

ИНДУЦИРОВАННАЯ ЗАВИСИМОСТЬ, ВЗАИМОДЕЙСТВИЕ ФАКТОРОВ И ДИСКРИМИНАЦИЯ КАУЗАЛЬНЫХ СТРУКТУР

Аннотация. Исследованы свойства модели (коллайдера), где две независимые переменные (причины) влияют на третью переменную (эффект). Показано, что в результате кондиционирования эффекта его причины, как правило, становятся взаимно зависимыми. Провоцированная (индуцированная) зависимость изучена количественно. Выявлена тесная связь спровоцированной зависимости с взаимодействием факторов; предложен индекс взаимодействия двух факторов. Показана эффективность применения спровоцированной зависимости для идентификации направлений связей и обнаружения маскированных связей (ребер) модели.

Ключевые слова: условная независимость, спровоцированная (наведенная) зависимость, коллайдер, взаимодействие факторов, маскированные ребра, идентификация связей, ориентация связей.

ВВЕДЕНИЕ

В статье анализируется система зависимостей, где две взаимно независимые переменные (факторы, причины) влияют на третью переменную — их общий эффект (отклик). Известно, что в результате селекции (или стратификации) данных по общему эффекту факторы становятся взаимозависимыми, т.е. возникает условная зависимость между причинами. Этот феномен иногда называют «фальшивой ассоциацией», парадоксом Берксона [1] и т.д. Рассматриваемый феномен тесно связан с проблемой подбора ковариат в анализе данных. Провоцированная (индуцированная) зависимость приводит к смещению оценок влияния факторов (selection bias). Смещение оценок необходимо учитывать в эконометрических, эпидемиологических, медико-биологических исследованиях (для оценки каузального эффекта на основе данных наблюдений [2–5] и в целом в каузальном моделировании [4–7]). В определенных условиях удастся устранить смещение из оценки коэффициентов влияния в линейной модели [2, 5]. В [7–9] феномен спровоцированной зависимости используется как инструмент в идентификации структуры модели. В настоящей статье изучаются свойства спровоцированной зависимости и показана ее возможная роль.

1. ФОРМУЛИРОВКА ПРОБЛЕМЫ

Явление спровоцированной (наведенной) зависимости можно пояснить на тривиальном примере. Пусть задано множество троек значений переменных X, Y, Z и известно, что в каждой тройке значение Y вычислено по формуле $y = x + z$, а значения переменных X и Z случайно и независимо выбраны из некоторых распределений плотностей $f(x)$ и $f(z)$. Сообщение о том, что значение переменной X равно x_i , не изменяет знаний о значении переменной Z , оно по-прежнему распределено согласно $f(z)$. Пусть известно, что $Y = y_k$. Тогда, дополнительно получив сообщение, что $X = x_i$, сразу делаем вывод, что $Z = y_k - x_i$. Наша неопределенность относительно Z полностью снимается. Можно привести примеры, где в подобной ситуации значение Z уточняется до подмножества значений. В более реалистических ситуациях ввиду влияния ненаблюдаемых случайных факторов и «шума» снижение (изменение) неопределенности значения Z выражается в форме изменения вероятности соответствующих значений. В общем виде указанный феномен описывается в вероятностной форме: условное распределение $p(Z|Y, X)$ отличается от условного распреде-

ления $p(Z|Y)$. (Заметим, что введение условия на значение Y не моделирует воздействия на Y ; воздействие на носитель Y в объекте не повлияло бы на взаимосвязь X и Z (см. [6])).

Введем необходимые понятия и обозначения. Пусть имеем набор переменных $\mathbf{S} \cup \{X, Z\}$.

(Для краткости будем писать XYS вместо $\mathbf{S} \cup \{X, Y\}$.) В общем виде условная независимость переменных X и Z при условии \mathbf{S} определяется формулой $p(XZ|\mathbf{S}) \cdot p(\mathbf{S}) = p(X|\mathbf{S}) \cdot p(Z|\mathbf{S})$. Такое определение остается в силе и для вероятностей (для дискретных переменных), и для плотностей вероятностей (для непрерывных переменных). Безусловная независимость есть частный случай условной (при $\mathbf{S} = \emptyset$). Обозначим факт условной независимости предикатом $\text{Ind}(X, Z|\mathbf{S})$. Когда независимость $\text{Ind}(X, Z|\mathbf{S})$ не выполняется, считают, что переменные X и Z условно зависимы при \mathbf{S} ; обозначаем этот факт как $\text{Dep}(X, Z|\mathbf{S})$.

Если $p(\mathbf{S}) \neq 0$, то для независимости $\text{Ind}(X, Z|\mathbf{S})$ получаем часто используемую форму

$$p(XZ|\mathbf{S}) = p(X|\mathbf{S}) \cdot p(Z|\mathbf{S}). \quad (1)$$

Далее, если $p(X|\mathbf{S}) \neq 0$, то получаем форму $p(Z|X\mathbf{S}) = p(Z|\mathbf{S})$. Факт $\text{Ind}(X, Z|\mathbf{S})$ можно выразить через взаимную информацию между переменными (по Шеннону) как $\text{Info}(X, Z|\mathbf{S}) = 0$. Положив $\mathbf{S} \equiv Y$, получим частный случай условной независимости $p(XZ|Y) = p(X|Y) \cdot p(Z|Y)$.

Определение. Провоцированная зависимость между переменными X и Z с условием Y — это паттерн $\text{Ind}(X, Z) \& \text{Dep}(X, Z|Y)$, т.е. сочетание отношений

$$p(XZ) = p(X) \cdot p(Z); \quad p(XZY) \cdot p(Y) \neq p(XY) \cdot p(ZY). \quad (2)$$

Такую зависимость провоцирует Y .

Наиболее системно и точно проблема описывается в аппарате структуральных уравнений, вероятностных графовых моделей и каузальных сетей [4–6]. Будем рассматривать ситуацию, когда имеем $\text{Ind}(X, Z)$, $\text{Dep}(X, Y)$ и $\text{Dep}(Y, Z)$. Эту ситуацию можно проиллюстрировать каузальными сетями (рис. 1). Пару смежных дуг (ребер), «сталкивающихся» своими стрелками (рис. 1, а), принято называть коллайдером, или коллизором. Такая модель количественно описывается через параметры в виде $p(XYZ) = p(X) \cdot p(Z) \cdot p(Y|XZ)$. Аналогично в структурах моделей, показанных на рис. 1, б, в, переменная Y провоцирует зависимость между X и Z . Для всех этих структур обязательность спровоцированной зависимости имплицитруется предположением каузальной необманчивости [6]. Это предположение распространяется на разнообразные ситуации, но в особых случаях может не выполняться. (Практически проблема возникает уже тогда, когда предположение каузальной необманчивости почти не выполняется.) В случае, когда модель имеет структуру коллайдера, предположение каузальной необманчивости имплицитует паттерн $\text{Ind}(X, Z)$, $\text{Dep}(X, Z|Y)$, $\text{Dep}(X, Y)$, $\text{Dep}(Z, Y)$.

2. НЕКОТОРЫЕ ХАРАКТЕРНЫЕ СВОЙСТВА МОДЕЛИ

Если в коллайдере (см. рис. 1, а) одну или две дуги развернуть (реверсировать), то получится цепь зависимостей (рис. 2). В частности, цепочка, отображенная на рис. 2, а, описывается через параметры в виде $p(XYZ) = p(X) \cdot p(Y|X) \cdot p(Z|Y)$. Все эти цепочки подчиняются марковскому свойству $\text{Ind}(X, Z|Y)$. Предположе-

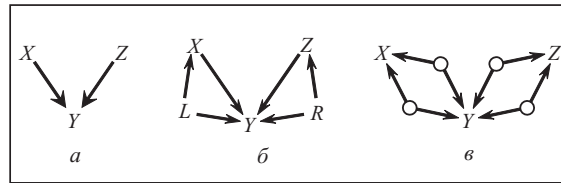


Рис. 1. Каузальные сети: простой коллайдер (а); варианты, где также формируется провоцированная зависимость (б, в)

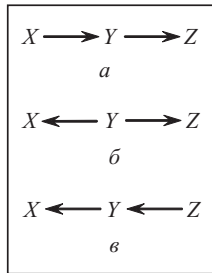


Рис. 2. Цепи зависимости: однонаправленные (а, в) и дивергентная (б) цепи

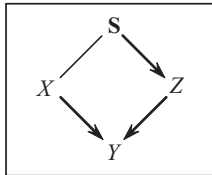


Рис. 3. Коллайдер в «контексте»

ние каузальной необманчивости для этих структур имплицитно определяет $\text{Dep}(X, Z)$. Констатируем, что структура коллайдера и структура цепочки принципиально различаются своими свойствами; их паттерны (не)зависимости — противоположны. Паттерн в виде цепочки характеризуется $\text{Dep}(X, Z) \& \text{Ind}(X, Z | Y)$. Коллайдерный паттерн — $\text{Ind}(X, Z) \& \text{Dep}(X, Z | Y)$. Ясно, что третье сочетание, а именно $\text{Ind}(X, Z) \& \text{Ind}(X, Z | Y)$ легко получить (например, когда все три переменные не взаимосвязаны). Однако если в модели присутствуют два ребра: $(X - Y)$ и $(Y - Z)$, то сочетание $\text{Ind}(X, Z) \& \text{Ind}(X, Z | Y)$ есть аномальный паттерн. Такая аномалия может возникнуть либо в цепочке (в результате коллапса «транзитной» зависимости), либо в коллайдере (в результате коллапса спровоцированной зависимости).

Пример 1. Рассмотрим модель со структурой цепи $X \rightarrow Y \rightarrow Z$, где переменные X и Z — бинарные, переменная Y — трехзначная и заданы параметры

$$p(Y | x_1) = (1/4; 5/16; 7/16), \quad p(Y | x_2) = (1/2; 3/8; 1/8),$$

$$p(Z | y_1) = (2/3; 1/3), \quad p(Z | y_2) = (7/8; 1/8),$$

$$p(Z | y_3) = (17/24; 7/24).$$

Тогда для любого распределения $p(X)$ получим $p(Z) = (3/4; 1/4)$, т.е. справедлива независимость $\text{Ind}(X, Z)$, происходит коллапс «транзитной» зависимости, который объясняется специфическим сочетанием значений параметров.

Далее покажем, насколько нетипичны обстоятельства, в которых спровоцированная зависимость не формируется. Помимо «простого» коллайдера (см. рис. 1, а) будем рассматривать коллайдер, «вложенный в контекст» (рис. 3). В этом случае переменные X и Z ассоциированы посредством других переменных и существует множество S , обеспечивающее условную независимость $\text{Ind}(X, Z | S)$.

Будем исключать из рассмотрения те значения переменных, которые имеют нулевую маргинальную вероятность. Тогда (подразумеваемая независимость X и Z) имеем $p(XZ) \neq 0$.

Лемма 1. Если в совместном условном распределении вероятностей $p(XYZ | S)$ выполняется $\text{Ind}(X, Z | S) \& \text{Ind}(X, Z | YS)$, то справедливо равенство

$$p(Y | XZS) = p(Y | XS) \cdot p(Y | ZS) / p(Y | S).$$

Доказательство. Из факта $\text{Ind}(X, Z | YS)$ согласно (1) следует

$$p(XZYS) = p(X | YS) \cdot p(Z | YS) \cdot p(YS).$$

Отсюда получаем

$$p(Y | XZS) = p(X | YS) \cdot p(Z | YS) \cdot p(Y | S) / p(XZ | S). \quad (3)$$

Факт $\text{Ind}(X, Z | S)$ означает, что

$$p(XZ | S) = p(X | S) \cdot p(Z | S). \quad (4)$$

Подставляем (4) в (3) и с помощью тождеств условной вероятности получаем

$$p(Y | XZS) = p(X | YS) \cdot p(Z | YS) \cdot p(Y | S) / (p(X | S) p(Z | S)) =$$

$$= p(X | YS) \cdot p(Z | YS) \cdot p^2(Y | S) / (p(X | S) p(Z | S) p(Y | S)) =$$

$$= p(Y | XS) \cdot p(Y | ZS) / p(Y | S).$$

В частном случае, когда $\mathbf{S} = \emptyset$, получаем следствие

$$p(Y | XZ) = p(Y | X) \cdot p(Y | Z) / p(Y). \quad (5)$$

В анализе данных гипотеза независимости верифицируется с помощью статистических тестов. В тестовую статистику встроены определенные показатели величины зависимости, который оценивает суммарную степень отклонения от независимости [10, 11]. В случае непрерывных переменных выбрать эффективный показатель зависимости — задача сложная. (Для вычисления взаимной информации между переменными на основе выборки данных с непрерывными переменными требуется аппроксимация неизвестной плотности вероятностей.)

3. СЛУЧАЙ ДИСКРЕТНЫХ ПЕРЕМЕННЫХ

Уточним и обобщим известное свойство дискретных моделей [8]. Рассмотрим коллаидер в «контексте» \mathbf{S} .

Лемма 2. Если в совместном условном распределении вероятностей $p(XYZ | \mathbf{S})$, где переменная Y — бинарная, выполняется $\text{Ind}(X, Z | \mathbf{S}) \& \text{Ind}(X, Z | Y\mathbf{S})$, то справедливо равенство

$$(p(Y | \mathbf{S}) - p(Y | X\mathbf{S})) \cdot (p(Y | \mathbf{S}) - p(Y | Z\mathbf{S})) = 0. \quad (6)$$

Доказательство. Согласно лемме 1 из фактов $\text{Ind}(X, Z | \mathbf{S})$ и $\text{Ind}(X, Z | Y\mathbf{S})$ следует

$$p(Y | XZ\mathbf{S}) = p(Y | X\mathbf{S}) \cdot p(Y | Z\mathbf{S}) / p(Y | \mathbf{S}). \quad (7)$$

Пусть переменная Y имеет значения $\{y_1, y_2\}$. Тогда

$$p(y_1 | XZ\mathbf{S}) + p(y_2 | XZ\mathbf{S}) = 1. \quad (8)$$

Запишем два варианта уравнения (7) для каждого значения y_1 и y_2 и просуммируем их:

$$p(y_1 | \mathbf{S})p(y_2 | \mathbf{S}) = p(y_2 | \mathbf{S})p(y_1 | X\mathbf{S}) \cdot p(y_1 | Z\mathbf{S}) + \\ + p(y_2 | \mathbf{S})p(y_1 | X\mathbf{S}) \cdot p(y_1 | Z\mathbf{S}).$$

Отсюда, выразив $p(y_2 | \cdot)$ через $p(y_1 | \cdot)$ с помощью (8), после алгебраических преобразований получим

$$(p(y_1 | \mathbf{S}) - p(y_1 | X\mathbf{S})) \cdot (p(y_1 | \mathbf{S}) - p(y_1 | X\mathbf{S})) = 0. \quad (9)$$

Выразив $p(y_1 | \cdot)$ через $p(y_2 | \cdot)$, преобразуем (9) к виду $(p(y_2 | \mathbf{S}) - p(y_2 | X\mathbf{S})) \cdot (p(y_2 | \mathbf{S}) - p(y_2 | X\mathbf{S})) = 0$. Итак, получили (6).

В частном случае, когда $\mathbf{S} = \emptyset$, получаем следующий результат.

Следствие 1. Если в совместном распределении вероятностей $p(XYZ)$, где переменная Y — бинарная, выполняется $\text{Ind}(X, Z) \& \text{Ind}(X, Z | Y)$, то справедливо равенство $(p(Y) - p(Y | X)) = 0$ или $(p(Y) - p(Y | Z)) = 0$.

Тонкое, но важное различие частного ($\mathbf{S} = \emptyset$) и общего случаев заключается в следующем. В частном случае из посылок леммы 2 можно сделать вывод, что отсутствует зависимость между переменными Y и X или между Y и Z . В общем случае такой вывод сделать нельзя, поскольку при разных значениях «контекста» \mathbf{S} в (6) первый или второй множитель может попеременно обращаться в нуль. Поэтому нельзя утверждать, что всегда $p(Y | \mathbf{S}) = p(Y | X\mathbf{S})$ или всегда $p(Y | \mathbf{S}) = p(Y | Z\mathbf{S})$.

Следствие 1 означает, в частности, что невозможно построить такую цепь $X \rightarrow Y \rightarrow Z$ с бинарной Y , чтобы транзитная зависимость $Z(X)$ коллапсировала, т.е. невозможно получить $\text{Ind}(X, Z)$. Поэтому для конструирования примера 1 необходима не-бинарность переменной Y .

Сопоставим частные случаи леммы 1 (формула (5)) и леммы 2, получим следующий результат.

Следствие 2. Если в совместном распределении вероятностей $p(XYZ)$, где переменная Y — бинарная, выполняется $\text{Ind}(X, Z)$ и $\text{Ind}(X, Z | Y)$, то справедливо равенство $p(Y | XZ) = p(Y | Z)$ или $p(Y | XZ) = p(Y | X)$.

Содержательно это означает, что в указанных ситуациях нет зависимости переменной Y от обеих X и Z одновременно. Иными словами, невозможен такой коллаيدر $X \rightarrow Y \leftarrow Z$ с бинарной переменной Y , для которого не возникает провоцированной зависимости. Полученные результаты можно суммировать в следующем виде.

Утверждение 1. Пусть в совместном распределении вероятностей $p(XYZ)$, где переменная Y — бинарная, выполняется $\text{Ind}(X, Z)$, $p(Y | XZ) \neq p(Y | Z)$ и $p(Y | XZ) \neq p(Y | X)$. Тогда будет выполняться $\text{Dep}(X, Z | Y)$.

В [12] показано: в цепи $X \rightarrow Y \rightarrow Z$ с бинарной Y транзитная зависимость не коллапсирует. Итак, в модели с двумя ребрами: $(X - Y)$ и $(Y - Z)$, где Y — бинарная, невозможен аномальный паттерн $\text{Ind}(X, Z) \& \text{Ind}(X, Z | Y)$.

Возвратимся к более общему случаю, когда значность переменной Y не специфицирована. Для отображения сути изучаемого феномена необходимо исключить вырожденные случаи выполнения паттерна $\text{Ind}(X, Z) \& \text{Ind}(X, Z | Y)$ и формулы (5). В частности, целесообразно исключить случаи, когда одна из переменных — изолирована. Если верно $\text{Ind}(X, Y)$, т.е. $p(Y | X) = p(Y)$, то (5) редуцируется до формулы $p(Y | XZ) = p(Y | Z)$. Поэтому следует сосредоточить анализ на ситуации, которую можно назвать парадоксальным паттерном зависимостей:

$$\text{Ind}(X, Z) \& \text{Ind}(X, Z | Y) \& p(Y | XZ) \neq p(Y | Z) \& p(Y | XZ) \neq p(Y | X).$$

В принципе это описание совместимо с двумя разными типами структуры модели — и с цепью (см. рис. 2), и с коллаидером (см. рис. 1). Совместное распределение $p(XYZ)$ с такими свойствами можно получить, подобрав соответствующие параметры. При этом в структуре типа цепь возникнет аномалия в виде коллапса транзита (см. пример 1), а в структуре типа коллаيدر аномалия будет проявляться в виде отсутствия провоцированной зависимости. Если поставить целью дать описание коллайдера, не допускающее эквивалентного представления в виде цепи, то придется исключить особые случаи сочетаний параметров, которые порождают свойства, не характерные (аномальные) для коллаидерной структуры. Тогда коллаидерная модель должна характеризоваться помимо независимости $\text{Ind}(X, Z)$ также свойством, что вероятностное описание зависимой переменной Y невозможно декомпозировать по факторам X и Z , т.е. описание $p(Y | XZ)$ нельзя представить как композицию некоторых функций от $p(Y | X)$ и $p(Y | Z)$. Необходимость трехмерных распределений вероятности для описания условного распределения зависимой переменной свидетельствует о взаимодействии факторов. Взаимодействие факторов характеризует невырожденный коллаيدر и имплицитно подразумевает присутствие провоцированной зависимости. Итак, невырожденный коллаيدر характеризуется взаимодействием двух факторов при влиянии на совместный эффект. Тогда $p(Y | XZ)$ нельзя выразить формулой (5).

Различие между указанными альтернативными структурами модели состоит также в том, что в невырожденной цепи всегда подразумевается $\text{Dep}(X, Y)$ и $\text{Dep}(Z, Y)$, в то время как для коллайдера эти зависимости необязательные (несущественные). Рассмотрим пример.

Пример 2. Модель имеет структуру коллайдера $X \rightarrow Y \leftarrow Z$, где все переменные бинарные, распределение $p(X)$ — произвольное, а остальные параметры определены так: $p(Z) = (2/7; 5/7)$; $p(y_1 | x_1, z_1) = 0,3$; $p(y_1 | x_2, z_1) = 0,8$; $p(y_1 | x_1, z_2) = 0,4$; $p(y_1 | x_2, z_2) = 0,2$.

Легко убедиться, что в этой модели выполняется $\text{Ind}(X, Y)$. Таким образом, одна из дуг коллайдера не идентифицируется изолированно, вне контекста. Тем

не менее, дуга $X \rightarrow Y$ существует и необходима для описания модели. Назовем такую дугу (связь) маскированной.

В невырожденном коллаидере обе дуги одновременно могут быть маскированными, что иллюстрируется следующим примером.

Пример 3. Имеем модель со структурой $X \rightarrow Y \leftarrow Z$ и с параметрами из примера 2, но с заданным распределением $p(X) = (6/7; 1/7)$. Тогда будет $\text{Ind}(X, Y)$ и $\text{Ind}(Z, Y)$.

Информационную меру провоцированной зависимости в коллаидере $X \rightarrow Y \leftarrow Z$ естественно определить как $\text{Info}(X, Z | Y)$. Выразим взаимную информацию через энтропию $H(\cdot)$ (по определению):

$$\text{Info}(X, Z | Y) = H(X | Y) + H(Z | Y) - H(XZ | Y). \quad (10)$$

Энтропию $H(XZ | Y)$ можно выразить как

$$H(XZ | Y) = H(XYZ) - H(Y) = H(Y | XZ) + H(XZ) - H(Y). \quad (11)$$

Подставляя (11) в (10), получаем

$$\text{Info}(X, Z | Y) = H(X | Y) + H(Z | Y) - H(Y | XZ) - H(XZ) + H(Y). \quad (12)$$

Пусть теперь имеем $\text{Ind}(X, Z)$, т.е. $H(XZ) = H(X) + H(Z)$. Тогда (12) принимает вид

$$\text{Info}(X, Z | Y) = H(Y) - H(Y | XZ) - H(X) + H(X | Y) - H(Z) + H(Z | Y). \quad (13)$$

Термы этой формулы можно преобразовать, используя эквивалентную форму взаимной информации $\text{Info}(Q, W) = H(Q) - H(Q | W)$. Тогда формула (13) принимает вид

$$\text{Info}(X, Z | Y) = \text{Info}(Y, XZ) - \text{Info}(X, Y) - \text{Info}(Z, Y). \quad (14)$$

Эта формула предполагает только факт $\text{Ind}(X, Z)$. Величину $\text{Info}(Y, XZ)$ естественно интерпретировать как информационную меру «коллаидерной зависимости» для $X \rightarrow Y \leftarrow Z$. Иными словами, $\text{Info}(Y, XZ)$ — это «семейная» информация, естественно подразумеваемая под семей совокупность зависимой переменной и двух ее аргументов («родителей»). Термы $\text{Info}(X, Y)$ и $\text{Info}(Z, Y)$ — это частные (реберные) зависимости, или «проекции» коллаидерной зависимости на соответствующий фактор.

Формула (14) выявляет связь провоцированной зависимости с взаимодействием факторов. Она дает полный «объем» взаимодействия X и Z с точки зрения их влияния на Y . Провоцированная зависимость коллапсирует в коллаидере только тогда, когда семейная информация равна сумме «однофакторных» зависимостей, т.е. в случае отсутствия взаимодействия факторов.

Ранее было показано, что дуга коллаидера может быть маскированной (когда $\text{Info}(X, Y) = 0$ или $\text{Info}(Z, Y) = 0$). Если обе дуги коллаидера маскированы, то получим $\text{Info}(X, Z | Y) = \text{Info}(Y, XZ)$, т.е. провоцированная зависимость достигает «наибольшего» значения, равного семейной информации. В частности, для примера 3 получаем $\text{Info}(X, Z | Y) = \text{Info}(Y, XZ) \approx 0,0266$.

Сравнение величины провоцированной зависимости с «реберной» зависимостью иногда целесообразно, однако в случае маскированных дуг теряет смысл. Для определения, насколько велика провоцированная зависимость, целесообразно сравнить ее с полной зависимостью от двух факторов. Предлагаем следующий индекс взаимодействия двух факторов:

$$\lambda = \text{Info}(X, Z | Y) / \text{Info}(Y, XZ).$$

Величина λ варьирует между 0 и 1 и может служить показателем силы взаимодействия факторов X и Z . Другим показателем силы взаимодействия факторов может служить величина $\kappa = \text{Info}(X, Z | Y) / (\text{Info}(X, Y) + \text{Info}(Z, Y))$.

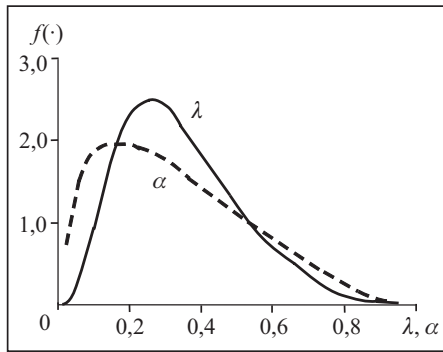


Рис. 4. Оценка распределения λ и α в коллаидере с трехзначными переменными

Для количественного анализа дискретного коллаидера была использована стохастическая симуляция. Случайно сгенерировано по 100000 случаев параметризации коллаидера с бинарными и трехзначными переменными. Для переменных X и Z генерировались безусловные распределения вероятности, для зависимой Y — таблицы условных распределений. Каждая строка таблицы распределения заполнялась независимо; компоненты строки генерировались последовательно, используя равномерное распределение. Первая компонента бралась из распределения

в интервале $[0, 1)$. Последующие компоненты брались из интервалов, размер которых определялся «остатком» (единица минус сумма сгенерированных компонент). Затем выполнялась пертурбация (компоненты случайно переставлялись.) Исключались вырожденные случаи (когда компонента вероятности близка к единице или семейная информация близка к нулю). Оказалось, что величина провоцированной зависимости сильно варьирует, но в большинстве случаев остается довольно высокой. На рис. 4 показана сглаженная эмпирическая плотность распределения величины λ и «удельного веса ребра» $\alpha = \text{Info}(X, Y) / \text{Info}(Y, XZ)$ для моделей с трехзначными переменными. Среднее значение λ равно 0,29 и 0,35 соответственно для бинарных и трехзначных моделей. Провоцированная зависимость составляет около 1/3 полной (семейной) информации коллаидера. В представленной таблице даны вес ребра $M[E1]$ и вес провоцированной зависимости $M[Pvk]$, а также частоты выполнения соответствующих соотношений. В таблице обозначено $Pvk \equiv \text{Info}(X, Z | Y)$, $E1 \equiv \text{Info}(X, Y)$, $E2 \equiv \text{Info}(Z, Y)$.

В трехзначных моделях максимум величины λ составил 0,972, а в бинарных моделях — 0,9998993. Ясно, что большие значения λ достигаются, когда ребра почти маскированы. В одном случае провоцированная зависимость оказалась приблизительно в 10000 раз больше суммы весов обоих ребер.

Таблица 1

| Переменные | Средняя величина | | Частота выполнения соотношений, % | | | |
|-------------|------------------|----------|-----------------------------------|------------------------|-------------------|---------------------------|
| | $M[E1]$ | $M[Pvk]$ | $Pvk < \min\{E1, E2\}$ | $Pvk > \max\{E1, E2\}$ | $Pvk > (E1 + E2)$ | $Pvk > 5 \cdot (E1 + E2)$ |
| Бинарные | 0,0559 | 0,0472 | 34,9 | 23 | 19 | 4,5 |
| Трехзначные | 0,0838 | 0,0890 | 22 | 33 | 18 | 0,4 |

4. МОДЕЛИ С ДЕЙСТВИТЕЛЬНЫМИ ПЕРЕМЕННЫМИ

Рассмотрим сначала модели с линейными зависимостями. В этом случае зависимость можно измерять коэффициентом корреляции Пирсона, а условную зависимость — частной корреляцией. Коэффициент корреляции переменных, «крайних» в структуре цепи, равен произведению коэффициентов корреляции для всех отрезков цепи. (Имеются в виду теоретические значения.) Следовательно, коллапс транзитной зависимости в цепи невозможен.

В коллаидере (см. рис. 1, а) с линейными зависимостями провоцированная зависимость выражается как $r_{XZ,Y} = (r_{XZ} - r_{XY} \cdot r_{YZ}) / ((1 - r_{XY}^2)(1 - r_{YZ}^2))^{1/2}$. Тогда из фактов $r_{XZ} = 0$, $r_{XY} \neq 0$ и $r_{YZ} \neq 0$ вытекает $r_{XZ,Y} \neq 0$. Более того, из указанной фор-

мулы можно видеть, что провоцированная зависимость по абсолютной величине больше, чем произведение реберных зависимостей. Рассмотрим простой случай.

Пример 4. Модель имеет структуру коллайдера и описывается уравнениями $X = \varepsilon_1$, $Z = \varepsilon_2$, $Y = X + Z + \varepsilon_3$, где $\varepsilon_1, \varepsilon_2, \varepsilon_3$ — нормальны и взаимно независимы, причем дисперсия величин ε_1 и ε_2 равна σ_0^2 , а дисперсия ε_3 равна σ_3^2 .

Для этой модели получим $r_{XY} = r_{YZ} = (2 + (\sigma_3^2 / \sigma_0^2))^{-1/2}$; $r_{XZ.Y} = (-\sigma_0^2) / (\sigma_0^2 + \sigma_3^2)$. Если положить $\sigma_0 = \sigma_3$, то $r_{XY} = r_{YZ} = 1/\sqrt{3}$; $r_{XZ.Y} = -1/2$. В результате получаем отношение величины провоцированной зависимости к «реберной» зависимости $|r_{XZ.Y} / r_{XY}| = \sqrt{3}/2 \approx 0,87$. В крайнем случае, когда дисперсия σ_3^2 равна нулю (и зависимость становится детерминистической), получим $|r_{XZ.Y} / r_{XY}| = \sqrt{2} \approx 1,4$.

Поскольку реберная зависимость в линейной модели измеряется корреляцией r_{XY} , то в качестве меры коллайдера (семейной) зависимости вместо $\text{Info}(Y, XZ)$ естественно использовать множественный коэффициент корреляции R_* (корень из коэффициента детерминации). Имеем $R_{Y(X,Z)}^2 = 1 - \sigma_{Y|XZ}^2 / \sigma_Y^2$, где $\sigma_{Y|XZ}^2$ — дисперсия остатков. Силу взаимодействия факторов $\lambda^{(L)}$ в линейной модели определим (по аналогии с λ) как $\lambda^{(L)} = r_{XZ.Y} / R_{Y(X,Z)}$.

Для примера 4 будет $R_{Y(X,Z)}^2 = 1 - \sigma_3^2 / (2\sigma_0^2 + \sigma_3^2)$. Для случая $\sigma_0 = \sigma_3$ получим $R_{Y(X,Z)}^2 = 2/3$. Тогда $\lambda^{(L)} = -\sqrt{3} / (2\sqrt{2}) \approx -0,61$. В линейной модели каждый фактор вносит свой индивидуальный вклад в эффект, тем не менее, как видим, суммирование, присходящее в линейной модели, также порождает взаимодействие факторов.

В специальной литературе понятие взаимодействия часто трактуется более узко, чем определено выше. Взаимодействием называют явно выраженный «синергизм» факторов. Предельный случай синергизма можно найти в зависимостях балансового типа таких, например, как исключяющее «ИЛИ» с бинарными переменными.

В общем случае анализ зависимости действительных переменных намного сложнее. Заметим, что вышеприведенная характеристика провоцированной зависимости в терминах взаимной информации сохраняется и для действительных переменных. Но необходимо иметь в виду, что тогда используется дифференциальная энтропия, которая может быть отрицательной и бесконечной в особых случаях. Проанализируем транзитную зависимость в цепи. Для ясности начнем с детерминистических зависимостей. Поскольку суперпозиция функций есть функция, цепь невырожденных зависимостей дает невырожденную «транзитную функцию». Тем не менее, попытаемся сконструировать параметризацию цепи с коллапсом транзитной зависимости. Отталкиваемся от варианта цепи вида $X \leftarrow Y \rightarrow Z$, где $X = g(Y)$ и $Z = h(Y)$ — невырожденные функции, а Y варьирует в некотором диапазоне (a, b) . Чтобы добиться независимости между X и Z (нивелировать транзитную зависимость), наложим на функции $g(Y)$ и $h(Y)$ следующие требования. Везде, где $\frac{dg}{dy} \neq 0$, требуем $\frac{dh}{dy} = 0$, и наоборот. Неформально говоря, там, где Y влияет на X ,

переменная Z не должна изменяться (локальная константа). Но этого недостаточно. Чтобы добиться равенства $p(Z | X) = p(Z)$, необходимо наложить на функции $g(Y)$ и $h(Y)$ дополнительные требования: на всех интервалах Y , где функция $h(Y)$ не постоянна, функция $g(Y)$ должна быть не только постоянной, но и принимать одно и то же значение. И симметрично: там, где $g(Y)$ изменяется, $h(Y) = \text{const}$. Локальные константы каждой функции должны быть идентичны на соответствующих интерва-

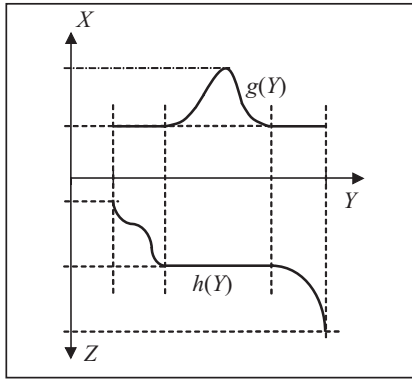


Рис. 5. Иллюстрация к построению цепи функций

лах. Пара функций, удовлетворяющих названным требованиям, показана на рис. 5 (графики функций $X = g(Y)$ и $Z = h(Y)$ совмещены по оси Y). Чтобы трансформировать такую модель к структуре цепи вида $X \rightarrow Y \rightarrow Z$, требуется найти функцию, обратную для $g(Y)$. Но для однозначности функции $Y = g^{-1}(X)$ необходима монотонность $g(Y)$. При указанных ограничениях добиться монотонности можно только в случае, если диапазон изменения Y распадается на два интервала (а не три, как на рис. 5), на одном из которых функция $g(Y) = \text{const}$.

На рис. 6, *a* изображены функции $X = g(Y)$ и $Z = h(Y)$, которые удовлетворяют всем названным требованиям. На рис. 6, *б* отображена карта локализации совместной плотности $f(xz)$ для сконструированной модели. Место, где плотность $f(xz)$ не равна нулю, концентрируется на линии вдоль левой и нижней границы квадрата; значит, совместная плотность $f(xz)$ вырождена. Совместная плотность $f(xz)$ равна нулю почти на всей плоскости значений переменных. Значит, мы не добились независимости X и Z . При изменении переменной X в интервале $(t, u]$ распределение переменной Z не изменяется. И симметрично: при изменении переменной Z в интервале $(m, l]$ распределение X не изменяется. Когда X достигает значения t , происходит скачкообразное изменение распределения переменной Z . Аналогичный скачок происходит при достижении m .

Добавив к переменным X и Z независимый шум, равномерно распределенный в интервалах $[t, u]$ и $[m, l]$, можно заполнить указанный квадрат значений равномерной плотностью. Но и тогда по-прежнему не будет достигнута независимость X и Z , поскольку образуются «линии-выступы» ненулевой плотности $f(xz)$ на продолжении левой и нижней границы квадрата (рис. 6, *в*).

Наконец, предположим, что удалось построить желаемую модель со структурой $X \leftarrow Y \rightarrow Z$. Ввиду константности функции $g(Y)$ на соответствующем интервале попытка развернуть дугу $X \leftarrow Y$ и получить связь вида $X \rightarrow Y$ приводит к функции с бесконечной производной. В данном примере интервалы значений Y , где две функции константны, взаимно пересекаются на (c, d) . Поэтому при попытке репараметризовать эту модель для структуры

коллайдера $X \rightarrow Y \leftarrow Z$ функция $Y = \varphi(X, Z)$ будет неоднозначной. Для однозначности необходимо исключить пересечение интервалов, где функции $g(Y)$ и $h(Y)$ — константны, т.е. положить $c = d$.

Проведенный анализ позволяет предположить, что на цепи невырожденных функций с непрерывными переменными транзитная зависимость не может коллапсировать. (При анализе данных вместо непрерывности пере-

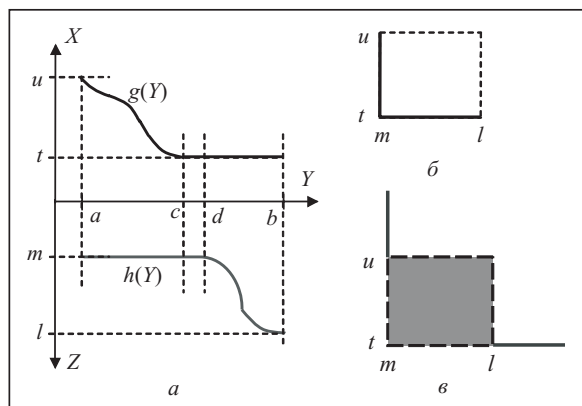


Рис. 6. Иллюстрация цепи функций: *a* — графики функций $g(Y)$, $h(Y)$; *б* — карта локализации плотности $f(x, z)$; *в* — плотность для модели с шумом

менных более уместно говорить о гладкости аппроксимированных плотностей.) Итак, можно заключить, что паттерн $\text{Dep}(X, Y) \& \text{Dep}(Z, Y) \& \text{Ind}(X, Z)$ не может возникнуть в цепи.

Феномен маскированных связей (дуг) в определенном смысле возможен и в коллайдере с действительными переменными (и нелинейными зависимостями). Проиллюстрируем это следующим примером (подразумеваем, что переменные имеют средние значения).

Пример 5. В коллайдере $X \rightarrow Y \leftarrow Z$ переменная Y определяется формулой $Y = X \cdot Z + \varepsilon$, где плотности распределения переменных X и Z — любые симметричные (вокруг нуля) функции, а плотность распределения ε — произвольная.

Поскольку зависимость в этой модели — нелинейная, для анализа реберных зависимостей обычная корреляция непригодна. Возьмем для этого корреляционное отношение η^2 (которое показывает, насколько дисперсия средних условных значений зависимой переменной приближается к полной дисперсии). В примере 5 каждая переменная X и Z в отдельности не влияет на среднее значение Y . Взяв в качестве условия любой набор значений фактора, получим, что теоретические значения корреляционного отношения $\eta_{Y(X)}^2$ и $\eta_{Y(Z)}^2$ равны нулю. Это дает повод заключить, что в рассмотренной модели обе дуги коллайдера — маскированные. В действительности это обманчивый результат, вызванный неадекватностью корреляционного отношения как средства оценивания зависимости в данной ситуации. В примере 5 каждый фактор в отдельности влияет только на величину дисперсии Y .

Количественный анализ выполнен с помощью симуляции. Модель уточнена: X , Z и ε имеют стандартное нормальное распределение. Для измерения реберных зависимостей используется корреляционное отношение, причем в качестве аргумента берется переменная Y . Из модели генерируем выборку данных и разбиваем ее на страты (10–12 страт) по значению переменной Y . Использовались выборки данных объемом 125000 записей. Для зависимости X от Y величина $\eta_{X(Y)}^2$ оказалась меньше значения 0,0001. Значение $\eta_{Z(Y)}^2$ также близко к нулю. Имеем иллюзию маскированных дуг. Как контраст, провоцированную зависимость можно измерять на основе линейной корреляции. А именно, мы оценили провоцированную зависимость с помощью модуля условной (по Y) корреляции X и Z . Модули корреляции усреднялись по всем стратам выборки (с весами). Условная корреляция X и Z составила около 0,37.

5. РОЛЬ ПРОВОЦИРОВАННОЙ ЗАВИСИМОСТИ В ИДЕНТИФИКАЦИИ СТРУКТУРЫ МОДЕЛИ

Техника провоцирования зависимости полезна для идентификации ориентаций (направленности) связей. Рассмотрим известное (стандартное) правило ориентации ребер, имплементированное, в частности, в алгоритме РС [6]. Алгоритм принимает решение ориентировать концы дуг в виде $X \rightarrow Y \leftarrow Z$ в ситуации, когда выведена пара соседних ребер $X-Y-Z$ и отсутствует ребро $X-Z$, причем отсутствие этого ребра идентифицировано на основании факта независимости $\text{Pr}(X; \mathbf{S}; Z)$, где $Y \notin \mathbf{S}$. Практическая ненадежность такого правила ориентации ребер объясняется тем, что в процессе вывода моделей из реальных выборок данных статистические тесты вынуждены не отвергать гипотезу независимости в ситуации, когда зависимость довольно слабая. (Слабая зависимость может быть проявлением выборочного шума.) Таким образом, фрагмент генеративной модели может иметь структуру $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$ или $X \leftarrow Y \leftarrow Z$, причем транзитная зависимость между X и Z — слабая, и тогда тест трактует ее как шум.

Для упрощения анализа далее рассмотрим частный случай, когда $S = \emptyset$. Для повышения надежности процедуры ориентации связей (ребер) прибегнем к помощи провоцированной зависимости. Действительно, в цепи $X-Y-Z$ кондиционирование переменной Y устраняет транзитную зависимость. Наоборот, в случае коллайдера $X \rightarrow Y \leftarrow Z$ кондиционирование переменной Y создает провоцированную зависимость между X и Z . Таким образом, альтернативные модели ведут себя по-разному. Критическим вопросом для различения моделей становится величина провоцированной зависимости.

Идея привлечения провоцированной зависимости для идентификации ориентаций связей появилась давно. В явном виде использование провоцированной зависимости для повышения надежности ориентации предлагалось в [8, 13, 14]. Более полно эта техника описана и оценена в [15]. В раннем алгоритме SGS [6] для ориентации фрагмента модели $X-Y-Z$ вызывается тест в формате $\text{Ind}(X; S \cup \{Y\}; Z)$.

Попытаемся количественно оценить эффективность техники провоцированной зависимости как средства дискриминации альтернативных структур модели. Будем вычислять характеристики непосредственно из параметров генеративной модели, что соответствует асимптотическим оценкам для большой выборки. По-видимому, наименьший эффект можно ожидать в случае линейных моделей. (В линейных моделях со структурой коллайдера факторы суммируются, а суммирование относят к слабому взаимодействию.) Для наглядности анализа примем следующую упрощенную технику верификации независимости. Определим единый для всех тестов порог t значимости зависимости. Безусловную независимость переменных X и Z не отвергаем в случае $r_{XZ} < t$, а условную — в случае $r_{XZ.Y} < t$.

Пусть имеется линейная модель со структурой цепи (см. рис. 1, а). Тогда если $r_{XZ} < t$, «стандартное» правило ориентации ошибочно идентифицирует коллаيدر. Без потери общности можно считать все коэффициенты линейной связи положительными. Вероятность ошибки определяется как частота случаев $(r_{XY} > t) \& (r_{YZ} > t) \& (r_{XZ} < t)$ среди всех случаев $(r_{XY} > t) \& (r_{YZ} > t)$. Площадь «области ошибок» определяется формулой $-t(1-t) + \int_t^1 (t/u) du$. Площадь всех рассматриваемых случаев равна $(1-t)^2$. Величину порога t значимости берем от 0,005 до 0,1. Для простоты предположим, что множество анализируемых вариантов моделей генерируется так, что значения r_{XY} и r_{YZ} берутся независимо из равномерного распределения в интервале (0, 1). Тогда вероятность ошибки изменяется от 2 % до 17 %. (При $t = 0,05$ вероятность ошибки составляет 11 %.)

«Усиленное» правило идентификации ориентаций связей (с подтверждением в виде значимости провоцированной зависимости) сужает область ошибок. Тогда для линейной модели со структурой цепи ошибка усиленного правила ориентации исключается, так как $r_{XZ.Y} = 0$. Однако появляется другой вид ошибок — невыполнение ориентации связей в случае, когда генеративная модель является коллаидером $X \rightarrow Y \leftarrow Z$ и при этом величина провоцированной зависимости ниже порога значимости t . Ошибки происходят, когда выполняются соотношения $(r_{XY} > t) \& (r_{YZ} > t) \& (r_{XZ.Y} < t)$. Граница областей идентификации/неидентификации коллаيدرной структуры определяется условием $r_{XZ.Y} = t$. Тогда, используя формулу частной корреляции, получаем формулу, описывающую эту границу: $r_{YZ} = t((1-r_{XY}^2)/(t^2 + r_{XY}^2 - t^2 \cdot r_{XY}^2))^{1/2}$. При $t = 0,05$ вероятность того, что коллаيدرные ребра не будут ориентированы, равна 9 %.

Выше отмечено, что в линейном коллаидере провоцированная зависимость больше (по абсолютной величине), чем произведение реберных зависимостей. Поэтому вероятность того, что ребра коллайдера не ориентируются усиленным

правилom, меньше, чем вероятность ошибочной ориентации ребер цепи стандартным правилom.

Оценим ожидаемую эффективность усиленного правила ориентации связей по отношению к дискретным моделям. Сохраним «упрощенную» методику анализа и решений, но используем синтетические данные имитации. (Генерация параметров описана в разд. 3.) Коллайдерная ориентация ребер подтверждается, если провоцирующая зависимость превышает порог значимости. В моделях с бинарными переменными усиленное правило ориентировало 51 % коллайдерных ребер, если включать случаи с «почти маскированными ребрами», и 52 % — если не включать эти случаи. Для коллайдеров с трехзначными переменными эффективность ориентации составила 77 % и 83 % соответственно.

В дискретных моделях транзитная зависимость в цепи может сильно варьировать относительно веса ребер, образующих цепь (в частности, может коллапсировать). Поэтому ошибочная идентификация коллайдера стандартным правилom ориентации (когда генеративная модель является цепью) происходит чаще. Согласно результатам симуляции моделей с бинарными переменными риск таких ошибок составил 50–70 %; в моделях с трехзначными переменными риск ошибок — 25–75 % (в зависимости от порога значимости).

Таким образом, ужесточение процедуры идентификации направления связей (ориентация ребер только при условии подтверждения в виде провоцирующей зависимости) приводит к двум эффектам. С одной стороны, некоторые коллайдеры остаются неориентированными (ослабление эффективности). С другой стороны, предотвращаются ошибочные ориентации, которые потенциально могут породить ложные практические выводы. Изложенные выше эмпирические результаты дают основания надеяться, что частота «недо-идентификации» модели будет меньше, чем частота «ложной» идентификации. Принято считать, что лучше воздержаться от необоснованных выводов, чем ошибиться. Привлечение провоцирующей зависимости в процедуру идентификации ориентаций связей следует признать целесообразным.

По-видимому, еще более полезным направлением применения техники провоцирования зависимости будет обнаружение маскированных ребер коллайдера в дискретных моделях. Практически проблема возникает в ситуации почти маскированных ребер, когда вес ребра оказывается статистически незначимым (хотя и не нулевым). Результаты симуляции для коллайдера с дискретными переменными показали, что «почти маскированные» ребра типичны. Такие ребра фиксировались с помощью двух методик: 1) использовалось фиксированное абсолютное значение порога t для величин $\text{Info}(X, Y)$ и $\text{Info}(Z, Y)$; 2) использовался «плавающий» порог с относительной величиной, равной 10 % семейной информации, т.е. $t = \text{Info}(Y, XZ) / 10$. При использовании фиксированного порога число почти маскированных ребер составило 44 % для бинарных переменных и 33 % для трехзначных переменных. При использовании «плавающего» порога число таких ребер составило 27 % для бинарных переменных и 13 % для трехзначных переменных.

Маскированное и почти маскированное ребра считаются обнаруженными (демаскированными), если провоцирующая зависимость превысила установленный порог значимости. Согласно результатам экспериментов при использовании фиксированного порога восстанавливаются 56 % и 75 % (почти) маскированных ребер соответственно для бинарных и трехзначных переменных. При использовании плавающего порога успешность восстановления ребер составила 87 % и 97 % соответственно. В тех случаях, когда оба ребра коллайдера оказались (почти) маскированными, все они (100 %) восстанавливаются. Это понятно, поскольку в таких случаях величина провоцирующей зависимости приближается к семейной информации согласно формуле (14). Таким образом, показана новая роль провоцирующей зависимости — выявление маскированных связей мо-

дели. (Заметим, что высокую оценку частоты маскированных ребер нельзя экстраполировать на общий случай, т.е. на коллаидеры в контексте.)

На первый взгляд, могло бы показаться, что если переменная Y безусловно не зависит от переменной X , то кондиционирование Y не должно влиять на связи X с третьими переменными. Ошибочность такой логики заключается в подмене понятия «абсолютно не зависит» понятием «безусловно не зависит». Если имеем три попарно независимые переменные и в результате кондиционирования одной из них две другие становятся зависимыми, можно утверждать, что в модели существует не менее двух связей.

ЗАКЛЮЧЕНИЕ

Согласно полученным результатам для идентификации модели полезно не только выявление условных независимостей (т.е. блокирование зависимости), но и провоцирование зависимости (т.е. разрушение независимости). Наведение (провоцирование) зависимости эффективно и для ориентации, и для демаскирования статистических связей. Провоцированная зависимость в коллаидере также предоставляет принципиальное средство измерения силы взаимодействия факторов.

СПИСОК ЛИТЕРАТУРЫ

1. Berkson J. Limitations of the application of fourfold table analysis to hospital data // *Biometrics Bulletin*. — 1946. — 2, N 3. — P. 47–53.
2. Heckman J.J. Sample selection bias as a specification error // *Econometrica*. — 1979. — 47, N 1. — P. 153–161.
3. Elwert F., Winship Ch. Endogenous selection bias: The problem of conditioning on a collider variable // *Annual Review of Sociology*. — 2014. — 40. — P. 31–53.
4. VanderWeele T.J., Robins J.M. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect // *Amer. Journ. Epidemiol.* — 2008. — 166, N 9. — P. 1096–1104.
5. Bareinboim E., Tian J., Pearl J. Recovering from selection bias in causal and statistical inference // *Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014*. — P. 2419–2416 (July 27–31, 2014, Québec Convention Center, Québec City, Québec, Canada).
6. Spirtes P., Glymour C., Scheines R. *Causation, prediction and search*. — New York: MIT Press, 2001. — 543 p.
7. de Campos L.M., Huete J.F. On the use of independence relationships for learning simplified belief networks // *Intern. Journal of Intelligent Systems*. — 1997. — 12, N 7. — P. 495–522.
8. Балабанов А.С. К выводу структур вероятностных зависимостей из статистических данных // *Кибернетика и системный анализ*. — 2005. — № 6. — С. 19–31.
9. Балабанов А.С. Реконструкция модели вероятностных зависимостей по статистическим данным. Инструментарий и алгоритм // *Проблемы управления и информатики*. — 2009. — № 6. — С. 90–103.
10. Schweizer B., Wolff E.F. On nonparametric measures of dependence for random variables // *The Annals of Statistics*. — 1981. — 9, N 4. — P. 879–885.
11. Lehmann E.L. Some concepts of dependence // *The Annals of Mathematical Statistics*. — 1966. — 37, N 5. — P. 1137–1153.
12. Андон П.І., Балабанов О.С. До відкриття латентного бінарного фактора в статистичних даних категорного типу // *Доповіді НАН України*. — 2008. — № 9. — С. 37–43.
13. Bada L. Determining the direction of causal influence in large probabilistic networks: A constraint-based approach // *Proc. of the 16th European Conf. on Artificial Intelligence (ECAI-2004), Valencia, Spain, 2004*. — Amsterdam: IOS Press, 2004. — P. 263–267.
14. Kuroki M., Kikuchi T., Miyakawa M. The graphical condition for identifying arrows in recovering causal structure // *Journal of Japan Statistical Society*. — 2001. — 31, N 2. — P. 175–185.
15. Ramsey J., Spirtes P., Zhang J. Adjacency-faithfulness and conservative causal inference // *Proc. of the 22nd Conf. on Uncertainty in Artificial Intelligence*. — Oregon: AUAI Press, 2006. — P. 401–408.

Поступила 28.05.2015