрессионная модель Кокса** является классическим методом, использующимся при анализе процессов повторяющихся событий [11]. В этом случае в качестве зависимой переменной регрессии выступает функция риска, которая тесно связана с функцией интенсиности процесса [11]:

$$h(B(t)) = u_i \rho_i(t) = u_i \rho_0(t;\alpha)\exp(x_i(t)\beta),$$

где $B(t)$ – интервал времени с момента последнего эпизода, $\rho_0(t;\alpha)$ – базовая функция, отражающая поведение функции интенсивности при отсутствии воздействия ковариант, $x_i(t)$ – коварианты процесса, β –вектор коэффициентов регрессии.

Подгонка регрессии осуществляется при помощи метода максимального правдоподобия, и в случае включения в модель случайного эффекта, обуславливающего индивидуальные различия в форме гамма-распределения вероятности (1), строится также оценка его параметра ϕ. Однако, как показывают вычислительные эксперименты [13], мощность статистического теста на равенство параметра гамма-распределения нулю невысока, особенно если для каждого индивида наблюдаются менее 20 эпизодов. Отметим также, что в этом случае оценка максимума правдоподобия оказывается смещенной.

Применимость подхода обусловлена, с одной стороны, распространенностью соответствующего программного обеспечения, и, с другой стороны, тем, что этот метод опирается на подходы анализа времени жизни, которые широко применяются при оценке риска и язык которых известен в этой сфере деятельности. Также этот подход позволяет учитывать внешние факторы и их зависимость только в форме линейной регрессии.

Однако согласно [14], чтобы параметр регрессии был идентифицируемым, необходимо наблюдение нескольких ковариант процесса эпизодов, и потому в вычислительном эксперименте этот подход не участвует. Подход регрессии Кокса может использоваться в ситуациях, когда отсутствует нехватка данных и кроме самих эпизодов поведения наблюдаются иные факторы, которые объясняют вариабельность между индивидами.

Оценка параметра копулы, связывающей длины интервалов между последовательными эпизодами поведения в гамма-пуассоновской модели, является параметрическим методом получения искомого параметра. В основе этого подхода лежит свойство смешанного гамма-пуассоновского процесса, которое утверждает, что копулой зависимости нескольких длин интервалов между последовательными интервалами [15] выступает *копула Клейтона* [16].

В этом случае функцию их совместного распределения можно представить в виде [17]:

$$P(W_1^i > w_1, \ldots, W_m^i > w_m) = \left(\sum_{j=1}^{m} S_j^i(w_j)^{-\phi} - (m-1) \right)^{-\phi^{-1}} = \\ = C_{\text{clayton}}(S_1^i(w_1), \ldots, S_m^i(w_m)),$$

где $S_j^i(w_j) = P(W_j^i > w_j)$ – маржинальные распределения, ϕ – параметр гамма-распределения (1), $C_{\text{clayton}}(\cdot)$ – функция, объединяющая маржинальные распределения в совместное распределение вероятности

$$C_{\text{clayton}}(z_1, \ldots, z_m) = \Psi(\Psi(z_1) + \ldots + \Psi(z_m)),$$

$$\Psi(z) = \frac{1}{\phi}(z^{-\phi} - 1).$$

Таким образом, оценка параметра копулы, связывающей длины наблюдаемых интервалов между эпизодами поведения, является искомой оценкой параметра гамма-распределения вероятности u_i. Существуют различные методы построения оценки параметра копулы, в том числе классический метод максимальное правдоподобия или обращения рангового коэффициента корреляции, и для небольших выборок возможно смещение оценки. Отмечается [18], классические методы позволяют получить достоверные оценки, если наблюдается от 25 индивидов; однако для оценки параметра копулы Клейтона были предложены методы [19], позволяющие оценивать параметр всего по 10 наблюдениям. Особенностью применения этого метода является необходимость перехода к псевдо-наблюдениям, получаемым посредством интегрального преобразования вероятности.

Последовательность шагов построения оценки параметра при помощи аппарата копул можно сформулировать при помощи следующей последовательности шагов.

1. Собрать данные об m эпизодах для n объектов наблюдения, причем число m должно быть одинаковым для всех, $m \geqslant 2$ (в вычислительном эксперименте используется $m = 2$). Число объектов наблюдения может быть относительно невелико, порядка 10 или 50, и зависит от конкретного метода оценивания параметра копулы на последнем шаге. Собранные данные представляют собой набор длин интервалов между последовательными эпизодами поведения W_j^i.

2. Привести значения сформированной на шаге 1 выборки к равномерному распределению при помощи преобразования:

$$u_j^i = r_j^i/(n+1), i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\},$$

где r_j^i обозначает ранг наблюдения W_j^i среди всех $i \in 1, ..., n, j \in 1, ..., m$.

3. Оценить значение параметра функции Клейтона при помощи полупараметрического метода обращения коэффициента корреляции Кендалла [18] (или любым иным методом).

4. Для сравнения на основе полученной оценки параметра были сгенерированы 1000 реализаций гамма-распределенной случайной величины.

Полученная оценка является оценкой искомого параметра смешивающего гамма-распределения вероятности. В вычислительном эксперименте на основе этой оценки были сгенерированы значения гамма-распределенной случайной величины.

2.3. Методы приближения смешивающего распределения с использованием байесовских сетей доверия. Однако методы, опирающиеся на оценку параметра, обладают ограниченными возможностями для обработки данных с неопределенностью, которые часто возникают при обращении к самоотчетам и интервью [20]. Гибким инструментом для работы с такими данными являются БСД.

Классическая БСД в качестве модели оценивания интенсивности по ограниченным данным была предложена в работах [21,22]. Графическая структура сети формируется следующим образом: центральным узлом является переменная, моделирующая различия между индивидами в выборке u_i (она же позволяет вычислить среднее число эпизодов в промежутке времени), а детьми этого узла являются узлы, сопоставленные переменным-длинам интервалов между последовательными эпизодами. Таким образом, возникает звездчатая структура, которая обуславливает

разбиение совместного распределения вероятности (например, для трех интервалов между последовательными эпизодами поведения):

$$P(u^i < u, W_1^i < t_1, W_2^i < t_2, W_3^i < t_3) = P(u^i < u)P(W_1^i < t_1 \mid u^i) \\ P(W_2^i < t_2^i \mid u^i)P(W_3^i < t_3^i \mid u^i).$$

Подгонка искомого гамма-распределения в рамках этого подхода осуществляется с помощью проведения апостериорного вывода с использованием ранее опредленных параметров сети при заданной экспертным образом структуре сети [21]. Оценка параметров БСД осуществляется при помощи метода максимального правдоподобия [23, 24]. При этом априорные параметры модели могут быть заданы на основе статистической информации или же предоставлены экспертами.

Этот подход широко применяется в различных ситуациях, связанных с неопределенностью данных [24]. Подход классических БСД обладает высокой применимостью в приложениях благодаря развитому программно-аналитическому обеспечению процесса байесовского рассуждения, структура БСД отражает интуитивно понятные причинно-следственные связи в предметной области. Основным его недостатоком для построения оценок в рамках гамма-пуассоновской модели поведения является необходимость дискретизации поступающих данных о длинах инетрвалов между последовательными эпизодами, что значительно увеличивает число параметров, которое требуется оценить/задать для модели. Например, если для оценки функции интенсивности гамма-пуассоновского процесса используются k переменных, отвечающих длинам интервалов между последовательными эпизодами, каждая из которых разбита на 5 интервалов, то общее число параметров БСД для ядра гамма-пуассоновской модели, которые необходимо или оценить или задать экспертным путем составляет $24 * k$, что представляет большую нагрузку на данные или экспертов.

В качестве оценки искомого смешивающего распределения выступает апостериорное дискретное распределение вероятности, получаемое при пропагации свидетельств о длинах интервалов между эпизодами, и объем выборки не важен вследствие использования байесовского подхода.

Подход гибридных БСД обладает гибкостью классических БСД к моделированию неопределенности, и при этом отражает особенности гамма-пуассоновской модели эпизодического процесса,

как непрерывность составляющих ее переменных. Возможны различные подходы к заданию гибридной БСД, далее рассмотрено приближение дискретных распределений при помощи смесей усеченных базисных функций [25, 26]. Основной идеей метода является приближение совместной плотности n-мерного вектора $\mathbf{X}$, отражающего распределение переменных в БСД, при помощи наборов вещественных функций $\Psi = \{\psi_i\}_{i=0}^{\infty}, \psi_i : \mathbb{R} \to \mathbb{R}$:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{k} \prod_{j=1}^{c} a_{i,\mathbf{y}}^{(j)} \psi_i(z_j),$$

где $\mathbf{Y} = (Y_1, \ldots, Y_d)$ дискретная часть вектора $\mathbf{X}$, $\mathbf{Z} = (Z_1, \ldots, Z_c)$ – непрерывная часть; $a_{i,\mathbf{y}}^{(j)}$ вещественные числа. В качестве базисных функций ψ_i часто используются полиномы $\psi_i(x) = x^i$ и экспоненты $\Psi(x) = \{1, \exp(-x), \exp(-2x) \ldots\}$.

Для рассмотренной выше структуры БСД для гамма-пуассоновской модели поведения с тремя интервалами между последовательными эпизодами поведения, разложение с помощью усеченных экспонент выглядит следующим образом:

$$\hat{f}(u, t_1, t_2, t_3) = a_0 + \sum_{i=1}^{m} a_i \exp\left(b_i^{(u)} u + b_i^{(W_1)} t_1 + b_i^{(W_2)} t_2 + b_i^{(W_3)} t_3\right), \quad (2)$$

где m представляет собой глубину приближения базисными функциями.

При задании априорных параметров модели этот подход является достаточно требовательным к данным, однако как и все модели на основе БСД, модель может быть задана один раз, и затем использована для построения оценок интенсивности поведения. При использовании этого подхода оценка интенсивности поведения представляет собой приближенное распределение вероятности, получаемое с помощью пропагации свидетельств о длинах интервалов между последовательными эпизодами поведения.

Сама процедура численного задания исходной модели опирается на решение задачи квадратичной оптимизации для определения параметров смесей, описывающих маргинальные и условные распределения, отвечающие графу БСД [26]. Количество параметров таких смесей может варьироваться, соответственно, варьируется и объем необходимых

данных. Кроме того, алгоритмы приближения могут опираться на априорные знания экспертов.

Последовательность шагов построения косвенной оценки при помощи байесовских сетей доверия можно сформулировать при помощи следующей последовательности шагов.

1. Задать структуру зависимости наблюдаемых переменных, ядром которой является звездчатая структура гамма-пуассоновской модели. На этом этапе возможно включение в модель факторов внешней среды, недоопределенность знаний о их влиянии может быть квантифицирована экспертно или же статистически.

2. Сформировать тренировочный набор данных, который содержит достаточно большое число наблюдений интервалов между последовательными эпизодами исследуемого поведения и оцениваемой частоты, а также иные наблюдаемые факторы. В рамках вычислительного эксперимента сгенерированный набор данных был разбит на тестовую и тренировочную выборки в отношении 2:1.

3. (шаг для применения подхода классической БСД) Дискретизовать непрерывные данные об интервалах между последовательными эпизодами W_j^i на большое число интервалов. Для проведения вычислительного эксперимента использовалось разбиение на 8 интервалов при помощи метода взаимной информации Хартеминка [27].

4. При помощи специализированных алгоритмов (для смесей усеченных базисных функций [26], для классических байесовских сетей [27]) обучить параметры байесовской сети доверия на тестовой выборке. В исследовании использовался метод максимума правдоподобия.

5. Провести байесовский вывод для тестового набора поступающих данных (произвольного объема, от одного наблюдения) для получения апостреиорного распределения искомой переменной ϕ, отражающей смешивающее распределение гамма-пуассоновского процесса. В рамках вычислительного эксперимента для получения апостериорного распределения использовалась тестовая выборка сгенерированного тестового набора данных.

Существуют и иные подходы к работе с непрерывными распределениями в рамках формализма БСД, позволяющие использовать меньшее количество наблюдений для оценки априорных параметров, например, подход лозы [28]. В таблице 2 приведен сравнительный анализ существующих подходов с точки зрения применимости в условиях ограниченных ресурсов. Проанализированы возможности моделей к учету внешних факторов, влияющих на поведение, требования к данным

для задания той или иной модели, а также число параметров модели оценивания.

3. Точность методов оценивания интенсивности поведения в гамма-пуассоновской модели поведения по ограниченным данным. Анализ поведения человека является трудоемкой задачей, сбор данных в которой часто связан со значительными сложностями. Так, для прямой оценки интенсивности рискованного поведения используют ресурсозатратные способы прямого наблюдения или же дневниковые методы, которые не могут охватить все интересующие виды поведения [3]. Поэтому использование сгенерированных данных часто является необходимым шагом при оценке численных показателей автоматизированных систем обработки информации о поведении.

Особого внимания требует вопрос о релевантности ключевого объекта работы: гамма-пуассоновской модели поведения, реально наблюдаемому поведению индивидов. Эта модель опирается на классические предпосылки пуассоновского процесса (подробнее этот вопрос освещен в работе [12]), однако модели оценивания кумулятивной характеристики, рассматриваемые в работе, могут быть адаптированы к возможным отклонениям от этих предполсылок. Так, функция интенсивности процесса может включать зависимость от различных ковариант в форме регрессии [11], и подобные модели широко используются при анализе риска. Аналогично, использование БСД допускает внедрение узлов, которые могут аккумулировать отклонения от исходных предположений. Отметим также, что в ситуациях острой ограниченности ресурсов, когда доступны данные лишь о нескольких эпизодах для небольшого числа индивидов, нет возможности установить принадлежность наблюдаемых эпизодов тому или иному классу случайных процессов. Рассматривые в работе методы позволяют получить предварительную косвенную оценку уже на сверхмалых выборках, что важно на ранних этапах мониторинга поведения индивидов.

Для достижения цели исследования был проведен вычислительный эксперимент, в основе которого лежит генерация набора данных, отвечающего свойствам гамма-пуассоновской модели поведения, согласно шагам псевдокода 1. Его шаги отражают структуру гамма-пуассоновской модели поведения (раздел 2.1): сначала генерируется значение ненаблюдаемой случайной интенсивности процесса (шаг 11), имеющее гамма-распределение вероятности; затем на основе этого значения генерируются пять случайных величин, имеющие экспоненциальное распределение вероятности (шаги 12-16), на основании которых рассчитывались значения оценок с помощью различных подходов. Для

применения подхода на основе копул сгенерированные значения были отнормированы согласно рангам (шаг 21, шаги 1-6).

Для достижения цели работы использовались несколько значений исходного гамма-распределения, которые наиболее характерны для повседневного эпизодического поведения (шаг 7). Объем датасета составил 10000 наблюдений, что удовлетворяет требованиям существующих подходов к оценке, которые описаны в разделе 2. При генерации случайных величин были отброшены очень длинные интервалы, которые естественно возникают при обращении к пуассоновскому процессу с низкой интенсивностью, однако наблюдение которых в рамках самоотчета крайне маловероятно (функция 1-7 и шаги 22-27).

Листинг 1. Последовательность шагов генерации набора данных, отражающих стурктуру гамма–пуассоновской модели

```
1: function Truncate(Data)
2:     NewData ← Data
3:     for i ∈ (1 : NumberOfColumns(Data) − 1) do
4:         NewData[, i + 1] ← Data[which(Data[, i + 1] < 1000), i + 1]
5:     end for
6:     return NewData
7: end function
8: S ← [0.2, 0.5, 0.8, 1.0, 1.2, 1.5]
9: SimData ← DataFrame[nrow = 10000, ncol = 36]
10: for s ∈ S do
11:     for i ∈ [1, 10000] do
12:         λ_s[i] ← RandomGamma(s, 1/s)
13:         τ_s^1[i] ← RandomExponential(λ)
14:         τ_s^2[i] ← RandomExponential(λ)
15:         τ_s^3[i] ← RandomExponential(λ)
16:         τ_s^4[i] ← RandomExponential(λ)
17:         τ_s^5[i] ← RandomExponential(λ)
18:     end for
19:     MiniData ← ColumnBind(λ_s, τ_s^1, τ_s^2, τ_s^3, τ_s^4, τ_s^5)
20:     SimData ← ColumnBind(SimData, MiniData)
21: end for
22: data1 ← Truncate(simdata[, 1 : 6])
23: data2 ← Truncate(simdata[, 7 : 12])
24: data3 ← Truncate(simdata[, 13 : 18])
25: data4 ← Truncate(simdata[, 19 : 24])
26: data5 ← Truncate(simdata[, 25 : 30])
27: data6 ← Truncate(simdata[, 31 : 36])
```

С использованием сгенерированного набора данных, в среде статистической обработки данных R были получены оценки искомого гамма-распределения для разного количества наблюдений (20, 40, 60, 80) трех существующих подходов:

1. $g_u^{\text{ClassicBBN}}$ для подхода, подразумевающего получение апостериорного распределения на основе классических БСД; для

получения оценок использовался алгоритм взвешенного правдоподобия, реализованный в пакете bnlearn [27];

2. g_u^{copula} для подхода, в основе которого лежит оценка параметра копулы, связывающей две переменные – длины интервалов в гамма-пуассоновской модели поведения; использовался классический алгоритм обращения рангового коэффициента корреляции, реализованный в пакете copula [1];

3. $g_u^{\text{HybridBBN}}$ для подхода, подразумевающего получение эмпирического апостериорного распределения на основе гибридных БСД, где непрерывные распределения приближены с помощью смесей усеченных экспонент, реализованные в пакете MoTBFs [2].

При этом все рассматриваемые методы являются различными по сути получаемых оценок. Метод на основе копул выдает оценку параметра смешивающего гамма-распределения, и потому итогом является непрерывное распределение вероятности. Ранее разработанный подход с ядром в виде классической БСД дает оценку в виде дискретного распределения вероятности с небольшим числом ступеней, в то время как гибридные БСД, приближенные с помощью смесей усеченных распределений, позволяют получить эмпирическое непрерывное распределение неизвестного класса. В силу разнообразной природы оценок, получаемых в различных моделях, для сравнения их точности было решено использовать расстояние Канторовича-Рубинштейна (метрику Вассерштайна), которая является мерой удаленности двух вероятностных распределений друг от друга. Метрика Канторовича-Рубинштейна [29] представляет собой естественное расстояние на пространстве распределений вероятности.

В диссертационном исследовании вычисляется следующим образом. Пусть a и b представляют собой значения некоторой случайной величины, тогда их эмпирические функции распределения есть $F(t) = \sum_{i=1}^{m} w_i^{(a)} \mathbb{1}\{a_i \leqslant t\}$ и $G(t) = \sum_{j=1}^{n} w_j^{(b)} \mathbb{1}\{b_j \leqslant t\}$ ($w_i^{(a)}$ и $w_j^{(b)}$ представляют собой веса). Расстояние между такими функциями может быть вычислено как

$$W_p(F, G) = \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p}.$$

[1] Hofert M., Kojadinovic I., Maechler M. and Yan J. (2023). copula: Multivariate Dependence with Copulas. R package version 1.1-3 URL:https://CRAN.R- project.org/package=copula

[2] Pérez-Bernabé I., Salmerón A., Nielsen T.D., Maldonado A.D. (2022). MoTBFs: Learning Hybrid Bayesian Networks using Mixtures of Truncated Basis Functions. R package version 1.4.1, https://CRAN.R- project.org/package=MoTBFs

Здесь F^{-1} и G^{-1} – обобщенные обратные функции. Для $p = 1$ также выполнено:

$$W_1(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|dx.$$

Таблица 1 отражает рассчитанные расстояния Канторовича-Рубинштейна для различных моделей оценивания и различных значений исходных параметров. Расстояние вычислялось с использованием пакета transport[3].

4. Обсуждение. Основой целью работы является определение точности существующих методов оценивания интенсивности поведения. В таблице 1 представлены результаты вычислительного эксперимента, проведенного для оценки точности существующих подходов к оцениванию искомой характеристики. Отметим, что самые точные косвенные оценки позволяет получить подход, опирающийся на оценку параметра копулы, который при этом не позволяет напрямую учитывать контекст реализации эпизодов. Более гибкий подход БСД позволяет получать достаточно точные оценки при использовании неперерывных переменных в узлах; классический же подход дает более далекие от истинных значения.

Точность получаемой оценки является одним из ключевых факторов, обуславливающих применимость существующих моделей оценивания кумулятивной характеристики эпизодического поведения в группе индивидов, различающихся по склонности к поведению. Кроме этого важно учитывать нагрузка на данные, т.е. сколько данных необходимо использовать для получения оценки. С другой стороны важно учитывать контекст реализации эпизодов и возможное влияние внешних факторов.

Таблица 2 обобщает результаты работы, и представляет возможность выбора подхода в зависимости от имеющихся в распоряжении исследователя данных и целей его исследования. Качественное сравнение подходов к построению оценки опирается на описание вводных данных, представленное в 2 (пункт 1 для алгоритма оценки параметра копулы и пункты 1–2 для алгоритма построения приближенного распределения на основе БСД). Под относительной точностью понимается отношение значения расстояния Канторовича-Рубинштейна между оценкой, получаемой тем или иными методом, и истинным значением и расстояния между оценкой, получаемой методом

[3] Schuhmacher D., Bähre B., Bonneel N., Gottschlich C., Hartmann V., Heinemann F., Schmitzer B., Schrieber J. (2024). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.15-2. https://cran.r- project.org/package=transport

классической БСД, и истинным значением в случае, когда оценка строится по 20 наблюдениям для различных значений параметров гамма распределения. m – число рассматриваемых эпизодов, k – число интервалов дискретизации.

Таблица 1. Среднее значение и 95% бутстреп доверительный интервал для расстояния Канторовича-Рубинштейна между оценкой частоты, полученной при помощи различных подходов по n наблюдениям, и истинным смешивающим гамма-распределением, для 500 репликаций исходной выборки

Модель оценивания	Значение параметра исходной модели					
	$\phi = 1.5$	$\phi = 1.2$	$\phi = 1$	$\phi = 0.8$	$\phi = 0.5$	$\phi = 0.2$
Классическая БСД (исходная модель)						
$n = 20$	0.55, (0.18, 1.3)	0.94, (0.28, 1.8)	0.87, (0.24, 1.85)	1.12, (0.3, 2.48)	1.37, (0.33, 2.85)	2.96, (0.56, 6.35)
$n = 40$	0.49, (0.2, 0.99)	0.88, (0.36, 1.62)	0.81, (0.29, 1.44)	1.17, (0.47, 2.03)	1.30, (0.42, 2.54)	2.83, (0.83, 5.17)
$n = 60$	0.48, (0.22, 0.87)	0.87, (0.41, 1.43)	0.78, (0.33, 1.34)	1.12, (0.5, 1.92)	1.26, (0.48, 2.24)	2.84, (1.21, 4.85)
$n = 80$	0.46, (0.23, 0.75)	0.88, (0.45, 1.36)	0.78, (0.35, 1.27)	1.12, (0.6, 1.76)	1.28, (0.56, 2.13)	2.88, (1.51, 4.42)
Оценка параметра копулы						
$n = 20$	0.21, (0.03, 0.52)	0.20, (0.03, 0.52)	0.20, (0.03, 0.57)	0.20, (0.04, 0.51)	0.19, (0.04, 0.48)	0.30, (0.07, 0.62)
$n = 40$	0.14, (0.03, 0.38)	0.14, (0.03, 0.36)	0.14, (0.03, 0.36)	0.14, (0.03, 0.37)	0.14, (0.04, 0.35)	0.29, (0.09, 0.52)
$n = 60$	0.12, (0.03, 0.31)	0.12, (0.03, 0.32)	0.12, (0.03, 0.31)	0.11, (0.03, 0.26)	0.13, (0.04, 0.29)	0.28, (0.11, 0.49)
$n = 80$	0.11, (0.03, 0.27)	0.10, (0.03, 0.25)	0.10, (0.03, 0.26)	0.10, (0.03, 0.25)	0.12, (0.04, 0.29)	0.29, (0.13, 0.46)
Гибридная БСД (приближение усеченными базисными функциями)						
$n = 20$	0.26, (0.12, 0.54)	0.24, (0.12, 0.51)	0.26, (0.13, 0.53)	0.33, (0.15, 0.73)	0.45, (0.19, 1.03)	0.94, (0.35, 2.66)
$n = 40$	0.19, (0.09, 0.39)	0.18, (0.08, 0.38)	0.19, (0.1, 0.37)	0.25, (0.11, 0.52)	0.35, (0.14, 0.86)	0.84, (0.28, 2.12)
$n = 60$	0.16, (0.08, 0.32)	0.15, (0.07, 0.29)	0.16, (0.08, 0.3)	0.20, (0.09, 0.42)	0.30, (0.13, 0.65)	0.76, (0.26, 1.68)
$n = 80$	0.14, (0.07, 0.26)	0.13, (0.06, 0.24)	0.14, (0.08, 0.25)	0.18, (0.08, 0.37)	0.27, (0.12, 0.62)	0.78, (0.24, 1.63)

Анализ таблицы 2 позволяет сказать, что если не предполагается учитывать внешние факторы, то меньше всего требований к данным предъявляется при использовании подхода к оцениванию параметра искомого гамма-распределения при помощи копул. В этом случае существуют методы, которые позволяют строить оценку параметра копулы Клейтона всего по 10 наблюдениям.

Если говорить о ситуациях, когда важно учитывать влияние внешних факторов, то в этом случае используют подходы на основе БСД, которые после полного задания априорных значений параметров позволяют получать подгонку искомого гамма-распределения вероятности по любому количеству имеющихся данных и наблюдаемых интервалов для индивидов. Однако для использования байесовского рассуждения необходимо полностью задать модель БСД. Структура сети определена ранее, и различные подходы к заданию параметров БСД имеют различное число параметров к оценке. В целом, обе модели имеют сопоставимое число априорных параметров для задания. В случае классической БСД эти параметры представляют собой значения в таблицах условных

вероятностей, соответствующие ребрам, которые соединяют ключевую переменную и переменные-длины интервалов между эпизодами. Таких переменных будет $(m^2 - 1)$ для каждого узла. В случае гибридной БСД оцениваются параметры линейной комбинации базисных функций из разложения в формуле (2), и финальное число параметров зависит от числа элементов этого разложения. В целом, подгонка такого разложения требует большого числа наблюдений, порядка 30 для каждого из m параметров в (2).

Таблица 2. Сравнительный анализ методов оценивания интенсивности поведения с точки зрения результативности их применения в условиях ограниченных данных

Модель	Возможности учета влияния внешних факторов	Возможности обработки неточности поступающих значений	Требования к данным	Относительная точность получаемой оценки*
Параметрические подходы				
Регрессия пропорциональных рисков (Кокса)	Только в форме линейной регрессии	Нет	Более 20 наблюдаемых эпизодов для каждого индивида + наличие наблюдаемых факторов	–
Оценка параметра копулы	Нет	Нет	Оценивается 1 параметр; есть алгоритмы оценки для 10 и более наблюдений	0.21
Байесовские подходы				
БСД: классическая дискретизация	Да	Да	Задание модели происходит однократно на большом наборе данных или экспертно $(m*(k^2-1)$ параметров); затем для получения оценки требования к данным отсутствуют	1
Гибридная БСД: приближение смесями базисных функций	Да	Да	Задание модели происходит однократно на большом наборе данных; затем для получения оценки требования к данным отсутствуют	0.33

Отметим также, что в целом различные особенности входных данных могут учитываться моделями на основе БСД. Однако, поступающие данные, которые содержат различное количество эпизодов для разных индивидов, может учитывать также и подход регрессии Кокса.

Таким образом можно заключить, что переход к неперерывным моделям оценивания гамма-распределения позволяет получать распределение вероятности, более близкое к истинному значению, чем распределение, получаемое с помощью классической БСД. Можно

сформулировать следующие рекомендации по применению имеющихся моделей построения косвенной оценки интенсивности поведения:

– Если наблюдений много, и для каждого индивида наблюдается много эпизодов поведения, то подход *регрессии Кокса* является предпочтительным в силу развитого аппарата.

– Если наблюдений много, но о каждом индивиде известны только несколько эпизодов поведения, которые могут быть сопряжены с неопределенностью (например, получены в результате самоотчета), то подход *гибридных БСД* является предпочтительным в силу гибкости к учету возникающей неопределенности данных и знаний. Кроме того, впоследствии построенная модель гибридной БСД может использоваться для поддержки принятия решений в условиях острой нехватки наблюдений.

– Если же наблюдается небольшое число индивидов и несколько эпизодов для каждого из них, то достаточно точную численную оценку можно получить на основе оценки параметра копулы. Однако если при этом необходимо учитывать дополнительные факторы, то возможно использование гибридных БСД, заданных с помощью экспертов [30].

Среди ограничений проведенного исследования можно отметить небольшой разброс параметров гамма-распределения, которые использовались для формирования набора данных. Это обусловлено тем, что рассматривали те параметры, которые могут встречаться при моделировании обыденного поведения. Также в исследовании не были рассмотрены иные модели задания гибридных БСД.

5. Заключение. Для получения кумулятивных характеристик последовательностей эпизодов в группах индивидов существуют несколько подходов, которые опираются на математическую модель процесса событий: два из них опираются на оценку параметра искомого распределения, два – на предварительно заданную БСД. Для достижения цели исследования были решены:

1. Проведен анализ требований к данным для существующих подходов к косвенному оцениванию искомой характеристики. Основные результаты представлены в таблице 2. В результате анализа было установлено, что большими возможностями к учету возможного влияния сопутствующих факторов обладают подходы, основанные на БСД. Наименьшее число параметров требуется оценить при использовании параметрического подхода на основе оценки параметра копулы.

2. При помощи вычислительного эксперимента определена точность существующих подходов к косвенному оцениванию интенсивности поведения. Значения расстояния Канторовича-

Рубинштейна, полученные в рамках эксперимента, представлены в таблице 1. Показано, что для типовых значений параметров подходы, опирающиеся на непрерывные данные, обладают более низкими значениями расстояния Канторовича-Рубинштейна.

3. В разделе 4 были сформулированы рекомендации для применения конкретных методов в рамках в зависимости от исходных данных.

В работе проанализирована близость получаемой оценки к истинному значению в терминах расстояния Канторовича-Рубинштейна. Наивысшей точностью в сочетании с наиболее низкими требованиями к количеству наблюдаемых индивидов обладает подход на основе оценки параметра копулы, связывающей длины интервалов между последовательными эпизодами в гамма-пуассоновской модели поведения. При этом подход на основе копул не приспособлен для учета различной неопределенности, часто сопутствующей ситуации сбора самоотчетов.

Практической значимостью обладают сформулированные в таблице 2 особенности применимости различных подходов к оцениванию искомой характеристики. Этот результат является новым в сфере создания модельно-алгоритмического обеспечения систем обработки информации о поведении, и позволяет осуществлять выбор подходящего метода оценивания исходя из имеющихся данных. Дальнейшим направлением исследований является расширение вычислительного эксперимента для определения точности оценки гамма-распределений с иными параметрами, а также исследование точности оценок в ситуации ограниченных данных.

Литература

1. Chavez K., Palfai T.P., Cheng D.M., Blokhina E., Gnatienko N., Quinn E.K., Krupitsky E., Samet J.H. Hazardous Alcohol Use, Impulsivity, and HIV-Risk Behavior Among HIV-Positive Russian Patients With a History of Injection Drug Use // The American journal on addictions. 2021. vol. 30. no. 2. pp. 164–172.
2. Hendrieckx C., Ivory N., Singh H., Frier B.M., Speight J. Impact of severe hypoglycaemia on psychological outcomes in adults with type 2 diabetes: a systematic review // Diabetic Medicine. 2019. vol. 6. no. 9. pp. 1082–1091.
3. Пащенко А.Е., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Соколовский Е.В. Косвенная оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Здравоохранение Российской Федерации. 2010. № 2. С. 32–35.
4. Wojciechowski T.W. Major depressive disorder as a moderator of the relationship between heavy-episodic drinking and anxiety symptoms // Journal of mental health. 2023. pp. 1–8. DOI: 10.1080/09638237.2023.2245889.
5. Lewer D., Freer J., King E., Larney S., Degenhardt L., Tweed E.J., Hope V., Harris M., Millar T., Hayward A., Ciccarone D., Morley K. Frequency of health-care utilization by

adults who use illicit drugs: a systematic review and meta-analysis // Addiction. 2020. vol. 115. no. 6. pp. 1011–1023.

6. Feldhege J., Moessner M., Bauer S. Who says what? Content and participation characteristics in an online depression community // Journal of Affective Disorders. 2020. vol. 263. pp. 521–527.
7. Jiotsa B., Naccache B., Duval M., Rocher B., Grall-Bronnec M. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness // International journal of environmental research and public health. 2021. vol. 18. no. 6. DOI: 10.3390/ijerph18062880.
8. Олисеенко В.Д., Хлобыстова А.О., Корепанова А.А., Тулупьева Т.В. Автоматизация оценки темперамента пользователей онлайн социальной сети // Доклады Российской академии наук. Математика, информатика, процессы управления. 2023. Т. 514. № 2. С. 235–241. DOI: 10.31857/S2686954323601471.
9. Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users // Intelligent Distributed Computing XIII. Springer International Publishing, 2020. pp. 272–277.
10. Grandell J. Mixed Poisson Processes. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC. 1997. 280 p.
11. Cook R.J., Lawless J.F. The statistical analysis of recurrent events. Springer New York, 2007. 404 p.
12. Stoliarova V.F., Tulupyev A.L., Cox regression in the problem of risky behavior parameter estimation based on the last episodes' data // St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 2021. vol. 14(4). pp. 202–217. DOI: 10.18721/JPM.14415.
13. Rahgozar M., Faghihzadeh S., Babaee Rouchi G., Peng Y. The power of testing a semi-parametric shared gamma frailty parameter in failure time data // Statistics in medicine. 2008. vol. 27. no. 21. pp. 4328–4339.
14. Balan T.A., Putter H. A tutorial on frailty models // Statistical methods in medical research. 2020. vol. 29. no. 11. pp. 3424–3454.
15. Czado C. Analyzing dependent data with vine copulas // Lecture Notes in Statistics, Springer. 2019. 242 p.
16. Nelsen R.B. An introduction to copulas (Springer Series in Statistics). Springer, 2006. 286 p.
17. Столярова В.Ф. Копулы и моделирование зависимости: косвенные оценки интенсивности рискованного поведения // Компьютерные инструменты в образовании. 2018. № 3. С. 22–37.
18. Kojadinovic I., Yan J. Comparison of three semiparametric methods for estimating dependence parameters in copula models // Insurance: Mathematics and Economics. 2010. vol. 47. no. 1. pp. 52–63.
19. Qian L., Zhao Y., Yang J., Li H., Wang H., Bai C. A new estimation method for copula parameters for multivariate hydrological frequency analysis with small sample sizes // Water Resources Management. 2022. vol. 36. no. 4. pp. 1141–1157.
20. Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В., Красносельских Т.В. Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. 2010. № 4. С. 30–38.
21. Suvorova A., Tulupyev A. Learning Bayesian network structure for risky behavior modelling // Proceedings of the Third International Scientific Conference "Intelligent

Information Technologies for Industry" (ITI'18). Springer International Publishing, 2019. pp. 58–65.

22. Суворова А.В., Тулупьев А.Л., Сироткин А.В. Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления. 2014. Т. 9. № 2. С. 115–129.
23. Тулупьев А.Л., Николенко С.И., Сироткин А.В. Основы теории байесовских сетей. СПб: СПбГУ, 2019. 399 с.
24. Koller D., Friedman N. Probabilistic graphical models: principles and techniques. MIT press, 2009. 1230 p.
25. Langseth H., Nielsen T.D., Rumı R., Salmeron A. Mixtures of truncated basis functions // International Journal of Approximate Reasoning. 2012. vol. 53. no. 2. pp. 212–227.
26. Perez-Bernabe I., Maldonado A.D., Nielsen T.D., Salmeron A. Hybrid Bayesian Networks Using Mixtures of Truncated Basis Functions // R Journal. 2020. vol. 12. no. 2. pp. 321–341.
27. Scutari M., Denis J.-B. Bayesian Networks with Examples in R. 2nd edition. Chapman and Hall, Boca Raton. 2021. 274 p.
28. Czado C., Nagler T. Vine copula based modeling // Annual Review of Statistics and Its Application. 2022. vol. 9. no. 1. pp. 453–477.
29. Kolouri S., Kolouri S., Park S.R., Thorpe M., Slepcev D., Rohde G.K. Optimal mass transport: Signal processing and machine-learning applications // IEEE signal processing magazine. 2017. vol. 34. no. 4. pp. 43–59.
30. Hanea A.M., Hemming V., Nane G.F. Uncertainty quantification with experts: present status and research needs // Risk Analysis. 2022. vol. 42. no. 2. pp. 254–263.

Столярова Валерия Фуатовна — младший научный сотрудник, лаборатория теоретических и междисциплинарных проблем информатики, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: анализ данных, байесовские сети доверия, искусственный интеллект. Число научных публикаций — 40. vfs@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3337.

Тулупьева Татьяна Валентиновна — канд. психол. наук, доцент, старший научный сотрудник, лаборатория теоретических и междисциплинарных проблем информатики, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН); профессор кафедры, кафедра государственного и муниципального управления, Северо-Западный институт управления РАНХиГС. Область научных интересов: психология личности, искусственный интеллект, социальная инженерия, методы обработки данных. Число научных публикаций — 170. tvt@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3337.

Вяткин Артем Андреевич — младший научный сотрудник, лаборатория теоретических и междисциплинарных проблем информатики, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: байесовские сети доверия, искусственный интеллект. Число научных публикаций — 10. aav@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3337.

Поддержка исследований. Исследование выполнено в рамках инициативной НИР СЗИУ РАНХиГС при Президенте РФ, номер в системе ЕГИСУ НИОКТР 122112900066-6, а также в рамках проекта по государственному заданию СПБ ФИЦ РАН СПИИРАН № FFZF-2022-0003.

DOI 10.15622/ia.23.6.6

V. Stoliarova , T. Tulupyeva , A. Vyatkin

APPROACHES FOR BEHAVIOR INTENSITY ESTIMATION IN GROUPS OF HETEROGENEOUS INDIVIDUALS: PRECISION AND APPLICABILITY FOR DATA WITH UNCERTAINTY

Stoliarova V., Tulupyeva T., Vyatkin A. **Approaches for Behavior Intensity Estimation in Groups of Heterogeneous Individuals: Precision and Applicability for Data with Uncertainty.**

Abstract. In socially oriented areas, there arises the problem of assessing the cumulative characteristics of behavior, such as intensity, that are realized in groups of individuals. All individuals vary in their behavior and the available data is limited and may be associated with significant uncertainty: only a few episodes may be known and only a few individualsi the group may be observed. Mathematical models of behavior are used for estimation of key characteristics of the behavior. One of them is based on the gamma–Poisson point process, that reflects the heterogeneity of individuals in a form of a mixing distribution. This general model allows to formulate several methods of frequency estimation: the Cox regression, estimation of the copula parameter, and a posteriori inference in Bayesian belief networks. The aim of the paper is to assess their determine the precision of these methods based on the Kantorovich–Rubinstein distance between estimates and the true distribution of the desired parameter. The analysis of assumptions of those methods allows to formulate rules, that allow to chose the appropriate method in various sutuations of data availability. It has been shown that the copula-based approach provides the most accurate estimates and has the mild assumptions for the number of observed objects, but it cannot take into account external factors that may influence the behavior. Among methods that can take into account process covariants, estimates based on a posteriori inference in hybrid Bayesian belief networks have the highest precision. The paper considers a method for quantification of a hybrid BBNs with the approximation of mixtures of truncated exponents, that is data-demanding at the stage of calculating a priori estimates. However, it is noted that there are other approaches to setting hybrid BSDs in which a priori estimates can be set completely expertly.

Keywords: last episodes, risky behavior, Bayesian belief networks, hybrid bayesian belief networks, copula.

References

1. Chavez K., Palfai T.P., Cheng D.M., Blokhina E., Gnatienko N., Quinn E.K., Krupitsky E., Samet J.H. Hazardous Alcohol Use, Impulsivity, and HIV-Risk Behavior Among HIV-Positive Russian Patients With a History of Injection Drug Use. The American journal on addictions. 2021. vol. 30. no. 2. pp. 164–172.
2. Hendrieckx C., Ivory N., Singh H., Frier B.M., Speight J. Impact of severe hypoglycaemia on psychological outcomes in adults with type 2 diabetes: a systematic review. Diabetic Medicine. 2019. vol. 36. no. 9. pp. 1082–1091.
3. Paschenko A., Tulupyev A., Tulupyeva T., Krasnoselskikh T., Sokolovsky E. [Indirectassessment of the probability of HIV infection based on data on recent episodes of risky behavior]. Zdravoohranenie Rossijskoj Federacii – Healthcare of the Russian Federation. 2010. no. 2. pp. 32–35. (In Russ.).
4. Wojciechowski T.W. Major depressive disorder as a moderator of the relationship between heavy-episodic drinking and anxiety symptoms. Journal of mental health. 2023. pp. 1–8. DOI: 10.1080/09638237.2023.2245889.

5. Lewer D., Freer J., King E., Larney S., Degenhardt L., Tweed E.J., Hope V., Harris M., Millar T., Hayward A., Ciccarone D., Morley K. Frequency of health-care utilization by adults who use illicit drugs: a systematic review and meta-analysis. Addiction. 2020. vol. 115. no. 6. pp. 1011–1023.
6. Feldhege J., Moessner M., Bauer S. Who says what? Content and participation characteristics in an online depression community. Journal of Affective Disorders. 2020. vol. 263. pp. 521–527.
7. Jiotsa B., Naccache B., Duval M., Rocher B., Grall-Bronnec M. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness. International journal of environmental research and public health. 2021. vol. 18. no. 6. DOI: 10.3390/ijerph18062880.
8. Oliseenko V., Khlobystova A., Korepanova A., Tulupyeva T. [Automatization of the assessment of the temperament of users of an online social networ]. Doklady Rossijskoj akademii nauk. Matematika, informatika, processy upravlenija – Reports of the Russian Academy of Sciences. Mathematics, computer science, management processes. 2023. vol. 514. no. 2. pp. 235–241. DOI: 10.31857/S2686954323601471. (In Russ.).
9. Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users. Intelligent Distributed Computing XIII. Springer International Publishing, 2020. pp. 272–277.
10. Grandell J. Mixed Poisson Processes. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC. 1997. 280 p.
11. Cook R.J., Lawless J.F. The statistical analysis of recurrent events. Springer New York, 2007. 404 p.
12. Stoliarova V.F., Tulupyev A.L., Cox regression in the problem of risky behavior parameter estimation based on the last episodes' data. St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 2021. vol. 14(4). pp. 202–217. DOI: 10.18721/JPM.14415.
13. Rahgozar M., Faghihzadeh S., Babaee Rouchi G., Peng Y. The power of testing a semi-parametric shared gamma frailty parameter in failure time data. Statistics in medicine. 2008. vol. 27. no. 21. pp. 4328–4339.
14. Balan T.A., Putter H. A tutorial on frailty models. Statistical methods in medical research. 2020. vol. 29. no. 11. pp. 3424–3454.
15. Czado C. Analyzing dependent data with vine copulas. Lecture Notes in Statistics, Springer. 2019. 242 p.
16. Nelsen R.B. An introduction to copulas (Springer Series in Statistics). Springer, 2006. 286 p.
17. Stoliarova V. Copula and dependency modeling: indirect estimates of the intensity of risky behavior. Komp'yuternye instrumenty v obrazovanii – Computer tools in education. 2018. no. 3. pp. 22–37. (In Russ.).
18. Kojadinovic I., Yan J. Comparison of three semiparametric methods for estimating dependence parameters in copula models. Insurance: Mathematics and Economics. 2010. vol. 47. no. 1. pp. 52–63.
19. Qian L., Zhao Y., Yang J., Li H., Wang H., Bai C. A new estimation method for copula parameters for multivariate hydrological frequency analysis with small sample sizes. Water Resources Management. 2022. vol. 36. no. 4. pp. 1141–1157.
20. Suvoriva A., Tulupyev A., Paschenko A., Tulupyeva T., Krasnoselskikh T. Analysis of granular data and knowledge in the problems of researching socially significant behaviors.

Komp'yuternye instrumenty v obrazovanii – Computer tools in education. 2010. no. 4. pp. 30–38. (In Russ.).
21. Suvorova A., Tulupyev A. Learning Bayesian network structure for risky behavior modelling. Proceedings of the Third International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'18). Springer International Publishing, 2019. pp. 58–65.
22. Suvorova A., Tulupyev A., Sirotkin A. [Bayesian belief networks for the problems of assessing the intensity of risky behavior]. Nechetkie sistemy i myagkie vychisleniya – Fuzzy Systems and Soft Computing. 2014. vol. 9. no. 2. pp. 115–129. (In Russ.).
23. Tulupyev A.L., Nikolenko S.I., Sirotkin A.V. Osnovy teorii bajesovskih setej – Basics of Bayesian network theory. SPb: SPbGU, 2019. 399 p. (In Russ.).
24. Koller D., Friedman N. Probabilistic graphical models: principles and techniques. MIT press, 2009. 1230 p.
25. Langseth H., Nielsen T.D., Rumı R., Salmeron A. Mixtures of truncated basis functions. International Journal of Approximate Reasoning. 2012. vol. 53. no. 2. pp. 212–227.
26. Perez-Bernabe I., Maldonado A.D., Nielsen T.D., Salmeron A. Hybrid Bayesian Networks Using Mixtures of Truncated Basis Functions. R Journal. 2020. vol. 12. no. 2. pp. 321–341.
27. Scutari M., Denis J.-B. Bayesian Networks with Examples in R. 2nd edition. Chapman and Hall, Boca Raton. 2021. 274 p.
28. Czado C., Nagler T. Vine copula based modeling. Annual Review of Statistics and Its Application. 2022. vol. 9. no. 1. pp. 453–477.
29. Kolouri S., Kolouri S., Park S.R., Thorpe M., Slepcev D., Rohde G.K. Optimal mass transport: Signal processing and machine-learning applications. IEEE signal processing magazine. 2017. vol. 34. no. 4. pp. 43–59.
30. Hanea A.M., Hemming V., Nane G.F. Uncertainty quantification with experts: present status and research needs. Risk Analysis. 2022. vol. 42. no. 2. pp. 254–263.

Stoliarova Valerie — Junior research fellow, Laboratory of theoretical and interdisciplinary problems of computer science, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: data analysis, probabilistic graphical models. The number of publications — 40. vfs@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-3337.

Tulupyeva Tatiana — Ph.D., Associate Professor, Senior research fellow, Laboratory of theoretical and interdisciplinary problems of computer science, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS); Professor of the department, Department of state and municipal management, North-West Institute of management (NWIM), the branch of the Russian Presidential Academy of National Economy and Public Administration. Research interests: personal psychology, artificial intelligence, social engineering, data analysis. The number of publications — 170. tvt@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-3337.

Vyatkin Artyom — Junior research fellow, Laboratory of theoretical and interdisciplinary problems of computer science, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: data analysis, probabilistic graphical models. The number of publications — 10. aav@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-3337.

Acknowledgements. The study was carried out within the framework of the initiative research NWIM branch RANEPA, № 122112900066-6 and within the framework of the project under the state order of the St. Petersburg Federal Research Center of the Russian Academy of Sciences SPIIRAN No. FFZF-2022-0003.