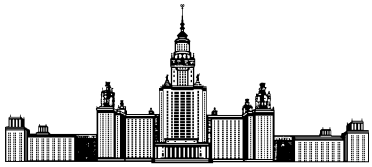
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Обзор методов восстановления ациклических зависимостей между случайными переменными»

Выполнил:
студент 3 курса 317 группы
Харациди Олег Александрович

Научный руководитель:
д.ф-м.н., профессор
Дьяконов Александр Геннадьевич

Москва, 2013

Содержание

1 Введение

В природе встречается множество стохастических событий, описываемых наборами наблюдаемых величин. Как правило, их несложно измерить. Однако очень часто бывает также необходимо выяснить взаимосвязь между ними, чтобы на основе этой информации сделать какие-либо выводы о причинах и следствиях наблюдаемых явлений.

Например, иногда нужно выяснить, как те или иные изменения окружающей среды влияют на здоровье человека, или же понять взаимосвязь между экономическими показателями, выделив причину и следствие.

При этом можно рассматривать как отдельные пары случайных переменных, так и целые наборы.

Иногда возникают трудности с проведением дополнительных экспериментов, позволяющих установить взаимосвязь. Например, они могут быть очень дорогостоящими или вовсе невозможными. Поэтому возникает потребность в методе, способном по ограниченной выборке значений восстанавливать зависимости между наблюдаемыми величинами.

Первые публикации на эту тему появились в 80-х годах. В 2000 году вышла книга [1], посвященная этой задаче. Авторы большинства статей, на основе которых написана данная работа, ссылаются на эту книгу как на полное собрание всех наработок по методам восстановления причинно-следственных связей на тот момент. С тех пор было предложено множество новых подходов, некоторые из которых изложены в данной работе.

Об актуальности задачи свидетельствует также тот факт, что на сайте Kaggle появилось посвященное ей соревнование [2]. Но на момент написания этого текста оно еще не закончено, и наработок участников в открытом доступе еще нет.

1.1 Определения и обозначения

Наблюдаемые переменные будем также называть случайными переменными или случайными величинами и, если не оговорено иное, обозначать $x_1, ..., x_n$. Функции зависимостей переменных будут обозначаться через f_i или f, а шум – через e_i или e.

Зависимости также будем называть причинно-следственными связями, а обозначать зависимость y от x записью вида $x \to y$.

Графом зависимостей набора случайных величин $x_1, ..., x_n$ назовем ориентированный граф, в котором вершины соответствуют случайным величинам, а ребра – зависимостям между ними.

Граф зависимостей будем обозначать как G.

Основная задача, которой посвящена эта работа, заключается в том, чтобы по заданному набору наблюдаемых случайных переменных $x_1, ..., x_n$ восстановить граф зависимостей G.

Если не оговорено иное, будет предполагаться, что случайные переменные заданы выборками значений из их совместного распределения.

2 Предобработка данных

Прежде чем перейти непосредственно к рассмотрению методов решения поставленной задачи, отметим, что работать с данными в исходном их виде не всегда удобно. Во-первых, во входных данных может быть большой разброс вещественных значений. Для методов, рассмотренных ниже, это довольно существенная проблема, поскольку они содержат много численно неустойчивых опраций. Во-вторых, в некачественных данных могут быть выбросы, которые также помешают получению требуемого результата. Наконец, хорошая нормировка выборок может «подогнать» данные под необходимую модель без потери информации о причинно-следственных связях.

Итак, полезными на практике оказываются следующие варианты препроцессинга данных:

- Фильтрация выбросов.
  Самый простой способ - удалить из выборки самые большие и самые маленькие значения. При этом нужно как можно точнее оценить долю выбросов. Если убрать не все выбросы, то проблема останется. Если же, наоборот, «выкинуть» слишком много, то данные могут исказиться, что приведет к плохим результатам.

- Центрирование.
  С центрированными данными удобнее работать, поскольку не нужно учитывать смещение. В рамках задачи восстановления связей центрирование, очевидно, не приведет к потере информации. При центрировании бывает полезно вычитать из выборки как ее среднее, так и медиану.

- Нормирование.
  Все исходные данные можно пронормировать либо на корень из дисперсии, либо на разность максимального и минимального значений. Нормирование также не приводит к потере информации. Чаще всего оно применяется после двух предыдущих пунктов.

- Дискретизация значений.
  Происходит потеря информации, но упрощаются некоторые операции, такие как, например, оценивание энтропии. Частоту дискретизации можно варьировать.

- Нумерация значений.
  Если занумеровать все встречающиеся значения одной из величин в порядке возрастания, а затем заменить их на свои номера, то изчезнут все «пробелы» в распределении значений. Причем разумно одинаковым значениям давать одинаковый номер, равный среднему встречающемуся среди них номеру. Нумерацию логично применять после дискретизации.

3 Модели с аддитивным шумом

Одна из наиболее простых моделей, используемых для восстановления нелинейных зависимостей, – модель с аддитивным шумом (Additive Noise Model, ANM). Она описывается следующим соотношением:

$$x_i = f_i(x_{pa(i)}) + e_i,$$

где x_i – случайная переменная, $x_{pa(i)}$ – множество случайных переменных, от которых она зависит, f_i – функция зависимости, а e_i – случайный шум.

В рамках данной модели делается очень важное предположение о независимости x_i и e_i. Рассмотрим два метода, основывающихся на этом предположении.

3.1 Нелинейная модель

Метод, предложенный в [3], сначала ограничивается рассмотрением пары случайных величин, а затем обобщает решение на случай ориентированного ациклического графа.

Итак, рассмотрим модель с аддитивным шумом для двух случайных переменных x и y:

$$y = f(x) + e$$

В общем случае задача восстановления зависимости для данной модели неразрешима. Например, если случайные переменные x, y и e имеют стандартное нормальное распределение с нулевым математическим ожиданием, n независима и с x, и с y, а функция f - тождественная (т.е. $f(x) = x$), то одновременно выполнены два равенства:

$$y = x + e$$

$$x = y - e,$$

то есть модель допускает зависимости в обе стороны. Однако в действительности можно считать, что таких случаев не так много. Предположим, что оба соотношения выполнены, плотности распределений $p(x)$, $p(y)$ и $p(e)$ строго положительны и, кроме того, все три плотности и функция f достаточно гладкие. В рамках этих вполне естественных ограничений можно показать [3], что при фиксированных $p(x)$ и f функция $\log p(y)$ должна являться решением обыкновенного дифференциального уравнения третьего порядка, то есть принадлежать некому трехмерному линейному многообразию. В то же время, пространство всех допустимых функций $\log p(y)$ имеет бесконечную размерность. Таким образом, можно неформально утверждать, что «в большинстве случаев» рассматриваемая задача однозначно разрешима.

Перейдем непосредственно к восстановлению зависимости. Сначала необходимо проверить, что наблюдаемые переменные x и y зависимы. Для этого предлагается использовать любой предназначенный для этой цели статистический критерий. Если

переменные оказываются независимыми, то дальнейших проверок не требуется. Если же x и y зависимы, то нужно проверить возможность зависимости y от x:

$$y = f(x) + e$$

Для этого необходимо с помощью метода нелинейной регрессии построить приближение $\hat{f}$ функции f, а затем с помощью статистического критерия проверить независимость шума $\hat{e} = y - \hat{f}(x)$ и x. После этого провести аналогичную проверку зависимости x от y.

Если переменные x и y статистически независимы, то между ними нет связи. Далее, если при проверке обоих видов зависимости получен положительный результат, то метод оказался неспособен определить дать ответ на поставленную задачу. Если оба результата - отрицательные, то зависимость между x и y имеет более сложную форму, и в рамках данной модели задача считается неразрешимой. Иначе, если подтвердился ровно один из видов зависимости, то именно он и принимается в качестве ответа.

Очень просто сделать обобщение этого метода на случай проверки зависимости одной переменной y от множества переменных $x_1, ..., x_n$:

$$y = f(x_1, ..., x_n) + e$$

Отличие от описанного выше метода в том, что здесь необходимо n тестов на независимость (для каждой пары (y, x_i)), а нелинейная регрессия производится сразу по n переменным $x_1, ..., x_n$.

Дальнейшее обобщение на случай бóльшего числа переменных делается следующим образом. Рассматриваются все возможные ориентированные ациклические графы зависимостей, и каждый из них проверяется отдельно. Для графа G_j проверяется возможность зависимости каждой вершины от каждой из ее родителей. Если все проверки подтвердились, то граф G_j полагается допустимым, иначе он отвергается.

По утверждению авторов данного метода, на практике это обобщение дает сравнительно неплохие результаты при общем количестве переменных, не превосходящем 7.

Одним из недостатков предложенного алгоритма является возможность переобучения при восстановлении функции f с помощью нелинейной регрессии. Если допустить сложный вид зависимости, то можно ошибочно принять неверную модель. Напротив, если сильно ограничить множество всех допустимых функций f, то метод может отвергнуть верную модель.

Кроме того, в силу множественного применения эвристик и статистических критериев, алгоритм может отвергнуть все рассматриваемые графы причинно-следственных связей.

3.2 Линейная негауссовская ациклическая модель

Принципиально другой подход - линейная негауссовская ациклическая модель (Linear Non-Gaussian Acyclic Model, LiNGAM) [4]. Она используется в рамках следующих предположений:

1. Существует такая перестановка наблюдаемых переменных $x_{k(1)}, ..., x_{k(n)}$, что $\forall i < j$ $x_{k(i)}$ не зависит от $x_{k(j)}$. Другими словами, отсутствуют циклические зависимости.

2. Все зависимости линейные, а шум – аддитивен:

$$x_{k(i)} = \sum_{j<i} b_{ij} x_{k(j)} + e_{k(i)} + c_{k(i)}$$

3. Все e_i независимы и имеют абсолютно непрерывное негауссово распределение.

Третье предположение играет ключевую роль в предложенной модели, которая станет ясна позже.

Будем считать, что все x_i имеют нулевые средние значения (в противном случае на стадии предобработки данных отцентрируем все x_i). Исходя из описанных предположений, можно записать следующее соотношение для векторов $\mathbf{x}$ и $\mathbf{e}$:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

или, что то же самое,

$$\mathbf{x} = \mathbf{A}\mathbf{e},$$

где $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$.

Отметим, что при перестановке пар случайных величин x_i будут меняться местами соответствующие пары строк и столбцов матриц $\mathbf{A}$ и $\mathbf{B}$. При применении к $\{x_i\}$ перестановки k получим нижнетреугольную матрицу $\mathbf{B}$ и, соответственно, нижнетреугольную матрицу $\mathbf{A}$, причем диагональ $\mathbf{A}$ будет единичная, а диагональ $\mathbf{B}$ – нулевая. Однако задача состоит как раз в поиске перестановки k. Следовательно, вместо этого можно сначала вычислить матрицу $\mathbf{A}$, а затем найти перестановку пар ее строк и столбцов (то есть одну перестановку, применяемую и к строкам, и к столбцам), которая приводит ее к диагональной форме. Рассмотрим каждую из двух подзадач отдельно.

Вычисление матрицы $\mathbf{A}$ производится с помощью анализа независимых компонент (Independent Components Analisys, ICA) [5, 6], описание которого выходит за рамки темы данной работы. Одной из задач анализа независимых компонент является поиск невырожденного линейного преобразования заданного случайного вектора, переводящее его в вектор с как можно менее зависимыми компонентами. Эта задача возникает, в частности, в обработке сигналов при выделении звуков отдельных объектов (например, голосов разных людей). В нашем же случае применение ее решения состоит в непосредственном вычислении матрицы $\mathbf{A}$ по заданному случайному вектору $\mathbf{x}$.

Одна из главных особенностей анализа независимых компонент состоит в том, что задача разделения на гауссовские компоненты в рамках него неразрешима. Именно этим обусловлено наше предположение о негауссовом распределении e_i.

Еще одна особенность состоит в том, что компоненты e_i, а значит, и столбцы матрицы $\mathbf{A}$, определяются с точностью до перестановки и домножения на константу.

Это позволяет преобразовать $\mathbf{A}$ (переставить столбцы и затем пронормировать их на соответствующие диагональные элементы) так, чтобы на диагонали стояли единицы. Если бы $\mathbf{A}$ была вычислена точно, то перестановка столбцов, после которой на диагонали не окажется нулей, была бы единственной в силу ацикличности зависимостей. Однако, поскольку $\mathbf{A}$ вычисляется с приближенно, такое преобразование может быть неоднозначным. Поэтому авторами метода предлагается искать перестановку, минимизирующую величину $\sum_i \frac{1}{|\mathbf{A_{ii}}|}$.

Далее, для нахождения перестановки k также необходимо воспользоваться некоторыми эвристическими соображениями. Снова, если бы матрица $\mathbf{A}$ была вычислена без погрешности, то поиск перестановки ее строк и столбцов, приводящей ее к нижнетреугольному виду, не представлял бы трудности. Проблему неточности авторы предлагают решать следующим образом. Сначала вычислим матрицу $\mathbf{B} = \mathbf{I} - \mathbf{A}^{-1}$, а затем найдем перестановку соответствующих строк и столбцов, минимизирующую функционал $\sum_{i \leq j} \mathbf{B}_{ij}^2$, приведя таким образом матрицу $\mathbf{B}$ к нижнетреугольной матрице с нулевой диагональю. Найденная перестановка будет равна k^{-1}, из которой тривиальным обращением может быть получена искомая перестановка k.

Мы ограничились рассмотрением базового варианта алгоритма LiNGAM, на практике же предлагается использовать его модификацию, в которую входит еще несколько эвристик, повышающих его точность и скорость [4].

4 Байесовский подход

Рассмотренные модели имеют один общий недостаток. Нелинейная модель основывается на необратимости функции зависимости f, а LiNGAM использует существование шума с достаточно большой дисперсией. Рассмотрим случай, для которого ни одно из ограничений не выполняется:

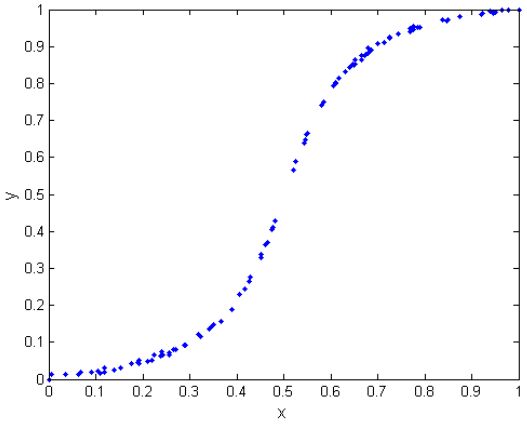


Рис. 1: $x \to y$

Ни одна из предложенных выше моделей не сможет дать ответ, зависит ли x от y, или же y зависит от x. Более того, на первый взгляд кажется, что такой ответ дать в принципе невозможно.

Однако существует подход, предоставляющий решение и в этом случае. Рассмотрим совместное распределение $p(x, y)$ двух наблюдаемых переменных x и y. Запишем два формальных равенства:

$$p(x, y) = p(x)p(y|x)$$
$$p(x, y) = p(y)p(x|y)$$

Нетрудно заметить, что в рамках данной задачи они имеют вполне определенную интерпретацию: первое описывает механизм порождения пары (x, y) в случае $x \to y$, а второе – в случае $y \to x$. Можно предположить, что имеет место та из двух порождающих моделей, которая описывается наиболее простым способом. Но для этого необходимо формализовать понятие «простоты» разложения

$$p(x, y) = p(y)p(x|y)$$

Именно это предлагается сделать авторами двух рассмотренных ниже методов.

Подробное описание методов в этой работе не приводится, изложены лишь основные идеи и принципы.

4.1 Информационно-геометрический принцип

Обозначим «сложность» модели $x \to y$ через $C_{x\to y}$. Информационно-геометрический принцип (Information Geometric Causal Inference, IGCI) [7] предлагает несколько способов определения этой величины. Рассмотрим один из них.

Обозначим через $\mathcal{E}_x$ и $\mathcal{E}_y$ классы «простых» распределений случайных величин x и y соответственно. Введем также функцию расстояния $\rho(p, q) \geq 0$ на множестве абсолютно непрерывных распределений и определим расстояние между p и множеством $\mathcal{E}$ как $\rho(p, \mathcal{E}) = \inf_{p'\in\mathcal{E}} \rho(p, p')$. Величину $C_{x\to y}$ можно определить следующим образом:

$$C_{x\to y} = \rho(p_x, \mathcal{E}_y) - \rho(p_y, \mathcal{E}_x)$$

Нетрудно видеть, что в этом случае $C_{x\to y} = -C_{y\to x}$, поэтому достаточно ограничиться вычислением $C_{x\to y}$. После этого, если $C_{x\to y} < 0$, то принимается гипотеза $x \to y$, иначе – $y \to x$.

Для полной определенности осталось ввести функцию ρ и классы $\mathcal{E}_x$ и $\mathcal{E}_y$. Авторами данного метода предлагается использование в качестве функции расстояния дивергенции Кульбака-Лейблера:

$$\rho(p, q) = D_{KL}(p, q) = \int \log \frac{p(t)}{q(t)} p(t)dt$$

При определенном выборе $\mathcal{E}_x$ и $\mathcal{E}_y$ и предположении, что $P(x \in [0, 1]) = P(y \in [0, 1]) = 1$ можно получить следующее выражение для $C_{x\to y}$:

$$C_{x\to y} = \mathcal{H}(p_y) - \mathcal{H}(p_x),$$

где $\mathcal{H}(p) = \int p(t) \log p(t)dt$ - дифференциальная энтропия.

Воспользуемся выборочной оценкой для дифференциальной энтропии [8]:

$$\hat{\mathcal{H}}(p_x) = \psi(m) - \psi(1) + \frac{1}{m-1}\sum_{i=1}^{m-1}\log(x_{i+1} - x_i),$$

где $\psi(x) = \frac{d}{dx}\log\Gamma(x)$, а $x_1, ..., x_m$ - упорядоченная по неубыванию выборка значений случайной величины x. Здесь полагается $\log 0 = 0$. Получим приближение для $C_{x\to y}$:

$$\hat{C}_{x\to y} = \frac{1}{m-1}\sum_{i=1}^{m-1}\log\frac{y_{i+1} - y_i}{x_{i+1} - x_i}.$$

Здесь предполагается, что выборки $x_1, ..., x_m$ и $y_1, ..., y_m$ упорядочены по неубыванию независимо.

Оценка сложности модели через разность энтропий имеет простую интерпретацию: независимая переменная должна быть распределена более равномерно, чем зависимая. Применительно к примеру, изображенному на Рис. 11, получим $\hat{C}_{x\to y} = -0.3967 < 0$, то есть y зависит от x. Плотности распределений переменных в этом примере выглядят следующим образом:

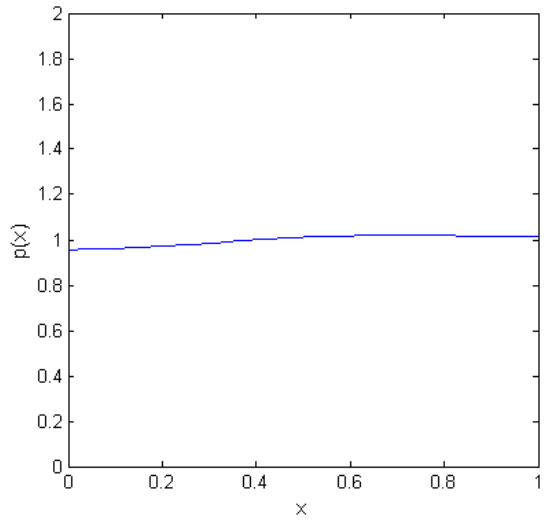


Рис. 2: график $p(x)$

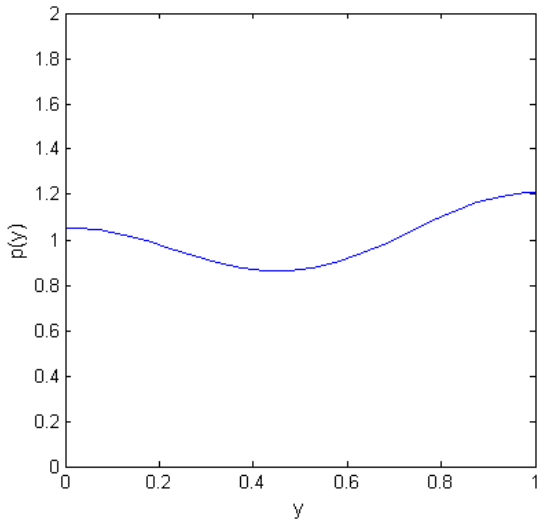


Рис. 3: график $p(y)$

Видно, что распределение переменной x более равномерное.

Итак, показан наиболее простой вариант использования информационно-геометрического принципа. Авторы метода предлагают и другие вариации данного метода, приводящие к менее тривиальным оценкам для $C_{x\to y}$. Стоит отметить, что абстрактность модели позволяет даже вывести аналогичный алгоритм для многомерных случайных векторов x и y.

4.2 Модель со скрытыми переменными

Байесовский подход позволяет также пойти дальше и использовать еще одну идею для формализации понятия «простоты» разложения совместного распределения $p(x, y)$.

Метод, использующий модель со скрытыми переменными [9], предполагает рассмотрение шума как отдельной скрытой переменной и предлагает для описания связи вида $x \to y$ модель, обладающую следующими свойствами:

1. Зависимость может быть произвольной:

$$y = f(x, e),$$

   где e - шум, в который входят все другие переменные, от которых зависит y.

2. x и e независимы.

3. Распределение x и функция f никак не зависят друг от друга. Это свойство здесь вводится неформально, и будет использовано ниже для вывода алгоритма. Его смысл в том, что в природе механизм, порождающий y из x, обычно никак не связан с самой переменной x.

4. Шум e имеет стандартное нормальное распределение: $e \sim \mathcal{N}(0, 1)$. Это свойство не накладывает ограничений на модель, поскольку любая случайная величина e представима в виде $e = g(\bar{e})$, где $\bar{e} \sim \mathcal{N}(0, 1)$, поэтому $y = f(x, g(\bar{e})) = \bar{f}(x, \bar{e})$.

Нетрудно заметить, что обе рассмотренные выше модели с аддитивным шумом подчиняются свойствам 1 и 2. При этом, чтобы избежать неразрешимости задачи, они накладывают довольно жесткие ограничения на функцию f, сужая таким образом область применения метода. В данном же подходе проблему неразрешимости предлагается устранить с помощью более слабых предположений о виде функции f и введения свойств 3 и 4.

Основная идея заключается в том, чтобы оценить «сложность» моделей $x \to y$ и $y \to x$, сравнив соответствующие правдоподобия $p(X, Y|x \to y)$ и $p(X, Y|y \to x)$. Ограничимся описанием алгоритма вычисления первого правдоподобия, второе может быть вычислено аналогично.

Итак, для зависимости $x \to y$, с учетом свойств 1-4, несложно показать, что имеет место следующее разложение:

$$p(X, Y) = p(Y|X)p(X) =$$

$$\left[\int \left(\prod_{i=1}^{m} p(x_i, \boldsymbol{\theta}_X)\right) p(\boldsymbol{\theta}_X) d\boldsymbol{\theta}_X\right] \left[\int \left(\prod_{i=1}^{m} \delta(y_i - f(x_i, e_i)) p_{\mathbf{e}}(e_i)\right) d\mathbf{e}\ p(f|\boldsymbol{\theta}_f) df\ p(\boldsymbol{\theta}_f) d\boldsymbol{\theta}_f\right],$$

где δ - дельта-функция Дирака, $\boldsymbol{\theta}_X$ - некоторые параметры совместного распределения выборки X, а $\boldsymbol{\theta}_f$ - некоторые параметры распределения для функции зависимости f.

Далее предлагается ввести распределение $p(x_i)$ как смесь гауссиан, ввести функцию f как случайный гауссовский процесс, а также наложить априрные распределения их на параметры $\boldsymbol{\theta}_X$ и $\boldsymbol{\theta}_f$. После этого, с помощью указанного выше разложения можно вывести приближенные оценки для $p(X)$ и $p(Y|X)$ и получить таким образом оценку для $p(X, Y)$. Подробный вывод этих шагов здесь не будет приведен по причине его громоздкости.

Отличительной особенностью данной модели является то, что функция «сложности» ассиметрична. То есть, в отличие от информационно-геометрического метода, соотношение $C_{x\to y} = -C_{y\to x}$ не выполняется. По мнению авторов модели со скрытыми переменными, это существенное преимущество. В доказательство они приводят в статье результаты экспериментов, в которых данный метод превзошел по качеству IGCI, а также LiNGAM и некоторые другие модели с аддитивным шумом.

Список литературы

[1] P. Spirtes, C. Glymour, R. Scheinesh. Causation, Prediction, and Search. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/learn-43/lib/photoz/.g/scottd/fullbook.pdf

[2] http://www.kaggle.com/c/cause-effect-pairs

[3] P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, B. Scholkopf. Nonlinear causal discovery with additive noise models. http://www.cs.helsinki.fi/u/phoyer/papers/pdf/hoyer2008nips.pdf

[4] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. http://www.cs.helsinki.fi/u/phoyer/papers/pdf/JMLR06.pdf

[5] P. Comon. Independent component analysis, A new concept? http://dev.ce.ucsb.edu/courses/ECE594/594C_F10Madhow/comon94.pdf

[6] A. Hyvärinen, E. Oja. Independent Component Analysis: Algorithms and Applications. http://mlsp.cs.cmu.edu/courses/fall2012/lectures/ICA_Hyvarinen.pdf

[7] D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, B. Schölkopf. Information-geometric approach to inferring causal directions. http://cs.ru.nl/j̃orism/articles/ai2012.pdf

[8] A. Kraskov, H. Stogbauer. Estimating Mutual Information. http://arxiv.org/abs/cond-mat/0305641v1

[9] J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. http://webdav.tuebingen.mpg.de/causality/NIPS2010-Mooij.pdf