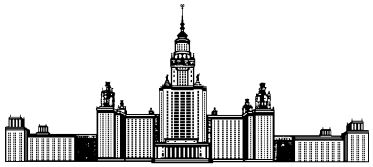
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА

«Исследование различных методов верификации моделей вероятностных распределений, основанных на байесовских сетях»

Выполнил:

Новиков Павел Александрович

Научный руководитель:

д.ф.-м.н.

Сенько Олег Валентинович

Москва, 2015

Содержание

1 Введение

Многие области человеческого знания являются плохо формализуемыми и не допускают на сегодняшний день прямого моделирования. Часто для таких областей недоступными являются также большие массивы однородных данных для анализа. Одной из таких областей является медицина. Аналогичные проблемы встают при изучении динамики популяции развивающихся стран, эпидемиологии, исследовании организованной преступности, экологии[14]. Названные ограничения делают неизбежным широкое использование экспертных знаний при решении любых задач.

В медицине можно выделить три большие группы методов, используемых для помощи принятия решений экспертами: методы на основе наборов решающих правил, часто с использованием аппаратов нечетких множеств, методы, основанные на корреляционных зависимостях(включая методы, использующие нейронные сети), и вероятностные методы с использованием байесовких сетей ([18], см. Рис. 1).

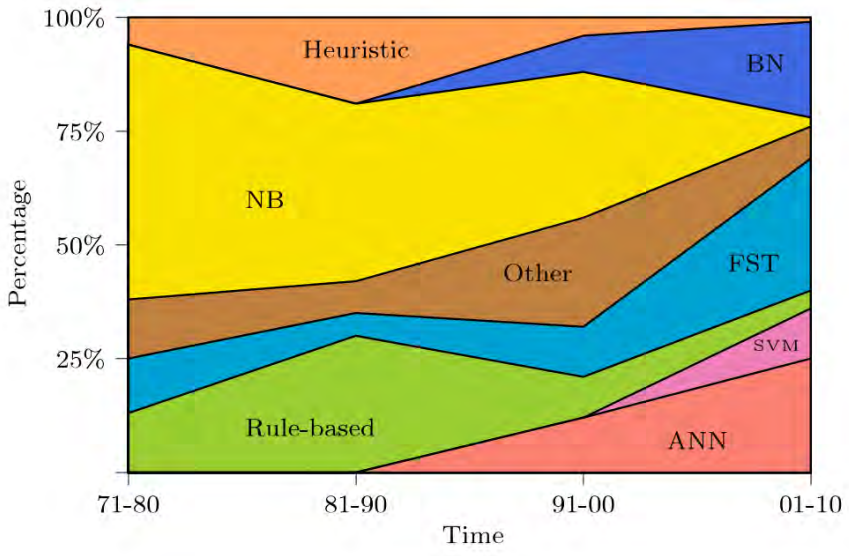


Рис. 1: График популярности методов на основе анализа статей из [18]. ANN - искусственные нейронные сети, SVM - метод опорных векторов, Rule-based - методы на основе решающих правил, FST - методы на основе нечетких множеств, NB - наивные байесовские методы, BN - байесовские сети, Heuristic - эвристические методы, Other - прочие методы.

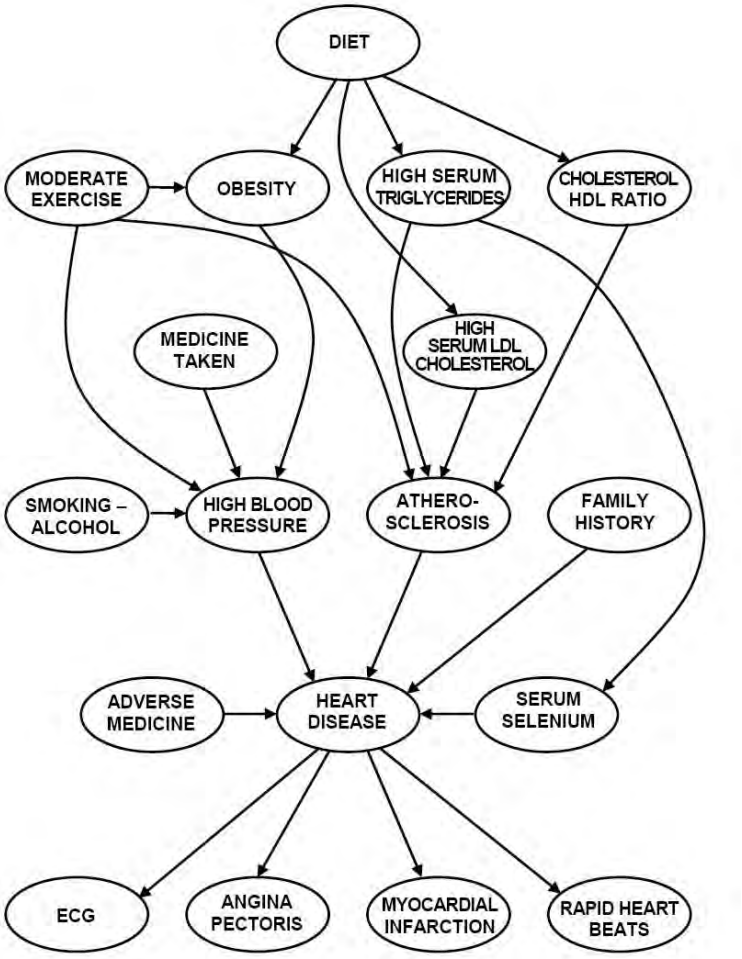


Рис. 2: Пример графа байесовской сети из [6]

Методы первых двух групп лучше согласуются с интуицией экспертов предметной области, однако в основе последней лежат более строгие математические модели - именно на этих методах мы фокусировались в данной работе.

В основе модели байесовских сетей лежит задание структуры зависимостей переменных многомерного распределения в виде графа (Рис. 2) и представление распределения в виде произведения условных распределений отдельных переменных при всевозможных фиксированных значения переменных-предков.

Теоретически этот метод позволяет моделировать произвольные многомерные вероятностные распределения, однако действительно эффективным он становится при наличии относительно разреженной структуры взаимосвязей между переменными.

Область применения байесовских сетей не ограничивается задачами, в которых доступно ограниченное количество данных. При наличии выборок достаточного размера ставятся задачи автоматического построения сетей по данным[17, 8, 16]. Эти задачи, однако, обладают большой вычислительной сложностью (часто NP-полнотой), поэтому используются эвристические алгоритмы.

Как построение сетей на основе экспертных оценок, так и использование эвристических алгоритмов даёт результаты, не позволяющие давать теоретических гарантий, что приводит к задаче валидации полученных результатов. При этом единого подхода к решению этой задачи не существует[14] - для отдельных задач строятся специфические меры достоверности, либо используются меры, основанные на правдоподобии. Последние позволяют эффективно сравнивать качество различных моделей, но не позволяют давать абсолютных оценок.

В данной работе исследуется возможность применения классических критериев согласия по подмножествам переменных для построения абсолютных оценок корректности вероятностных распределений, заданных байесовскими сетями.

1.1 План работы

В разделе 2 описываются известные понятия и результаты: понятие байесовских сетей и некоторые их свойства, методы исключения переменных и монте-карло построения маргинальных распределений, статистические тесты, включая задачу множественной проверки гипотез, случайные графы и генерация табличных распределений. Далее в разделе 3 излагаются детали предлагаемого метода. Затем в разделе 4 описывается алгоритм исключения переменных в байесовских сетях с сохранением свойств байесовской сети, и в разделе 5 приводятся результаты экспериментального исследования метода.

2 Теоретические основы

2.1 Понятие байесовских сетей

Рассмотрим многомерное распределение на множестве переменных $X = \{x_i\}, i = 1..n$. Количество различных комбинаций значений переменных растёт экспоненциально с ростом количества переменных, поэтому прямое задание всевозможных вероятностей может быть практически неосуществимым. Однако, если распределение обладает определённой структурой, оно может быть естественным образом представлено при помощи байесовской сети.

Для любого набора значений $\hat{X} = (\hat{x}_1, ..., \hat{x}_n)$ вероятность $P(X = \hat{X})$ может быть записана как произведение условных вероятностей: $P(x_1 = \hat{x}_1)P(x_2 = \hat{x}_2|x_1 = \hat{x}_1)P(x_3 = \hat{x}_3|x_1 = \hat{x}_1, x_2 = \hat{x}_2)...P(x_n = \hat{x}_n|x_1 = \hat{x}_1, x_2 = \hat{x}_2, ..., x_n = \hat{x}_n)$.

Символом $Pred(x_k)$ обозначим множество переменных $x_1, ...x_{k-1}$(заметим, что это множество зависит от фиксированного порядка переменных). Равенства выражений, в которых не указываются значения переменных будем понимать как выполненные для всевозможных комбинаций значений. Тогда равенство выше можно переписать как $P(X) = \prod_{i=1}^{n} P(x_i|Pred(x_i))$.

Часть переменных, входящих в запись некоторых множителей, может быть избыточной в том смысле, что существует такое подмножество $Par(x_k) \subset Pred(x_k)$, что $P(x_k|Pred(x_k)) = P(x_k|Par(x_k))$. Это позволяет уменьшить количество переменных, входящих в запись каждого из множителей. Если количество переменных велико, такое разрежение может сделать возможной на практике работу с распределением, простое представление которого иначе требовало бы недоступного количества ресурсов.

В теории байесовских сетей для представления обозначенной структуры используется ациклический ориентированный граф, каждая вершина которого соответствует одной из переменных, и в котором ребро (x,y) существует тогда и только тогда, когда $x \in Par(y)$.

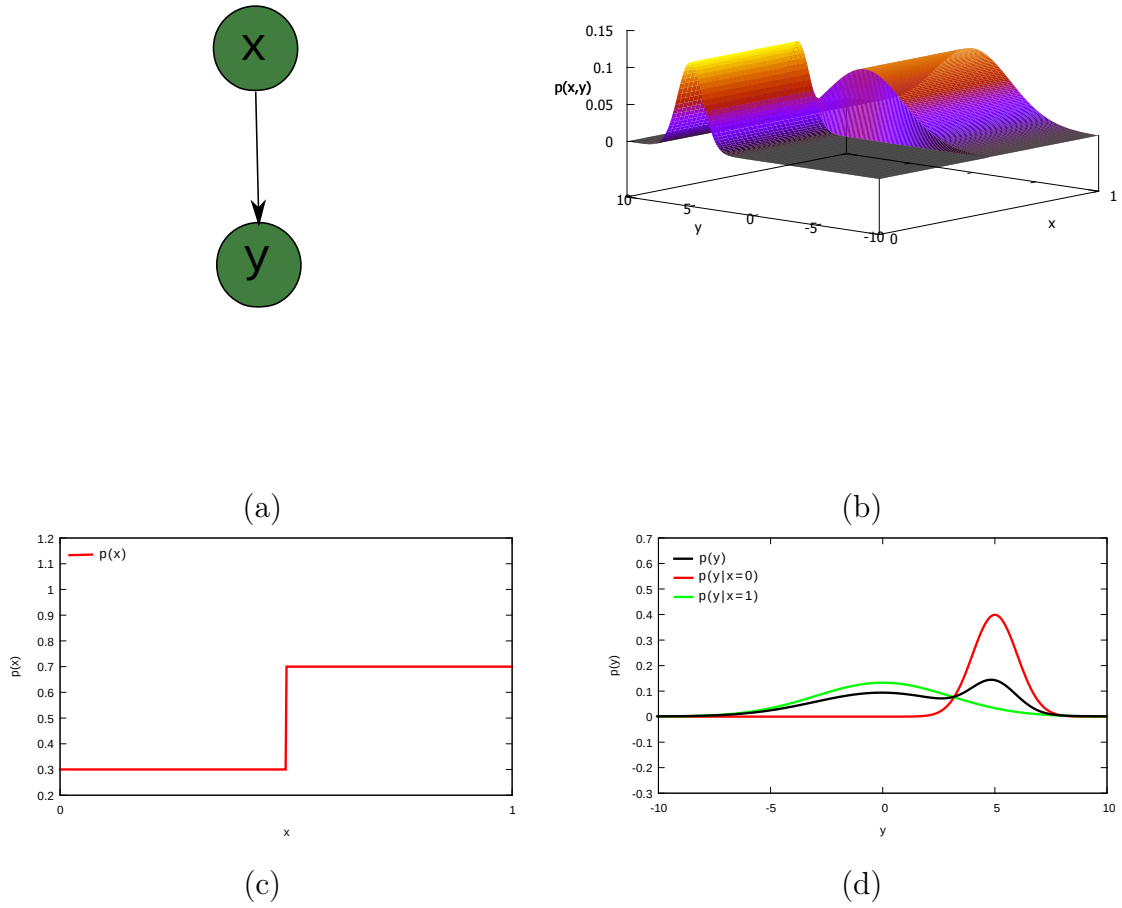


Рис. 3: Представление смеси нормальных распределений в виде байесовской сети. (a) - граф сети, (b) - совместное распределение, (c) - маргинальное распределение x, (d) - условные распределения $y|x$ и маргинальное распределение y.

Определение 1. *Байесовской сетью для n-мерного вероятностного распределения* $P(X)$ *называется ориентированный ациклический граф* $G(X, E)$ *вместе со множеством условных распределений* $P(x|Par(x))$. *Вершинами этого графа являются переменные* X *и* $(x_\alpha, x_\beta) \in E \Leftrightarrow x_\alpha \in Par(x_\beta)$.

Простыми примером байесовской сети является марковская цепь. Также можно представить в виде байесовской сети модель смеси распределений (Рис. 3).

Заметим, что вероятностному распределению может быть сопоставлено множество байесовских сетей (как минимум, одна для каждого фиксированного порядка

переменных; также добавление избыточных рёбер не влияет на корректность равенств). Более подробно с темой можно познакомиться в [9].

2.1.1 Независимость в графе

Будем рассматривать ориентированный граф без циклов $G(V, E)$ (формально, без установления связи с конкретными распределениями), и множество отображений $f : V \to \{o, u\}$(o - "наблюдаема"; u - "ненаблюдаема"). Будем использовать стандартное определение *пути* в графе - произвольная последовательность вершин, попарно соединённых рёбрами. Если для любой последовательной пары вершин пути (a, b) ребро направленно от a к b, то такой путь будем называть *направленным путём*. Будем также называть *путём из вершины* v_1 *в вершину* v_2 путь, первая вершина которого v_1 и последняя - v_2. Для фиксированного графа и отображения f обозначим символом $O_f(V)$ множество наблюдаемых вершин ($\{v \in V | f(v) = o\}$) и $U_f(V)$ множество ненаблюдаемых вершин ($V \setminus O_f(V)$). Символом $Desc(v)$ будем обозначать множество вершин-потомков вершины v (вершин v' для которых существует направленный путь из v в v').

Введём следующее

Определение 2. *Путь A из ненаблюдаемой вершины в ненаблюдаемую вершину называется* ***активным*** *(Рис. 4), если выполнены следующие два условия:*

1. *Для любой последовательной тройки вершин, образующих V-структуру вида $v_{i-1} \to v_i \leftarrow v_{i+1}$ центральная вершина v_i либо является наблюдаемой, либо имеет наблюдаемую вершину - потомка.*

2. *Ни одна вершина пути, не являющаяся центральной в V-структуре вида $v_{i-1} \to v_i \leftarrow v_{i+1}$ не является наблюдаемой.*

Определение 3. *Пусть заданы непересекающиеся подмножества X, Y, Z множества V, и функция f задана как индикатор принадлежности множеству Z: $f(v) = o \Leftrightarrow v \in Z$. Если ни для какой пары вершин $v_1 \in X, v_2 \in Y$ не существует активного пути, то говорят, что X и Y d-разделены при условии Z.*

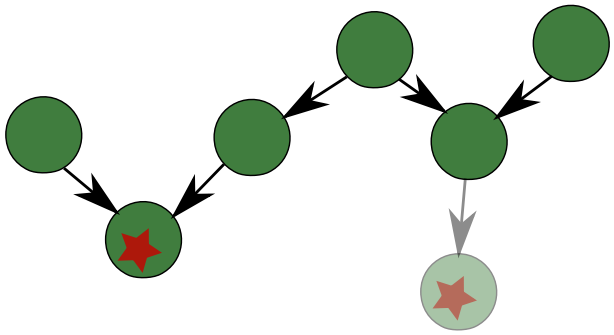


Рис. 4: Пример активного пути (ярким цветом обозначены вершины, входящие в путь, звездой - наблюдаемые вершины.

2.1.2 Факторизуемые распределения

Определение 4. *Если существует представление заданного распределения P(X) в виде байесовской сети с графом $G(X, E)$, то P(X)* ***факторизуемо*** *на графе $G(X, E)$.*

Выбор графа байесовской сети накладывает ограничение на множество факторизуемых распределений. Ключевым для описания этих ограничений является понятие условной независимости.

Пусть X_1, X_2, X_3 - непересекающиеся подмножества переменных X.

Определение 5. *X_1,X_2 условно независимы при условии X_3, если $P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3)$.*

Будем использовать для условной независимости X_1, X_2 при условии X_3 символ $X_1 \perp X_2|X_3$

Критерий факторизуемости даёт следующая

Теорема 1. *Распределение P(X) факторизуемо на графе $G(X, E)$ тогда и только тогда, когда для любых непересекающихся X_1,X_2,$X_3 \subset X$ из d-разделённости X_1 и X_2 при условии X_3 в G следует независимость X_1 и X_2 при условии X_3 в P.*

Доказательство этой теоремы можно найти в [9]. Проиллюстрировать же её можно следующими соображениями. Допустим, нам дана байесовская сеть, множество Z

наблюдаемых переменных и множества X, Y переменных, условную независимость которых требуется проверить. Несложно показать, что исключение путём суммирования (а также добавление) переменных, не входящих ни в одно из трёх заданных подмножеств, потомки которых также не входят в эти подмножества, никак не влияет на истинность утверждений об условной независимости, поэтому можно исключить их из рассмотрения.

Можно заметить, что в результате фиксации значения переменной x мы получаем новое факторизованное распределение, все множители которого, кроме $P(x|Par(x)$ и распределений дочерних переменных x остаются неизменными. Место $P(x|Par(x)$ занимает новый множитель, зависящий от $Par(x)$ (константу нормализации можно включить в этот множитель), а количество переменных, от которых зависят множители потомков уменьшается на единицу. Если все родительские вершины некоторой переменной y были зафиксированы, множитель, соответствующий y, не имеет общих существенных переменных со своими предками.

После этого произведём следующую процедуру группировки переменных. На каждом шаге будем рассматривать по одной вершине, не имеющей дочерних вершин, которые ещё не были рассмотрены . В группу этой переменной будем помещать все её ненаблюдаемые родительские вершины. В результате мы получим структуру, подобную изображенной на рис. 5.

Если для групп переменных X и Y в результате не найдётся последовательности попарно пересекающихся групп, соединяющих их, это будет означать, что распределение представимо в виде произведения двух множителей, один из которых не зависит от переменных X, а другой - от переменных Y. Можно показать, что из этого следует условная независимость. Если же такая последовательность найдётся, то можно построить последовательность путей с попарно общими последними вершинами, причём эти вершины принадлежат X, Y или Z. Можно показать, что из такой последовательности выделяется активный путь.

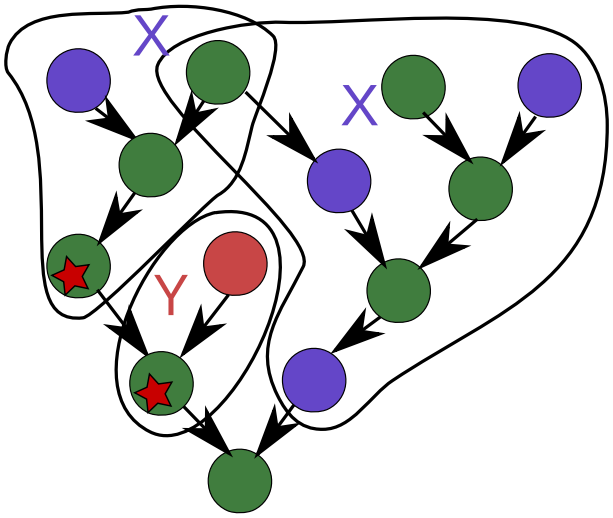


Рис. 5: Результат объединения переменных с общими существенными переменными.

2.2 Методы построения маргинальных распределений на основе байесовских сетей

Пусть нам дано многомерное распределение P на множестве переменных $X = \{x_i\}, i = 1..n$. Пусть также известна факторизация распределения, представимая в виде байесовской сети: $P(X) = \prod_{i=1}^{n} P(x_i|Par(x_i))$. Требуется найти маргинальное распределение $P(Y)$ для некоторого $Y \subset X$.

2.2.1 Последовательное исключение переменных

Введём обозначения. Символом $Child(y), y \in X$, обозначим множество дочерних вершин y. Символом $Child^*(y)$ обозначим $Child(y) \cup \{y\}$. Также введём следующее

Определение 6. *Марковским одеялом вершины y называется объединение множества родительских вершин y, множества дочерних вершин и множеств родительских вершин всех дочерних вершин (см. Рис. 6).*

Марковское одеяло для вершины $y \in X$ будем обозначать X^y.

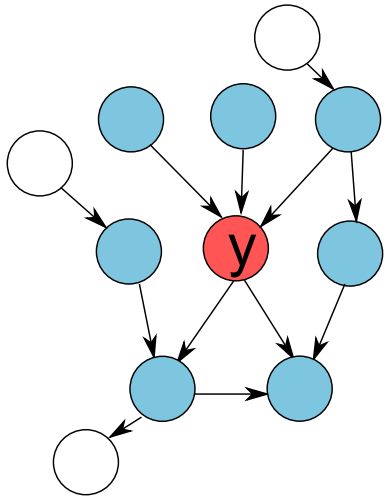


Рис. 6: Пример марковского одеяла переменной y. (Синие вершины).

$P(Y)$ можно получить из $P(X)$ непосредственно, просуммировав по всевозможным значениям каждой переменной $x \in X \setminus Y$:

$$P(Y) = \sum_{x_{r_1}} \sum_{x_{r_2}} ... \sum_{x_{r_k}} P(X), x_{r_t} \in X \setminus Y. \tag{1}$$

Если использовать разложение $P(X)$ на множители и вынести все множители, не зависящие от x_{r_k} из под знака внутреннего суммирования, получим

$$P(Y) = \sum_{x_{r_1}} \sum_{x_{r_2}} ... \sum_{x_{r_{k-1}}} \prod_{x_i \notin Child^*(x_{r_k})} P(x_i|Par(x_i)) \sum_{x_{r_k}} \prod_{x_i \in Child^*(x_{r_k})} P(x_i|Par(x_i)), x_{r_t} \in X \setminus Y. \tag{2}$$

Обозначим результат внутреннего суммирования $\sum_{x_{r_k}} \prod_{x_i \in Child^*(x_{r_k})} P(x_i|Par(x_i))$ как $T(X^{r_k})$ и перепишем равенство:

$$P(Y) = \sum_{x_{r_1}} \sum_{x_{r_2}} ... \sum_{x_{r_{k-1}}} \prod_{x_i \notin Child^*(x_{r_k})} P(x_i|Par(x_i)) T(X^{r_k}), x_{r_t} \in X \setminus Y. \tag{3}$$

Заметим, что аналогичную операцию можно было осуществить при любом разложении $P(X)$ на множители вне зависимости от того, имеют ли их значения смысл условных вероятностей. Это позволяет последовательно исключить все k переменных и получить $P(Y)$.

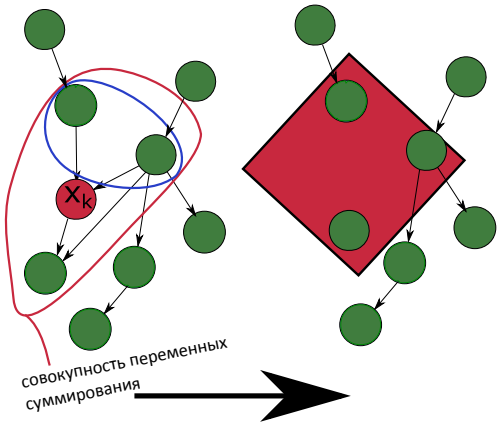


Рис. 7: Исключение переменной в байесовской сети.

Заметим, однако, что для получаемых в процессе исключения множителей, подобных $T(X^{r_k})$, не предполагается наличия структуры - то есть они представляют из себя таблицы, ставящие в соответствие каждой комбинации переменных из $T(X^{r_k})$ значение. Это означает, во-первых, что память, требуемая для их хранения и время, необходимое для совершения операций, растёт экспоненциально с ростом числа переменных в X^{r_k}. То же верно для результирующего представления $P(Y)$. Во-вторых, факторизация распределения $P(X \setminus \{x_{r_k}\})$, полученная после исключения первой переменной, уже не представляется в виде байесовской сети, и не обладает её преимуществами - например, возможностью эффективно оценить вероятность данного набора значений переменных, или эффективно сгенерировать выборку из распределения. В последующих разделах будет показано, что способ может быть модифицирован, чтобы избавиться от этих недостатков.

2.2.2 Монте-карло

Вероятности различных комбинаций значений интереса можно оценивать приближенно на основе предварительно сгенерированной выборки из распределения.

Для генерации случайного элемента выборки достаточно фиксировать произвольный линейный порядок, согласующийся с частичным порядком, заданным графом

($x < y \Leftrightarrow y \in Desc(x)$). После этого значения x_i генерируются последовательно из распределения $P(x_i|Par(x_i))$.

Следующие два неравенства позволяют дать гарантии точности приближения в зависимости от размера выборки [9]

Неравенство Хёффинга:

$$P(\hat{P}_y \notin [P(y) - \epsilon, P(y) + \epsilon]) \leq 2e^{-2M\epsilon^2} \tag{4}$$

Неравенство Чернова:

$$P(\hat{P}_y \notin [(1 - \epsilon)P(y), (1 + \epsilon)P(y)]) \leq 2e^{-MP(y)\epsilon^2/3} \tag{5}$$

Здесь $P(y)$ - оцениваемая вероятность, $\hat{P}_y$ - выборочная оценка, M - размер выборки. Можно также выразить из этих неравенств размер выборки, необходимый, чтобы вероятность ошибки (аддитивной для неравенства Хёффинга и мультипликативной для неравенства Чернова), превышающей ϵ была ограничена величиной δ.

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

$$M \geq 3\frac{\ln(2/\delta)}{P(y)\epsilon^2}$$

2.3 Проверка статистических гипотез

2.3.1 Тест χ^2 согласия Пирсона

Тест χ^2[2] используется для проверки согласия выборки X с табличным распределением $P(x)$ переменной x, принимающей значения $\alpha_1, ..., \alpha_K$.

Пусть N - размер выборки, $p_k = P(x = v_k)$, n_k - количество элементов выборки X, равных v_k.

Статистика критерия вычисляется как $t = \sum_{k=1}^{K} \frac{(n_k - Np_k)^2}{Np_k}$.

Гипотеза отвергается с уровнем значимости α, если величина t превышает $\chi^2_{1-\alpha,K-1}$ - $1 - \alpha$ - квантиль распределения χ^2 с $K - 1$ степенью свободы.

Плотность распределения χ^2 с d степенями свободы задаётся формулой

$$\frac{1}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}x^{\frac{d}{2}-1}e^{-\frac{x}{2}}$$

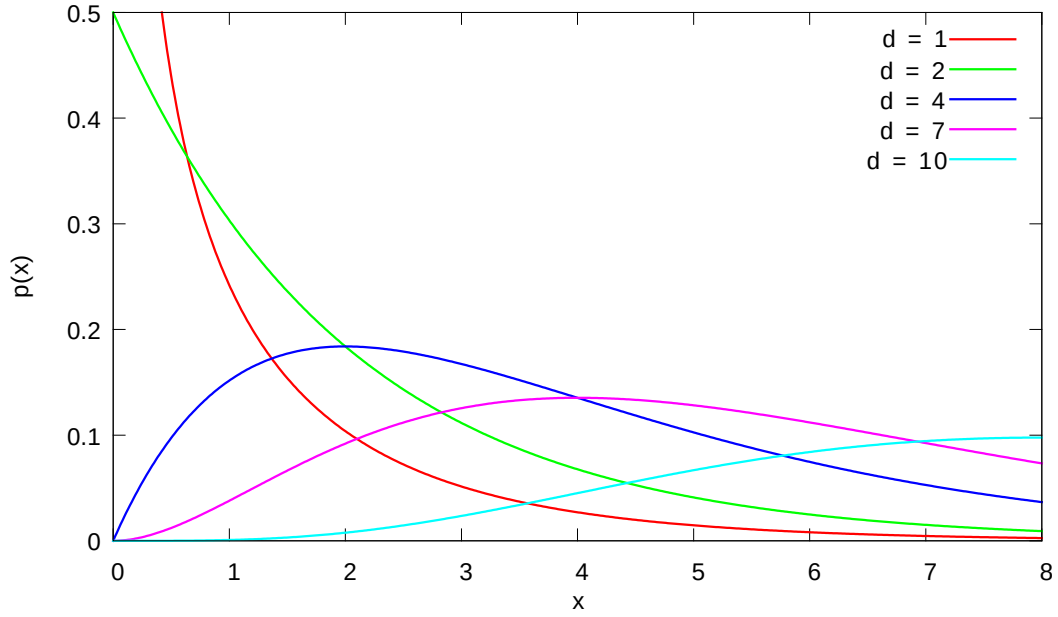


Рис. 8: Распределение χ^2 с d степенями свободы.

Тест χ^2 является приближенным - распределение статистики асимптотически стремится к распределению $\chi^2_{1-\alpha,K-1}$ при росте выборки, однако может давать неверные оценки при малых размерах выборки.

2.3.2 Тест отношения правдоподобия (G-тест)

G-тест является альтернативой теста χ^2. В аналогичных обозначениях статистика вычисляется как

$$G = 2\sum_{k=1}^{K} n_k \log(\frac{n_k}{Np_k}).$$

Статистика также имеет в пределе распределение χ^2[10].

2.3.3 Множественная проверка гипотез

В случаях, когда в рамках одного эксперимента производятся несколько статистических тестов, простой контроль уровня значимости для отдельных гипотез может вести к неадекватным выводам. Поэтому в задачах множественной проверки гипотез обычно фиксируются дополнительные общие параметры эксперимента, и уровни значимости отдельных тестов вычисляются на их основе.

	H_0 не отвергнута	H_0 отвергнута	Всего
H_0 истинна	N_{00}	N_{10}	M_0
H_0 ложна	N_{01}	N_{11}	M_1
всего	m-R	R	m

Существует два основных подхода к выбору параметров [7]:

1. Групповая вероятность ошибки (FWER) - вероятность допустить хотя бы одну ошибку первого рода $P(N_{10>0})$
2. Доля ложных отклонений нулевой гипотезы (FDP) $= \begin{cases} \frac{N_{10}}{R}, R > 0 \\ 0, R = 0 \end{cases}$

В рамках второго подхода обычно используются False Discovery Rate (FDR) - ожидаемая доля ложных отклонений $\mathbb{E}(FDP)$ и False Discovery Exceedence (FDX) - превышение доли ложных отклонений $P(FDP > c)$.

Методы построения уровней значимости делятся на одношаговые и многошаговые. Методы первой группы фиксируют единый уровень значимости для всех гипотез. В многошаговых методах гипотезы сортируются по достигаемому уровню значимости и используется последовательная процедура нахождения позиции, с которой все гипотезы отвергаются.

Простейшим одношаговым методом контроля групповой вероятности ошибки является использование поправки Бонферрони. В рамках этого метода используется уровень значимости $\frac{\alpha}{n}$, где n - количество проверяемых гипотез, α - базовый уровень

значимости. Преимуществом метода является то, что он не требует никаких предположений о зависимости гипотез. Однако этот метод обладает очень малой мощностью. Равномерно более мощным является многошаговый метод Холма. Ещё более мощным методом контроля групповой вероятности ошибки является многошаговый метод Хохберга, однако в этом методе делаются предположения о независимости распределений статистик. В методах Холма и Хохберга значения статистик предварительно сортируются по возрастанию достигаемого уровня значимости. Далее каждый достигаемый уровень значимости p_k сравнивается с величиной $\frac{\alpha}{n+1-k}$. Отвергаются все гипотезы до индекса r, где дла процедуры Холма $r = min(\{k : p_k > \frac{\alpha}{n+1-k}\}) - 1$, а для процедуры Хохберга $r = max(\{k : p_k \leq \frac{\alpha}{n+1-k}\})$.

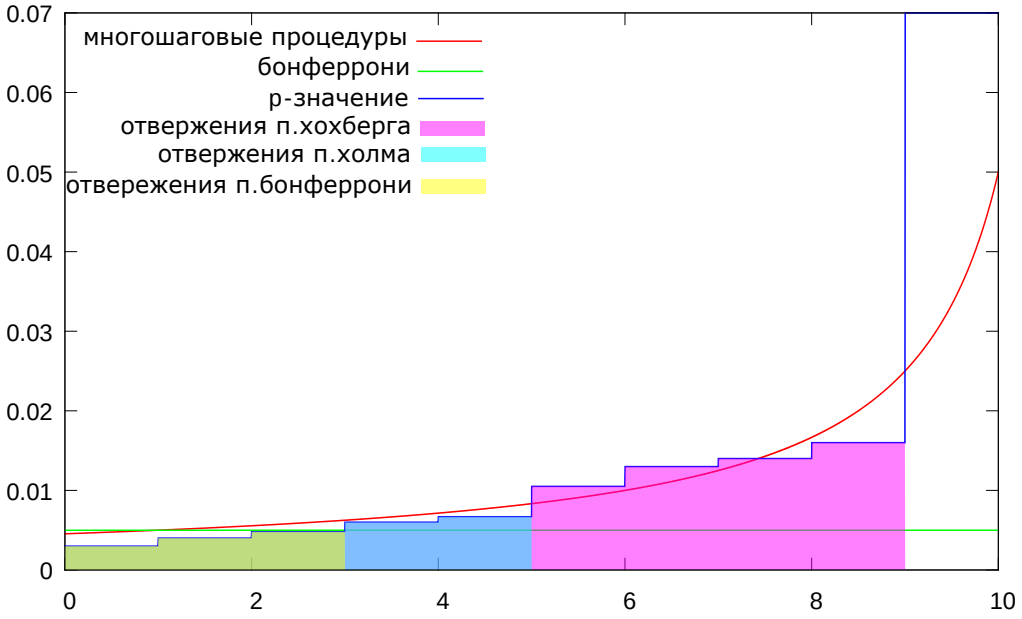


Рис. 9: Пример областей отвержения гипотез различными методами контроля FWER.

Методы контроля FDP в целом обладают большей мощностью за счёт того, что допускается небольшое количество ошибочных отвержений. С более детальным обзором методов множественной проверки гипотез можно познакомиться в [7].

2.4 Случайные графы

Генерация байесовской сети сводится к двум подзадачам:

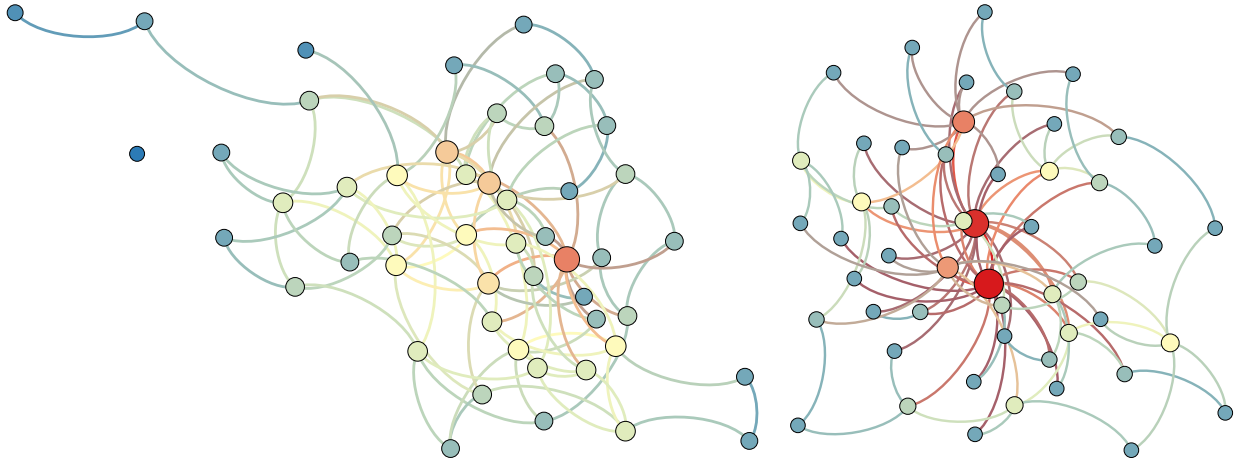
1. Генерация ациклического ориентированного графа, задающего систему зависимостей.

2. Генерация условных распределений индивидуальных переменных для всевозможных значений переменных-предков.

В этом разделе будут описаны два подхода к решению первой из подзадач. Эти подходы не исчерпывают всё множество вариантов[1], однако нам не удалось найти моделей методов генерации графов, которые были бы ориентированы на моделирование каких-либо реальных графов, кроме социальных сетей.

2.4.1 Модель Эрдеша-Реньи

Модель Эрдеша-Реньи (Рис. 10 (a)) является наиболее простой из возможных моделей случайных графов: при генерации графа фиксируется множество вершин и величина p, задающая вероятность вхождения каждого отдельного (неориентированного) ребра в граф. В альтернативной формулировке вместо вероятности p фиксируется число k рёбер, и выбирается равномерно случайное подмножество возможных рёбер мощности k .



(a) (b)

Рис. 10: Случайные графы: (a) - граф Эрдеша-Реньи с 50 вершинами и 100 рёбрами, (b) граф Барабаши-Альберт с 50 вершинами и 97 рёбрами

Результат описанной процедуры - неориентированный граф. Для получения на его основе ориентированного ациклического графа можно зафиксировать произвольный линейный порядок переменных и выбрать направление каждого ребра в сторону возрастания номера вершины.

2.4.2 Модель Барабаши-Альберт

В модели Барабаши-Альберт (рис. 10 (b)) фиксируется натуральное число m и моделируется последовательное добавление вершин к графу, начиная с полного графа с m вершинами. На каждом шаге генерируется m рёбер, соединяющих вновь добавленную вершину с существующими. Вероятность выбора существующей вершины для соединения с вновь добавленной пропорциональна её степени.

В этой модели при задании ориентации рёбер порядок существенно влияет на вид графа (степени вершин, добавленных ранее, в среднем больше, чем степени вершин, добавленных последними). Поэтому при экспериментах использовалось три разных способа задания порядка:

1. Вершины пронумерованы в порядке, в котором они добавлялись в модель.

2. Вершины пронумерованы в обратном порядке.

3. К вершинам применяется случайная перестановка с равными вероятностями всех порядков.

2.5 Генерация табличных распределений

Одномерное табличное распределение переменной с k возможными значениями представляет из себя k-мерный вектор неотрицательных вещественных чисел, сумма которых равна единице. Одним из возможных распределений на множестве таких векторов является распределение Дирихле[13].

$$Dir(x;\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

Распределение Дирихле параметризуется вектором неотрицательных вещественных чисел $\alpha = \alpha_1, ..., \alpha_k$. Заметим, что в области ненулевых вероятностей любое значение случайной величины, распределённой по закону Дирихле, однозначно определяется $k-1$ элементом вектора. Специальный случай распределения Дирихле при $k = 2$ носит название Бета-распределения. Также следует отметить, что значение параметра $\alpha = \vec{1}$ соответствует, фактически, равномерному распределению.

В экспериментах при генерации условных распределений для каждого набора значений наблюдаемых переменных распределения генерировались независимо.

3 Предлагаемый метод

Для нашего метода мы предполагаем, что дана байесовская сеть с дискретными переменными, принимающими конечное множество значений, и выборка векторов значений этих переменных. Мы рассматриваем гипотезу о том, что выборка была сгенерирована из распределения, определяемого заданной сетью, в качестве нулевой.

Суть метода заключается в том, чтобы провести серию статистических тестов согласия маргинальных распределений (χ^2 и теста отношения правдоподобий) под-

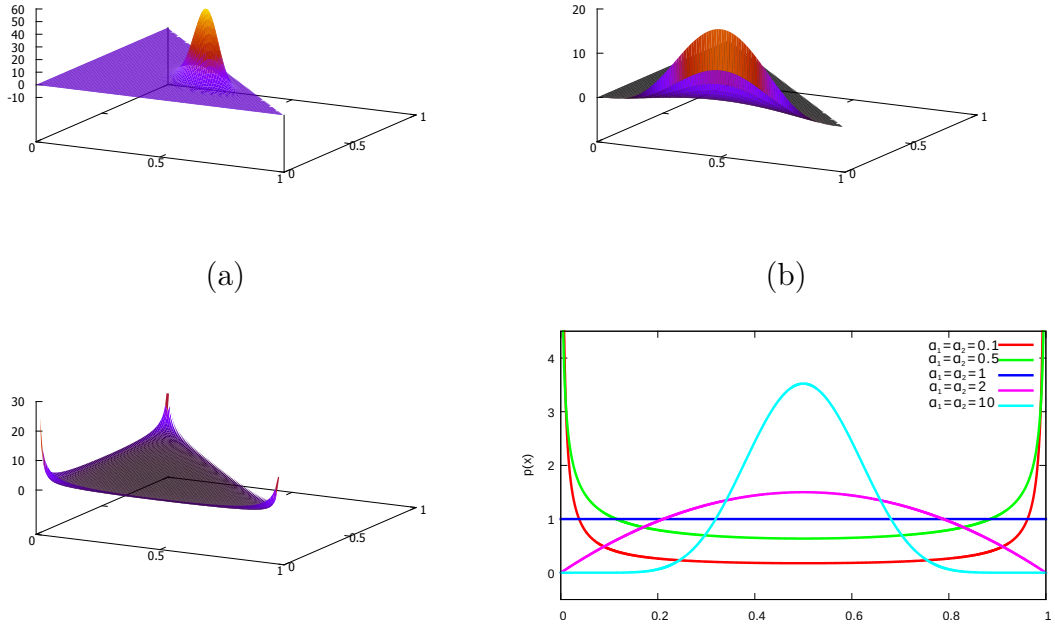


Рис. 11: Распределение Дирихле. (a) $Dir(x; 30, 20, 10)$,(b) $Dir(x; 5, 0.1, 5)$,(c) $Dir(x; 0.2, 0.2, 0.2)$, (d) Бета-распределение с равными параметрами.

множеств переменных с данными, и отвергнуть нулевую гипотезу для сети при отвержении хотя бы одной гипотезы для маргинальных распределений. При этом задаваемый уровень значимости корректируется при помощи процедур для множественной проверки гипотез.

4 Теоретические результаты

4.1 О теоретических границах применимости метода

В предложенном подходе к верификации проверяемое многомерное распределение, фактически, заменяется набором маргинальных распределений подмножеств переменных небольшой мощности. Возникает вопрос, эквивалентна ли подобная замена. Чтобы ответить на этот вопрос, рассмотрим совместное распределение векторной дискретной случайной величины $X = x_1, ..., x_n$, каждая из переменных x_i которых принимает s_i возможных значений. Множество таких распределений может быть взаимно однозначно закодировано множеством таблиц, ставящих в соответствие каждой комбинации значений переменных неотрицательное вещественное число. При этом накладывается дополнительное ограничение: сумма чисел в каждой таблице равняется единице. Количество чисел в такой таблице равно $S = \prod_{i=1}^{n} s_i$. Без ограничений неотрицательности и суммирования в единицу это множество представляет из себя линейное пространство размерности S. Суммирование элементов таблицы в единицу является линейным равенством, выделяющим линейное многообразие размерности $S - 1$. Множество ограничений неотрицательности дополнительно выделяет выпуклое подмножество этого многообразия.

Рассмотрим теперь множество всевозможных маргинальных распределений подмножеств переменных мощности k как множество ограничений на возможные распределения. Обозначим множество подмножеств таких переменных Sub_k

Каждое маргинальное распределение является результатом суммирования по всевозможным значениям прочих переменных $K \subset Sub_k$ и задаёт, таким образом, мно-

жество линейных уравнений. Количество уравнений равно количеству возможных комбинаций значений переменных, входящих в K. Мы можем посчитать, таким образом, количество полученных распределений как $\sum_{K \subset Sub_k} \prod_{i:x_i \in K} s_i$.

Рассмотрим простой случай, когда все переменные принимают равное количество значений s. Тогда $S = s^n$, а мощность множества ограничений, задаваемых маргинальными распределениями, равна $C_n^k s^k$. В итоге множество допустимых распределений представляет из себя подмножество линейного многообразия размерности s^n, задаваемое не более чем $C_n^k s^k + 1$ линейными ограничениями равенства и, ограничениями неравенства, задающими выпуклое множество. Заметим также, что существует как минимум одна точка, удовлетворяющая всем ограничениям - исходное распределение, из которого получены маргинальные распределения. Это позволяет сделать следующее

Утверждение 1. *Если $C_n^k s^k + 1 < s^n$, существует континуальное множество распределений, маргинальные распределения которых совпадают с заданными.*

Неравенство же, очевидно, выполняется при достаточно больших значениях n или s.

Таким образом, метод не является точным для произвольных распределений. Это не говорит, однако, о его неприменимости к решению практически важных задач.

4.2 Последовательное исключение переменных с сохранением свойств байесовской сети

Целью данной главы будет доказать следующее

Утверждение 2. *$P(X \setminus \{y\})$ можно представить в виде байесовской сети с графом $\hat{G}$, для любого ребра (a,b) которого выполнено одно из условий:*

1. (a, b) - ребро графа G исходной байесовской сети для $P(X)$.

2. $a \in X^y, b \in X^y, b \in Child(y), ind(b) > ind(a)$, где $ind(v)$ - индекс вершины v в фиксированном топологическом порядке (линейном порядке, в котором для любого направленного ребра (v_1, v_2) $ind(v_1) < ind(v_2)$).

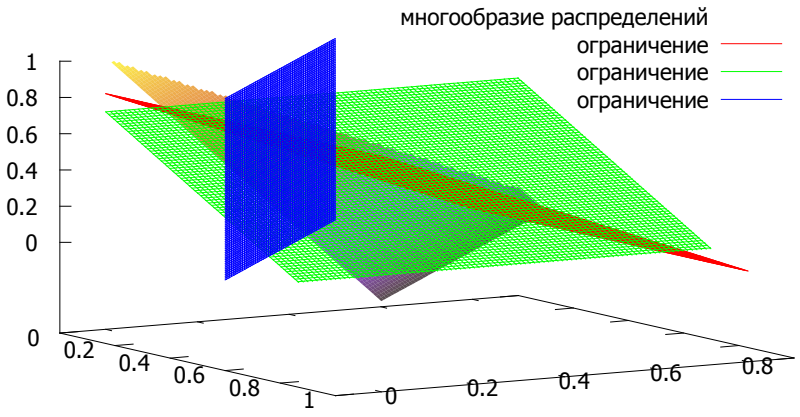


Рис. 12: Вид ограничений, накладываемых маргинальными распределениями на множество распределений

Доказательство. Доказательство состоит из последовательного использования двух вспомогательных утверждений ниже. □

Благодаря теореме 1, достаточно сравнить множество выполненных утверждений об условной независимости в распределении $P(X \setminus \{y\})$ со множеством выполненных утверждений о d-разделённости в полученном графе. Заметим, что первое является множеством всех утверждений об условной независимости исходного распределения, не включающих переменную y.

Пусть даны два графа $G_1(X, E_1)$ и $G_2(Y \subset X, E_2)$.

Определение 7. *Будем говорить, что граф G_2 сохраняет активные пути G_1 если для любой функции $f : X \to \{o, u\}$ такой, что $\forall z \in X \setminus Y : f(z) = u$ и любой пары ненаблюдаемых вершин $y_1, y_2 \in Y$ из существования активного пути между y_1 и y_2 в G_1 следует существование активного пути между ними в G_2.*

Вспомогательное утверждение 1. *Пусть даны два графа $G_1(X, E_1)$ и $G_2(X \setminus \{y\}, E_2)$ и распределение $P(X)$, факторизуемое на G_1. Достаточным условием фак-*

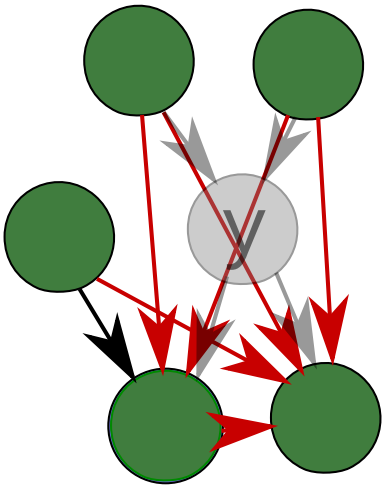


Рис. 13: Рёбра, которые достаточно добавить в граф при удалении вершины для сохранения свойств байесовской сети.

торизуемости $P(X \setminus \{y\})$ *на* G_2 *является сохранение графом* G_2 *активных путей* G_1.

Доказательство. Пусть дан произвольный граф G_2, сохраняющий активные пути G_1. Благодаря т.1, чтобы показать истинность утверждения, достаточно показать, что для любых непересекающихся множеств переменных $A, B, C \subset X^y$ из d-разделённости A и B при условии C в G_2 следует соответствующая независимость. Заметим, что множество троек (A,B,C) непересекающихся подмножеств $X \setminus \{y\}$ таких, что $A \perp B|C$ в распределении $P(X \setminus \{y\})$ является в точности множеством троек (A,B,C) непересекающихся подмножеств $X \setminus \{y\}$, таких, что $A \perp B|C$ в исходном распределении $P(X)$. Заметим также, что из определения d-разделённости и определения графа, сохраняющего активные пути следует, что любое утверждение о d-разделённости, выполненное в G_2, выполнено также в G_1. Это значит, что выполнено и утверждение об условной независимости в $P(X)$ и $P(X \setminus \{y\})$. □

Вспомогательное утверждение 2. *Граф* $G_2(X \setminus \{y\}, E_2)$, *полученный из графа* $G_1(X, E_1)$ *путём удаления вершины* y *вместе с инцидентными рёбрами и соединения рёбрами всех пар вершин (a,b) таких, что* $a, b \in X^y, b \in Child(y), ind(b) > ind(a)$, *сохраняет активные пути* G_1.

Доказательство. Рассмотрим пару вершин v_1,v_2, соединённых активным путём в G_1 для некоторого множества наблюдаемых переменных. Заметим, что, если существует активный путь между двумя вершинами, то всегда существует и активный путь, являющийся простым путём (путь, в который каждая вершина входит не более одного раза), поэтому будем рассматривать только простые пути.

Предположим этот активный путь не содержит вершину y. В таком случае единственный способ, которым удаление вершины y могло сделать этот путь неактивным - это нарушение свойства V-структуры (наблюдаемый потомок центральной вершины V-структуры - обозначим её v_c - перестал быть потомком). Это могло произойти только если направленный путь от v_c к потомку проходил через y. Однако в таком случае этот путь содержит родительскую и дочернюю вершину y, а они должны быть соединены ребром по условию утверждения.

Рассмотрим теперь произвольный активный путь из v_1 в v_2, содержащий y. Выберем минимальный активный подпуть S этого пути, содержащий y. Заметим, что любой подпуть активного пути, начинающийся и заканчивающийся ненаблюдаемой вершиной является активным. При этом любой активный подпуть ограничен ненаблюдаемыми вершинами по определению. Это означает, что S - минимальный подпуть исходного пути, ограниченный ненаблюдаемыми вершинами.

Заметим, что существует лишь 3 (поскольку порядок в пути не важен) возможных конфигурации рёбер, соединяющих y с соседями в этом пути:

1. $\rightarrow y \rightarrow$
2. $\leftarrow y \rightarrow$
3. $\rightarrow y \leftarrow$

Обозначая o любую наблюдаемую, а u - любую ненаблюдаемую переменную и пользуясь тем, что наблюдаемая вершина не может иметь исходящих рёбер в активном пути, мы можем описать возможные конфигурации окрестности y следующим образом:

1. $u \to y \to < u|o >$
2. $< u|o > \leftarrow y \to < u|o >$
3. $u \to y \leftarrow u$

Уже следующий шаг позволяет описать все возможные конфигурации S (штрихи добавлены, чтобы различить вершины одного типа):

1. $u \to y \to u'$
2. $u \to y \to o \leftarrow u'$
3. $u \leftarrow y \to u'$
4. $u \leftarrow y \to o \leftarrow u'$
5. $u \to o \leftarrow y \to o' \leftarrow u'$
6. $u \to y \leftarrow u'$

Нетрудно видеть, что все вершины S принадлежат марковскому одеялу y. Некоторые отношения индексов этих вершин в топологическом порядке мы можем определить на основании того, что для любой пары вершин v_1, v_2 таких, что в графе существует направленный путь из v_1 в v_2: $ind(v_1) < ind(v_2)$ (очевидное следствие определения топологического порядка). Рассмотрим последовательно 6 перечисленных вариантов, и покажем, что в каждом из них в графе G_2 обязательно присутствует структура, восстанавливающая активный путь (точнее, создающая новый активный путь, соединяющий те же вершины), который, возможно, перестал быть таковым при удалении y.

1. $u \to y \to u'$

   Здесь $u' \in Child(y)$, $ind(u) < ind(u')$, то есть в графе G_2 присутствует ребро (u, u'), восстанавливающее активный путь. (Рис. 14 (а))

2. $u \to y \to o \leftarrow u'$

   $o \in Child(y)$, $ind(u) < ind(o)$. Это означает, что в G_2 будет добавлено ребро (u, o), создающее V-структуру $u \to o \leftarrow u'$, и, тем самым, восстанавливающее активный путь. (Рис. 14 (g))

3. $u \leftarrow y \to u'$

   Вершины u и u' - дочерние вершины y. Вне зависимости от их относительного порядка, они соединены ребром, восстанавливающим активный путь. (Рис. 14 (c))

4. $u \leftarrow y \to o \leftarrow u'$

   $u, o \in Child(y)$. Если $ind(u') < ind(u)$, то в G_2 присутствует ребро (u', u), восстанавливающее активный путь. (Рис. 14 (e)). В противном случае $ind(u) < ind(u') < ind(o)$ и в G_2 присутствует ребро (u, o), и вместе с ним V-структура $u \to o \leftarrow u'$, восстанавливающая активный путь (Рис. 14 (d)).

5. $u \to o \leftarrow y \to o' \leftarrow u'$

   Вне зависимости от относительного порядка u и u', как минимум одна из двух наблюдаемых вершин из S имеет индекс, больший индексов обеих ненаблюдаемых вершин, таким образом в G_2 присутствуют рёбра, соединяющие u и u' с одной наблюдаемой вершиной, что гарантирует восстановление активного пути. (Рис. 14 (f))

6. $u \to y \leftarrow u'$

   Здесь u и u' - родительские вершины y. Существование между ними активного пути в G_1 говорит о том, что существует направленный путь, соединяющий y с некоторой наблюдаемой вершиной. Соединение u и u' с дочерней вершиной y, принадлежащей этому пути, создаёт V-структуру и восстанавливает активный путь. (Рис. 14 (b))

□

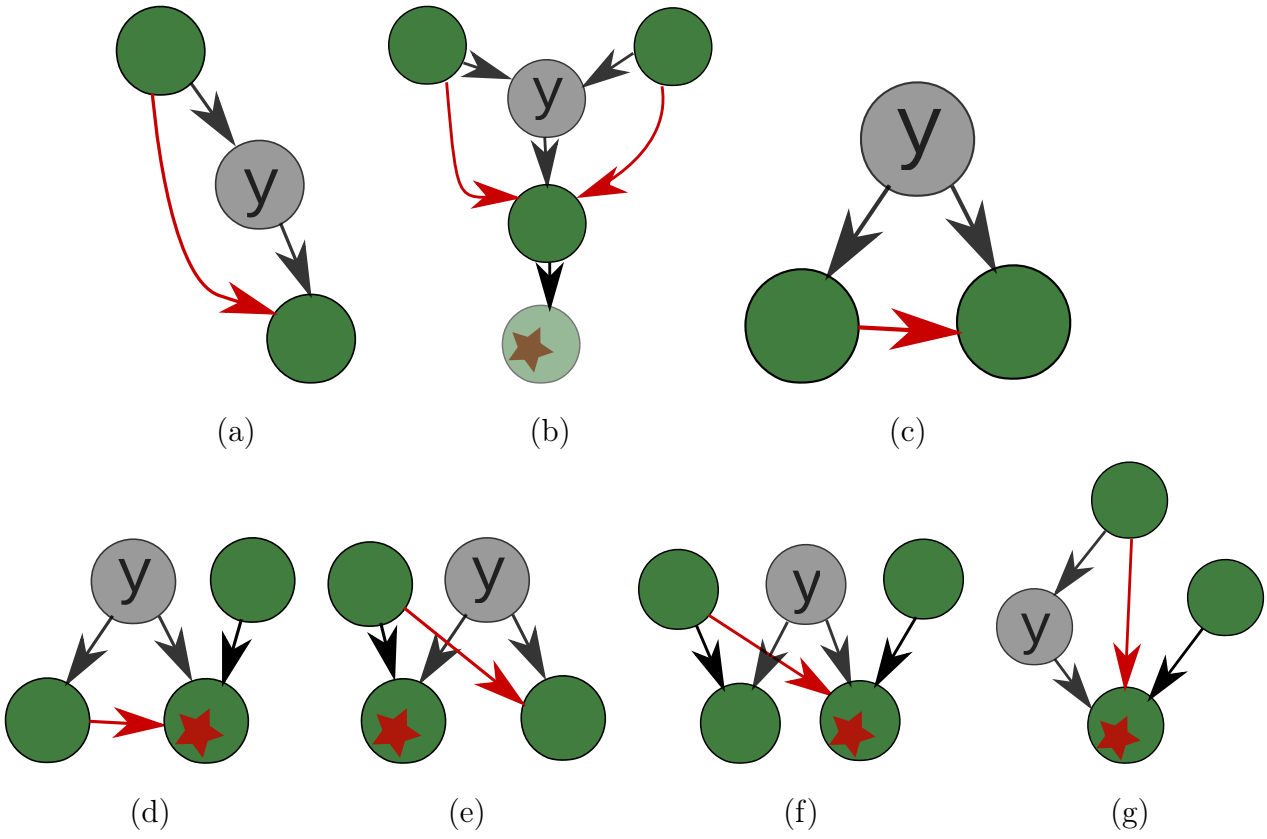


Рис. 14: Восстановление разомкнутых активных путей.

Мы показали, что для задания байесовской сети для распределения $P(X \setminus \{y\})$ достаточно соединить рёбрами дочерние вершины y с вершинами X^y, имеющими меньшие индексы. Это означает, что остаются неизменными все условные распределения для переменных из $X \setminus \{y\} \setminus Child(y)$, то есть все распределения, не вошедшие в множитель T в базовом алгоритме последовательного исключения. Таким образом, для перехода к новой байесовской сети достаточно найти распределения $P(x|\tilde{P}ar(x))$, где тильдой помечены множества, заданные на новом графе.

Искомое распределение $P(x|\tilde{P}ar(x)) = P(x|\tilde{P}red(x)) = \frac{P(x,\tilde{P}red(x))}{P(\tilde{P}red(x))} = \frac{\sum_y P(x,Pred(x))}{\sum_y P(Pred(x))} = \frac{\sum_y P(Pred(x))P(x|Par(x))}{\sum_y P(Pred(x))}$. Распределения в числителе и знаменателе описанной дроби представляют из себя произведения условных распределений исходной байесовской сети $P(t|Par(t))$, причем множители для $t \notin \{y\} \cup Child(y)$ не зависят от y и могут быть вынесены из под знака суммирования. Множество этих множителей совпада-

ет для числителя и знаменателя, и можно сократить на них дробь, исключив их, таким образом из, формулы. Отсюда следует, что условные распределения можно вычислять следующим образом:

1. Инициализировать $\mathcal{P}_0 = P(y|Par(y))$.
2. В порядке возрастания индексов k_i переменных x_{k_i} из $Child(y)$:
3. Положить $\mathcal{P}_i = \mathcal{P}_{i-1} P(x_{k_i}|Par(x_{k_i}))$.
4. $P(x_{k_i}|\tilde{P}ar(x_{k_i})) = \frac{\sum_y \mathcal{P}_i}{\sum_y \mathcal{P}_{i-1}}$.

5 Экспериментальные результаты

В данном разделе экспериментально исследуется вероятность совершения ошибки первого рода при использовании исследуемого метода. Эксперименты заключались в многократном применении процедуры проверки гипотез к выборке, сгенерированной из истинного распределения, где истинное распределение, в свою очередь, также генерировалось в виде случайной байесовской сети.

За основу были взяты маргинальные распределения для пар и троек переменных: в первом случае использовались всевозможные пары, во втором - случайное подмножество троек мощности 100. Байесовская сеть генерировалась в два этапа: сначала генерировался случайный граф при помощи одного из описанных выше типов (модель Эрдеша-Реньи или Барабаши-Альберт с использованием прямого, обратного или случайного порядка переменных для получения направленного графа), затем для каждой переменной и для каждой комбинации значений переменных-предков табличное условное распределение генерировалось из Бета-распределения, с равными параметрами α.

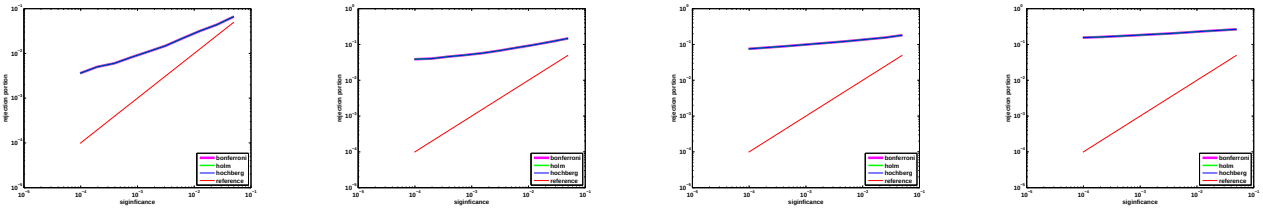
Во всех экспериментах генерировались байесовские сети с 10 бинарными переменными. Размер выборки для тестирующей процедуры во всех экспериментах был равен 100, что соответствует характерному размеру выборок в медицинских исследованиях.

Эксперименты производились для различных комбинаций параметров. Оценки вероятности ошибки первого рода получались как доля ложных отвержений для 5000 экспериментов, проведённых для фиксированных наборов параметров.

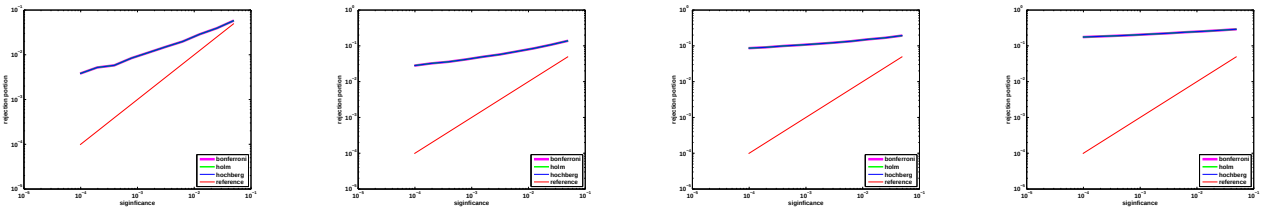
Все графики приводятся в двойной логарифмической шкале, нулевые значения заменены на 10^{-20}.

пары, $\alpha = 1$ тройки, $\alpha = 1$ пары, $\alpha = 0.2$ тройки, $\alpha = 0.2$

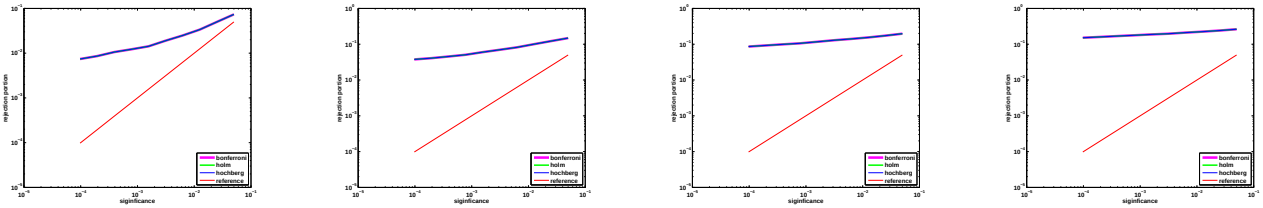
Модель Эрдеша-Реньи



Модель Барабаши-Альберт, прямой порядок



Модель Барабаши-Альберт, обратный порядок



Модель Барабаши-Альберт, случайный порядок

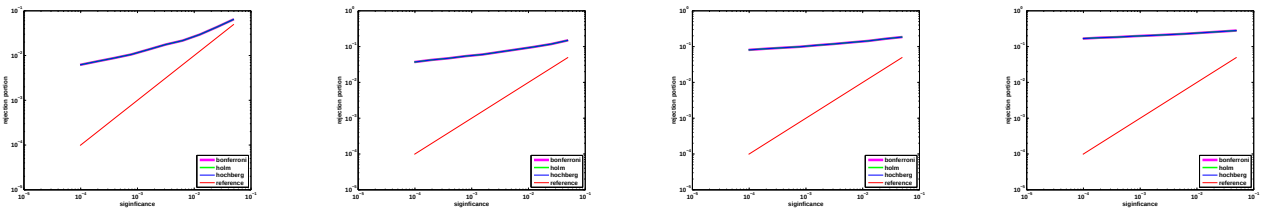
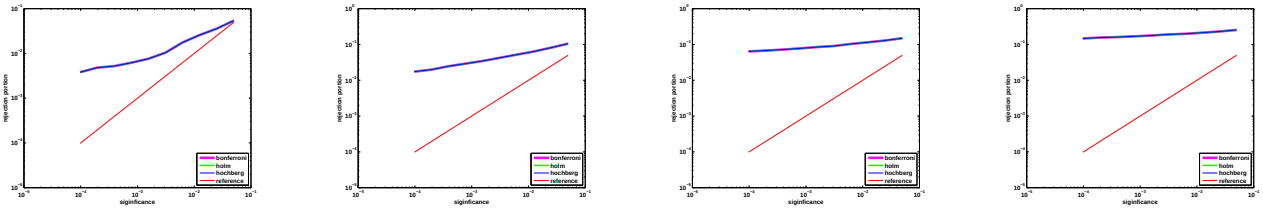


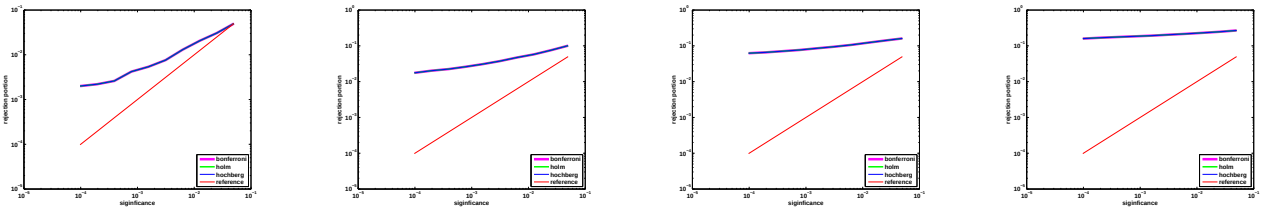
Рис. 15: Графики отклонения гипотез тестом χ^2- графы с 9 рёбрами.

пары, $\alpha = 1$ тройки, $\alpha = 1$ пары, $\alpha = 0.2$ тройки, $\alpha = 0.2$

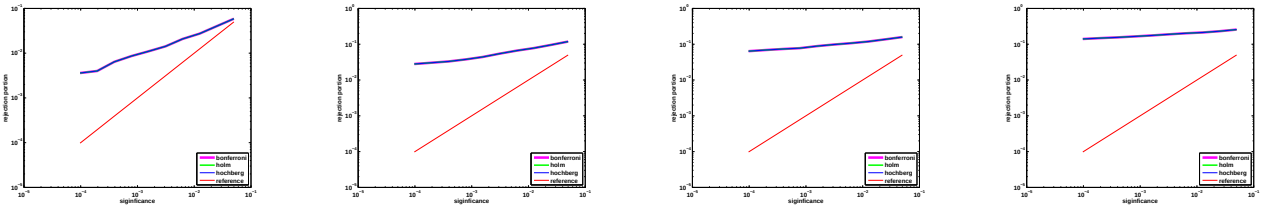
Модель Эрдеша-Реньи



Модель Барабаши-Альберт, прямой порядок



Модель Барабаши-Альберт, обратный порядок



Модель Барабаши-Альберт, случайный порядок

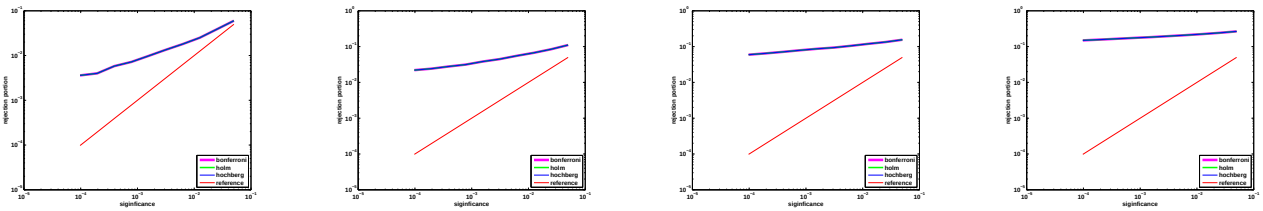
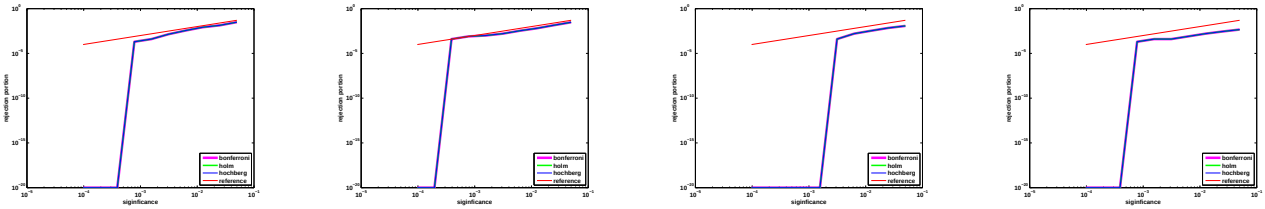


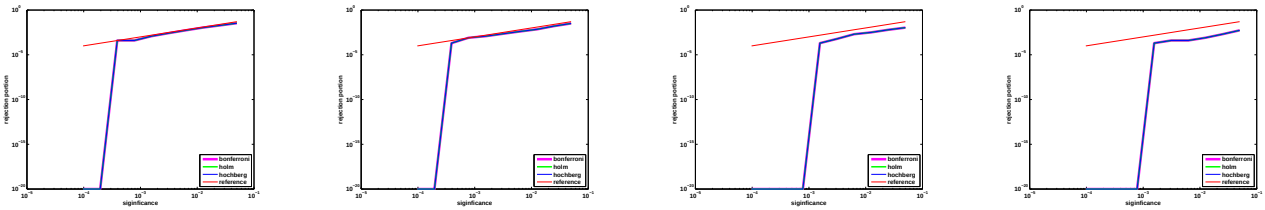
Рис. 16: Графики отклонения гипотез тестом χ^2- графы с 17-18 рёбрами.

пары, $\alpha = 1$ тройки, $\alpha = 1$ пары, $\alpha = 0.2$ тройки, $\alpha = 0.2$

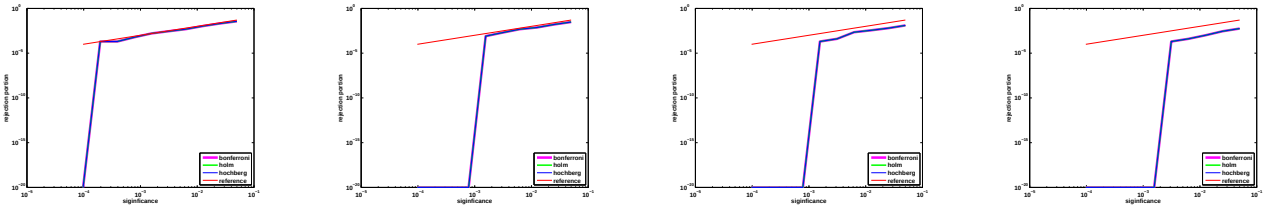
Модель Эрдеша-Реньи



Модель Барабаши-Альберт, прямой порядок



Модель Барабаши-Альберт, обратный порядок



Модель Барабаши-Альберт, случайный порядок

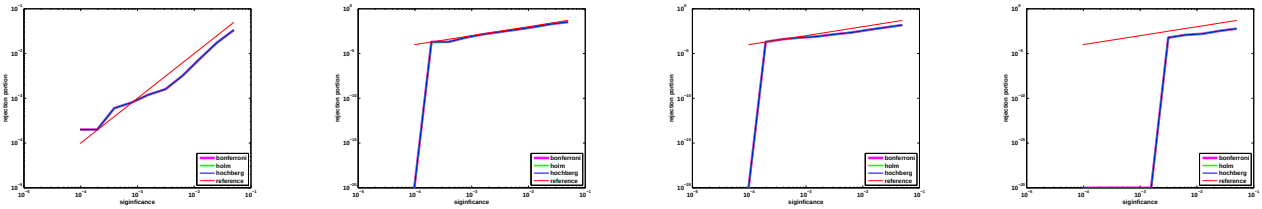
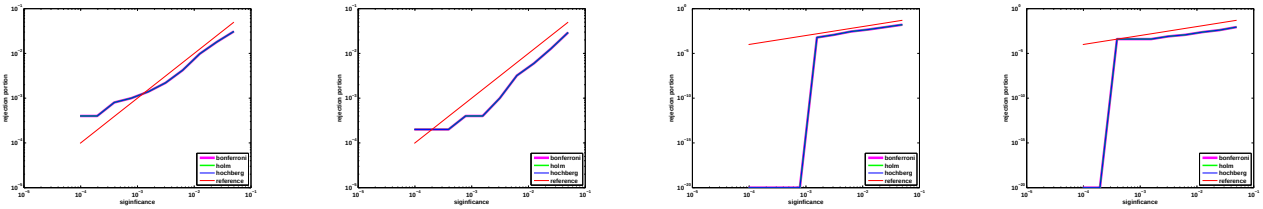


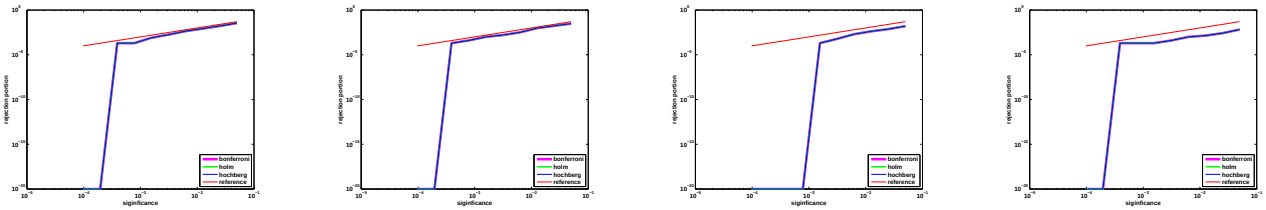
Рис. 17: Графики отклонения гипотез G-тестом - графы с 9 рёбрами.

пары, $\alpha = 1$ тройки, $\alpha = 1$ пары, $\alpha = 0.2$ тройки, $\alpha = 0.2$

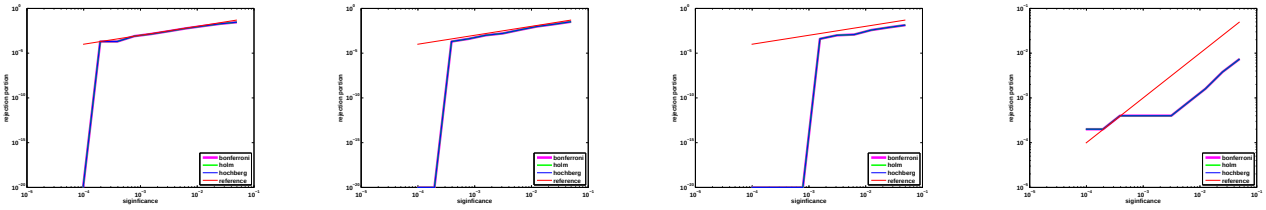
Модель Эрдеша-Реньи



Модель Барабаши-Альберт, прямой порядок



Модель Барабаши-Альберт, обратный порядок



Модель Барабаши-Альберт, случайный порядок

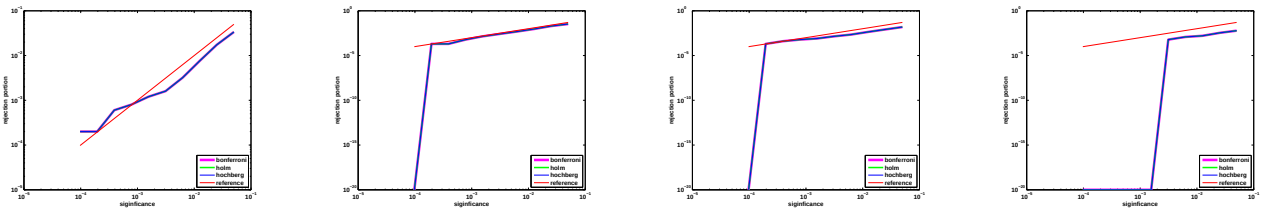


Рис. 18: Графики отклонения гипотез G-тестом - графы с 17-18 рёбрами.

На основе результатов проведённых экспериментов можно сделать следующие выводы:

1. Все процедуры коррекции множественной проверки гипотез из семейства методов контроля групповой вероятности ошибки дают одинаковые результаты.

2. Количество рёбер в графе и их распределение не оказывает значительного влияния на результаты метода.

3. Фиксированный уровень значимости близок к фактическому уровню значимости для сетей с большой энтропией условных распределений (Бета-распределение с параметром 1). Метод оказывается излишне консервативным, если условные распределения обладают малой энтропией (Бета-распределение с параметром 0.2).

4. Для малых размеров выборок, использованных в экспериментах, тест χ^2 гораздо больше подвержен ошибкам первого рода, чем тест отношения правдоподобий.

Первый вывод объясняется тем, что метод никак не зависит от количества отвергнутых гипотез, и результат, фактически, определяется результатом проверки гипотезы с наименьшим достигаемым уровнем значимости, который почти всегда сравнивается с одним и тем же пороговым значением (всегда для методов Бонферрони и Холма).

Второй и третий вывод можно объединить вместе, сказав, что влияние зависимости между проверяемыми гипотезами определяется не номинальным отсутствием независимости (которая определяется структурой графа), а степенью взаимного влияния переменных (которое зависит непосредственно от условных распределений).

Для выяснения причин, по которым доли ложных отклонений метода на основе теста χ^2 оказались стабильно больше заданного уровня значимости, был проведён дополнительный набор экспериментов без использования байесовских сетей. В рамках этих экспериментов генерировались мультиномиальные распределения, из каждого из которых генерировалась выборка размера 100, для которой производились

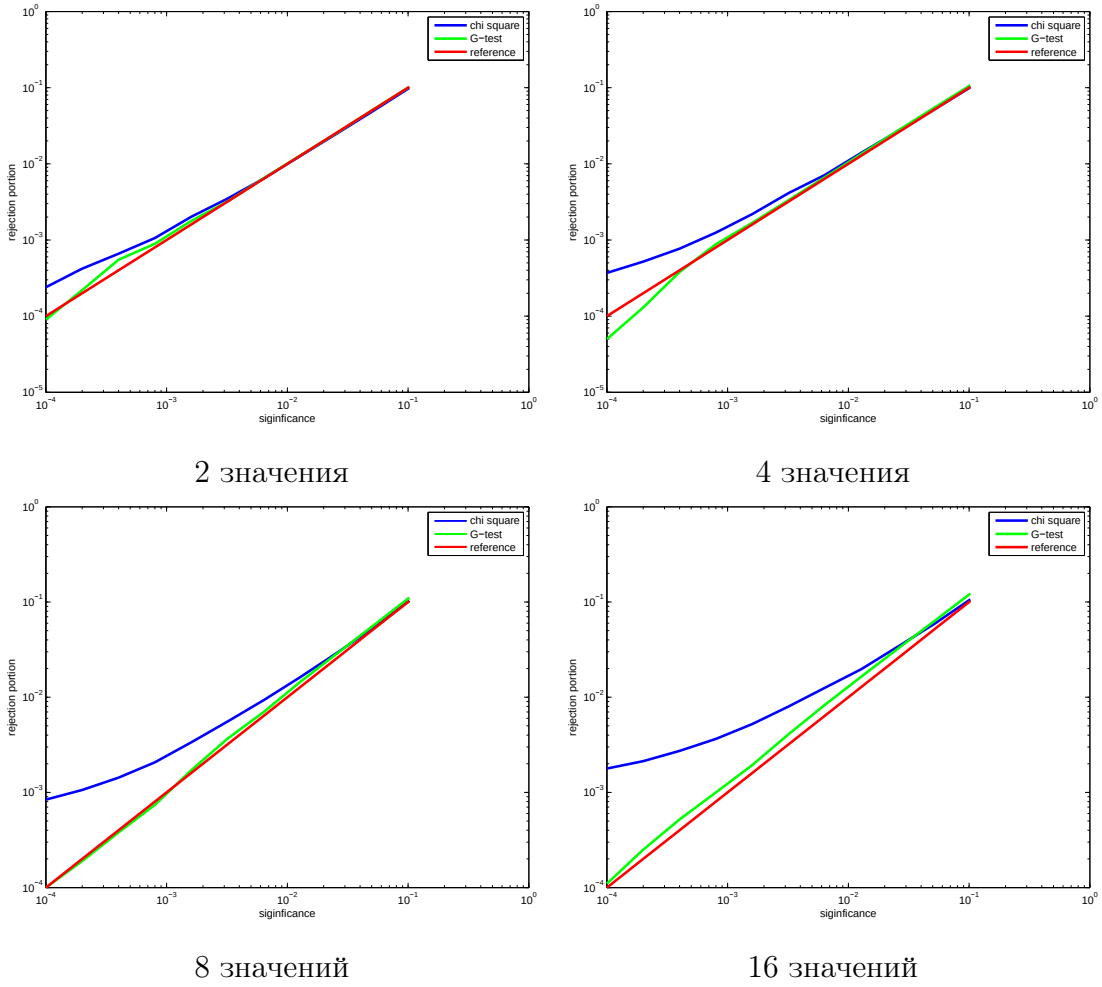


Рис. 19: Ложные отвержения в эксперименте с простыми табличными распределениями.

тесты согласия. Оценка вероятности отклонения получалась в результате усреднения по 100000 экспериментов. Таким образом, причиной плохих результатов теста χ^2 являются собственные ограничения теста, связанные с плохой устойчивостью при наличии малых вероятностей и при небольшом размере выборки.

6 Заключение и направления дальнейшей работы

В данной работе был исследован новый метод верификации распределений дискретных распределений, заданных в форме байесовских сетей в условиях недоступ-

ности больших выборок при помощи механизма множественной проверки гипотез. Была показана невозможность гарантировать заданный уровень значимости при использовании теста χ^2 и перспективность в этом отношении теста отношения правдоподобий в комбинации с любым из методов контроля групповой вероятности ошибки при множественной проверке гипотез. Было показано, что влияние структуры графа байесовской сети не оказывает значительного влияния на надежность метода, а также существует потенциал для уточнения метода при сохранении уровня значимости для распределений с сильной взаимной зависимостью переменных.

Предложенный метод может быть использован в медицинских исследованиях для проверки существующих представлений о взаимной зависимости переменных при получении новых данных, а также, в комбинации с методами оценки условных распределений, как инструмент для помощи эксперту в уточнении структуры графа.

Возможным направлением углубления результатов представляется систематическое исследование уровня ошибок второго рода. Также в текущей формулировке полностью игнорируются результаты проверки отдельных гипотез - их учёт может позволить локализовать найденные противоречия в байесовской сети. Использование в качестве основного метода теста отношения правдоподобий, статистика которого пропорциональна информационной дивергенции делает логичным переход от методов проверки гипотез к непосредственному использованию метрических методов. Ещё одним важным направлением расширения полученных результатов является учёт сетей с латентными переменными, и неоднородных выборок, порождённых их влиянием.

Кроме того, при разработке процедуры экспериментов была предложена модификация метода исключения переменных в байесовской сети. Эта процедура позволяет немного уменьшить размеры возникающих в процессе работы алгоритма клик (размер клик является определяющим параметром сложности алгоритма). Использование этой модификации делает выбор линейного порядка переменных существенным параметром для каждого шага алгоритма - таким образом, разработка хорошей эвристики для выбора этого порядка может потенциально привести к более суще-

ственному ускорению точного вывода. Можно отметить, что данная модификация помогает ответить на вопросы о значении структур, возникающих в процессе вывода - актуальность этого вопроса была отмечена, например в [19]. Также доступность метода построения байесовской сети, полученной из исходной исключением переменных может помочь в решении обратной задачи - поиска структуры, позволяющей получить более разреженную сеть за счёт введения дополнительных переменных.

7 Приложение

В данном приложении более подробно излагаются возможные шаги для развития результатов работы.

7.1 О локализации противоречий в байесовской сети

Любая ошибка в задании байесовской сети - как при определении рёбер в графе, так и непосредственно в задании условных распределений - представляет из себя набор ошибок в задании полных условных распределений при условии всех предшествующих переменных в графе. При отсутствии наблюдаемых переменных такие ошибки никак не влияют на корректность совместного распределения предшествующих переменных, но могут вносить вклад в распределения переменных-потомков. Логично предположить, что этот вклад различается для переменных, отстоящих от исходной переменной на различное расстояние в графе байесовской сети. Следующая теорема делает более вероятным предположение об убывании этого вклада.

Теорема 2. *Если переменные X, Y, Z образуют марковскую цепь $X \to Y \to Z$, то*

$$I(X;Z) \leq I(X;Y)$$

и

$$I(X;Z) \leq I(Y;Z),$$

где $I(\bullet;\bullet)$ - взаимная информация.

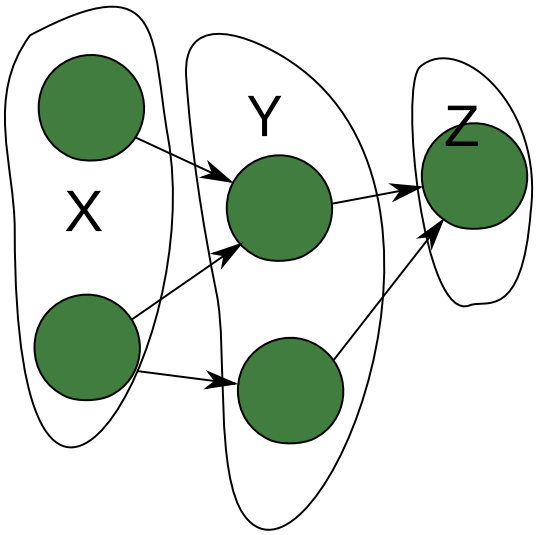


Рис. 20: Сведение к случаю марковской цепи.

Случай произвольной байесовской сети является несколько более сложным, однако некоторые выводы о зависимостях можно получить из этой теоремы путём группировки переменных (Рис. 20).

Таким образом можно ожидать, что изучение окрестности переменных, входящих в отвергнутые маргинальные распределения, позволит получить информацию для локализации противоречий. При этом, скорее всего, необходимо учитывать не только топологическое соседство, но и энтропию условных распределений.

7.2 Последовательное исключение V-структур.

Рассмотрим специальный случай байесовской сети (Рис. 21 (а)), в котором отсутствуют V-структуры вида $x \to y \leftarrow z$. Структура такого графа представляет из себя дерево с корнем в некоторой вершине y, все рёбра которого направлены от листьев к корню. Для вывода маргинального распределения в такой сети достаточно последовательно исключать листья. При этом размерность матриц, участвующих в таких операциях, ограничена размерностями исходных распределений. Назовём такую операцию *простым исключением сверху вниз*.

Теперь, в более общем случае, выделим подмножество предков A^y вершины y такое, что любой направленный путь из вершины $\in A^y$ проходит только по вершинам из A^y, причём V-структуры могут возникать только в начальных вершинах пути (Рис. 21 (b)). Заметим, что если фиксировать значения вершин, не имеющих предков в A^y, то задача вычисления маргинального распределения сводится к предыдущей. При этом совокупность таких маргинальных распределений при всевозможных фиксированных значениях задаёт, фактически, условное распределение (Рис. 21 (c)). Рассмотрев активные пути, проходящие через y, можно показать, что полученная сеть сохраняет d-разделённость.

Следует отметить, что эта процедура, в отличие от последовательного исключения переменных, не позволяет находить произвольные совместные распределения подмножеств переменных - достаточно рассмотреть задачу поиска совместного распределения на переменные x, z в сети, состоящей из единственной V-структуры $x \leftarrow y \rightarrow z$. Потенциально она может, однако, позволить ускорить общую процедуру и может быть полезной при решении других задач.

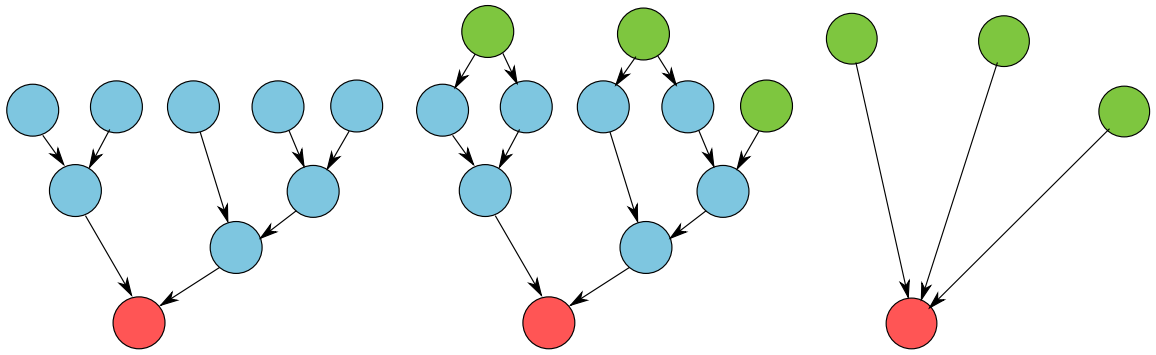


Рис. 21: Исключение v-структур.

7.3 Интервальное задание условных распределений

При экспертном задании байесовской сети точное численное определение распределений вряд ли возможно. Гораздо более удобным способом могло бы быть задание интервалов возможных значений вероятностей.

Будем считать, что все переменные распределения бинарные. Рассмотрим простейший граф байесовскй сети, состоящий из одной зависимой переменной y и её родителей $x_1, ..., x_n$ (Рис. 22). Будем считать, что для каждой из переменных x_i заданы ограничения $\underline{p}^1_{x_i} = \min(p(x_i) = 1) = 1 - \max(p(x_i) = 0) = 1 - \overline{p}^0_{x_i}$ и $\overline{p}^1_{x_i} = \max(p(x_i) = 1) = 1 - \min(p(x_i) = 0) = 1 - \underline{p}^0_{x_i}$. Аналогичные ограничения заданы на условные распределения $p_y^{1|X}$ и $p_y^{0|X}$. Требуется найти безусловные ограничения $\overline{p}^1_y$ и $\underline{p}^1_y$.

Для простоты ограничимся нахождением $\underline{p}^1_y$ - вторая задача аналогична. При фиксированных значениях вероятностей x_i соответствующие вероятности y находятся как $p^1_y = \sum_{X \in \{0,1\}^n} p_y^{1|X} \prod_{i=1}^n p_i^{X_i}$. Поскольку все множители неотрицательны, минимум будет достигаться при минимальных значениях всех $p_y^{1|X}$. Таким образом, задача сводится к нахождению $\min_{p_i^{X_i} \in [\underline{p}_i^{X_i}, \overline{p}_i^{X_i}] \forall i} \sum_{X \in \{0,1\}^n} \underline{p}_y^{1|X} \prod_{i=1}^n p_i^{X_i}$ с ограничениями $p^1_i + p^0_i = 1 \forall i$.

Эту задачу можно представить геометрически, как поиск точки внутри ограниченной подобласти единичного многомерного параллелепипеда, минимизирующей сумму взвешенных объемов параллелепипедов, определяемых проекциями этой точки на стороны исходного параллелепипеда. В качестве весов выступают $\underline{p}_y^{1|X}$, а X взаимно однозначно сопоставляется вершинам исходного параллелепипеда (Рис. 23).

Последовательно решая такие задачи, можно также осуществлять вывод сверху вниз в байесовской сети, представляемой деревом (Рис. 21(а)). Случай наличия V-структур $x \leftarrow y \rightarrow z$ оказывается более сложным, поскольку в каждой из последовательных задач оптимизации находятся конкретные значения вероятностей предыдущего уровня. Таким образом, при осуществлении вывода для дочерних вершин V-структуры могут быть выбраны разные значения вероятностей вершины-предка и

вывод уже не является точным. Вывод сверху-вниз, однако, может осуществляться при помощи последовательного исключения V-структур.

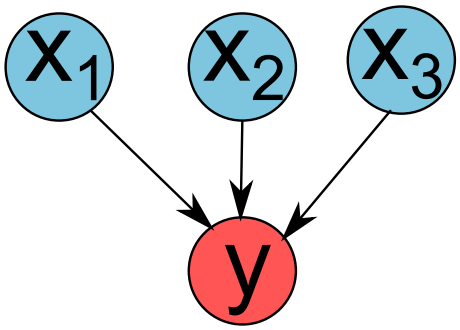


Рис. 22: Сеть с одной зависимой переменной.

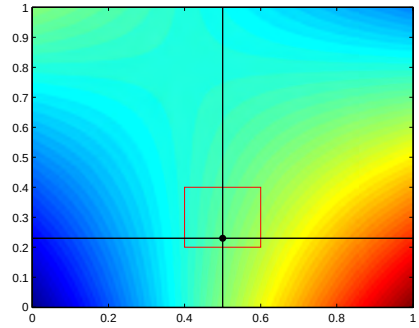


Рис. 23: Минимизация при выводе в интервально-заданной байесовской сети.

7.4 Обнаружение скрытой переменной - условная энтропия

Из обоснования метода исключения переменных с сохранением свойств байесовской сети можно видеть, насколько исключение переменных может усложнить структуру сети. При этом могут усложниться некоторые вычисления, а важные причинно-следственные связи оказаться скрытыми, поэтому поиск скрытых переменных представляет из себя отдельную важную задачу. На сегодняшний день существуют подходы к решению этой задачи на основе простых эвристик[5], а также методы, выводимые из сильных предположений[4].

Описанный в работе алгоритм исключения переменных может позволить более систематически подойти к локальной детекции скрытых переменных за счёт упрощения оценки результатов с использованием модельных данных. В качестве одной из локальных характеристик, способных потенциально оказаться полезными при решении этой задачи, можно рассмотреть условную энтропию переменных при условии переменных-родителей. Условная энтропия, как и обычная информационная энтропия, является мерой неопределённости и вычисляется по формуле

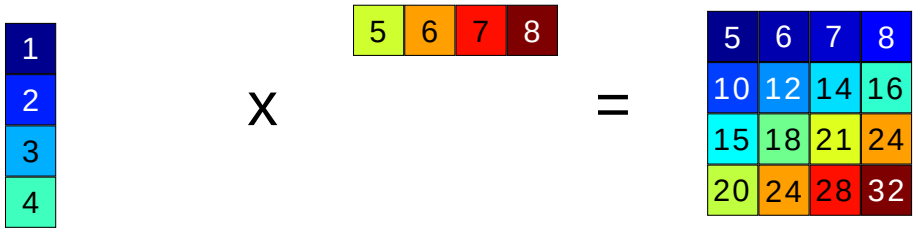
$$H(Y|X) = -\sum P(X) \sum P(Y|X) log_2 P(Y|X).$$

Рассмотрим пример. Пусть бинарные переменные $x_1, ..., x_n$ распределены независимо с вероятностью единицы, равной $\frac{1}{2}$. Пусть, далее $z = \oplus_{i=1}^{n} x_i$. В таком случае z принимает значения 0 и 1 с равными вероятностями, и, кроме того, можно показать, что исключение из рассмотрения любой из переменных x_i делает z независимой от оставшихся. В таком случае зависимость от родительских переменных принципиально нередуцируема к зависимости от их подмножеств. Значение условной энтропии при условии всех переменных x_i равно 0, для любого другого подмножества переменных значения условной энтропии равны $\frac{1}{2} log_2 \frac{1}{2}$.

Модифицируем этот пример. Пусть теперь $z = \oplus_{i=1}^{k} x_i \wedge \oplus_{i=k+1}^{n} x_i$. В этом случае можно упростить представление зависимости путём введения переменных $y_1 = \oplus_{i=1}^{k} x_i$ и $y_2 = \oplus_{i=k+1}^{n} x_i$. При этом для любого подмножества X переменных x_i, не включающего целиком переменных, задающих y_1 или y_2, условная энтропия по прежнему принимает значение $\frac{1}{2} log_2 \frac{1}{2}$. Если же включить в это множество все переменные $\{x_i | i \leq k\}$, условная энтропия уменьшается вдвое. На основании этого наблюдения можно предположить, что вычисление условной энтропии различных подмножеств родительских переменных для вершин байесовской сети может дать полезную информацию о местонахождении скрытых переменных.

7.5 Поиск скрытых переменных - ранги мультиматриц

Рассмотрим матричное умножение вектор-столбца на вектор-строку. Эту операцию можно представить при помощи копирования каждого из этих векторов вдоль размерностей, принадлежащих другому (Рис. 24). Каждый вектор при этом может рассматриваться как элемент расширенного пространства, размерности которого специальным образом сгруппированы. Часть значений не изменяется внутри групп размерностей, что и позволяет записать этот объект как вектор. Таким образом, мы вводим новый объект - назовём его мультиматрицей. Будем считать, что пары таких объектов всегда определяются в некотором общем объемлющем пространстве, и стандартные арифметические операции определены поэлементно. Можно заметить,



(a) стандартное матричное умножение

(b) матричное умножение с введением фиктивных групп размерностей.

Рис. 24: Матричное умножение векторов с разных точек зрения.

что при таком представлении нет причин, которые ограничивали бы нас во введении большего количества групп размерностей, и, в частности, значимых групп размерностей. По аналогии с обычными матрицами, можно определить ранг - будем считать, что он равен 1, если мультиматрица представима в виде произведения векторов, а ранг произвольной мультиматрицы равен минимальному количеству мультиматриц ранга 1, сумма которых равна данной.

С помощью мультиматриц удобно представлять байесовские сети дискретных распределений (Рис 25). Рёбра при этом задают значимые группы размерностей.

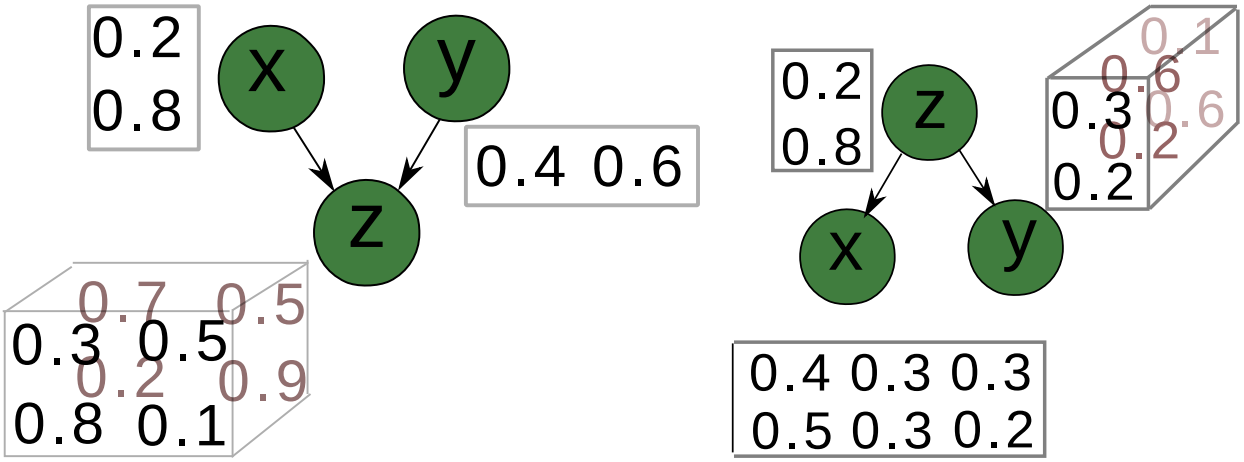


Рис. 25: Использование мультиматриц для задания байесовских сетей.

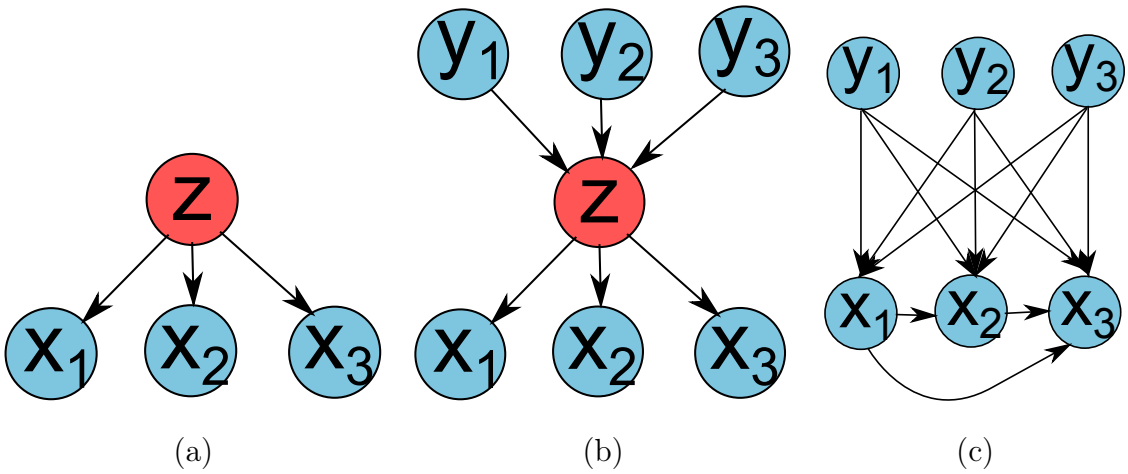


Рис. 26: Отдельные случаи исключения переменных в байесовских сетях.

Рассмотрим, что происходит при исключении переменных с точки зрения мультиматричного представления на примере сети с одним предком (Рис. 26 (b),25 (a)). Нетрудно видеть, что при каждом фиксированном значении переменной-родителя совместное распределение дочерних переменных представляет из себя мультиматрицу ранга 1. При исключении переменной-предка мы получаем, таким образом,

мультиматрицу, ранг которой ограничен сверху количеством значений, принимаемых родительской переменной.

В более общем случае на Рис. 25 (b),(c)) это свойство выполнено для условных распределений при заданных значениях исходных переменных-предков.

Изучение рангов различных маргинальных распределений может быть, таким образом, является другим возможным подходом для поиска скрытых переменных.

Список литературы

[1] Кузюрин Н.Н Берновский М.М. Случайные графы, модели и генераторы безмасштабных графов. *Труды Института системного программирования РАН*, 22:419–434, 2012.

[2] Ю.И.Медведев Г.И.Ивченко. Математическая статистика. *Москва*, 1984.

[3] Andrea Cerioli and Alessio Farcomeni. Error rates for multivariate outlier detection. *Computational Statistics & Data Analysis*, 55(1):544–553, 2011.

[4] Myung Jin Choi, Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning latent tree graphical models. *The Journal of Machine Learning Research*, 12:1771–1812, 2011.

[5] Gal Elidan, Noam Lotner, Nir Friedman, Daphne Koller, et al. Discovering hidden variables: A structure-based approach. In *NIPS*, volume 13, pages 479–485, 2000.

[6] Pietro Manzi Emanuela Barbini and Paolo Barbini. Bayesian approach in medicine and health management. In Dr. Alfonso Rodriguez-Morales, editor, *Current Topics in Public Health*. InTech, 2013.

[7] Alessio Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 2007.

[8] Mikko Koivisto and Kismat Sood. Exact bayesian structure discovery in bayesian networks. *J. Mach. Learn. Res.*, 5:549–573, December 2004.

[9] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[10] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[11] Erich Leo Lehmann and Joseph P Romano. *Generalizations of the familywise error rate*. Springer, 2012.

[12] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics*, pages 643–674. Springer, 2014.

[13] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons, 2011.

[14] Jegar Pitchforth and Kerrie Mengersen. A proposed validation framework for expert elicited bayesian networks. *Expert Systems with Applications*, 40(1):162–167, 2013.

[15] Armin Schwartzman and Xihong Lin. The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214, 2011.

[16] Ronen Talmon, Stéphane Mallat, Hitten Zaveri, and Ronald R Coifman. Manifold learning for latent variable inference in dynamical systems. Technical report, Yale University, 2014.

[17] Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Proceedings of the Twenty-first Conference on Uncertainty in AI*, pages 584–590, 2005.

[18] Kavishwar B Wagholikar, Vijayraghavan Sundararajan, and Ashok W Deshpande. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems*, 36(5):3029–3049, 2012.

[19] Wen Yan. *Identifying Structure and Semantics in Bayesian Network Inference*. PhD thesis, Faculty of Graduate Studies and Research, University of Regina, 2013.

[20] Raymond W Yeung. *Information theory and network coding*. Springer Science & Business Media, 2008.