Генеративный причинно-следственный подход к анализу данных ИМК

Владимиров Э.А.

Московский физико-технический институт

Научный руководитель: д. ф.-м. н. В. В. Стрижов

2023

Причинно-следственный анализ

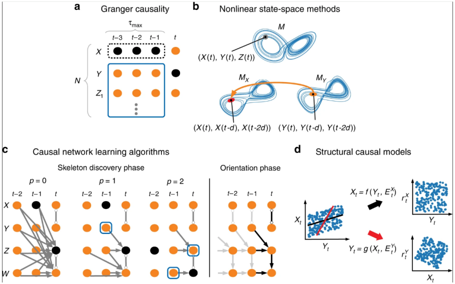
Проблема

- Традиционные методы (корреляция, линейная регрессия) неадекватны для сложных нелинейных связей
- Данные имеют высокую размерность, что усложняет поиск причинно-следственных связей
- Зависимости между переменными могут изменяться во времени

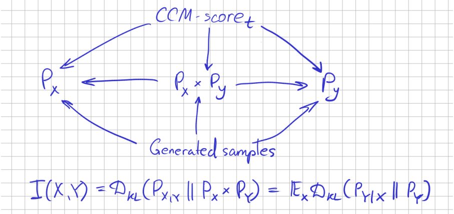
Решение

Построить устойчивую и интерпретируемую форму вероятностного анализа причинного влияния $\mathbf{X} \rightarrow \mathbf{Y}$

Различные подходы к поиску связей



Существующие подходы



Предлагаемый подход

Постановка задачи

Пусть $\mathbf{X}(t) = \{X_1(t), X_2(t), \dots, X_{n_x}(t)\}$ и $\mathbf{Y}(t) = \{Y_1(t), Y_2(t), \dots, Y_{n_y}(t)\}$ — два набора многомерных временных рядов, наблюдаемых в моменты времени $t = 1, \dots, T$.

Необходимо определить направленные причинные связи:

1. $X_i(t-\tau) \to Y_j(t)$ для $i = 1, \dots, n_x$, $j = 1, \dots, n_y$, и лагов $\tau \geqslant 0$,
2. $Y_j(t-\tau) \to X_i(t)$ для $i = 1, \dots, n_x$, $j = 1, \dots, n_y$, и лагов $\tau \geqslant 0$.

Предполагаем, что многомерные временные ряды $\mathbf{X}(t)$ и $\mathbf{Y}(t)$ генерируются следующим образом:

$$X_i(t) = f_i(\mathrm{Pa}_{X_i}(t), \varepsilon_{X_i}(t)),$$
$$Y_j(t) = g_j(\mathrm{Pa}_{Y_j}(t), \varepsilon_{Y_j}(t)),$$

где:

- $\mathrm{Pa}_{X_i}(t) \subseteq \{Y_1(t-\tau), \dots, Y_{n_y}(t-\tau)\}$ — множество родителей переменной $X_i(t)$ из $\mathbf{Y}$,
- $\mathrm{Pa}_{Y_j}(t) \subseteq \{X_1(t-\tau), \dots, X_{n_x}(t-\tau)\}$ — множество родителей переменной $Y_j(t)$ из $\mathbf{X}$,
- f_i и g_j — детерминированные функции, описывающие зависимость,
- $\varepsilon_{X_i}(t)$ и $\varepsilon_{Y_j}(t)$ — шумовые компоненты.

Оптимизационная задача:

$$\min_{G_{XY}, G_{YX}} \mathcal{L}(\mathbf{X}, \mathbf{Y} \mid G_{XY}, G_{YX}) + \lambda_1 \mathcal{R}(G_{XY}, G_{YX}) + \lambda_2 \mathcal{T}(G_{XY}, G_{YX}),$$

где:

- G_{XY} — граф зависимостей $X_i \to Y_j$,
- G_{YX} — граф зависимостей $Y_j \to X_i$,
- $\mathcal{L}(\mathbf{X}, \mathbf{Y} \mid G_{XY}, G_{YX})$ — правдоподобие наблюдаемых данных с учетом графов G_{XY} и G_{YX},
- $\mathcal{R}(G_{XY}, G_{YX})$ — регуляризатор, штрафующий за сложность графов,
- $\mathcal{T}(G_{XY}, G_{YX})$ — штраф за избыточную изменчивость графов во времени.

Independent Component Analysis

Предположим, что $\mathbf{X}(t)$ образуется из нескольких скрытых источников $\mathbf{S}(t) \in \mathbb{R}^{d_S}$:

$$\mathbf{X}(t) \; = \; A\,\mathbf{S}(t), \quad A \in \mathbb{R}^{d_X \times d_S}.$$

Каждая компонента $S_k(t)$ предполагается статистически независимой от других:

$$p(\mathbf{S}) = \prod_{k=1}^{d_S} p\big(S_k\big).$$

Задача оптимизации: Найти обратную матрицу $\widehat{A}^{-1}$, дающую

$$\widehat{\mathbf{S}}(t) \; = \; \widehat{A}^{-1}\,\mathbf{X}(t),$$

чтобы минимизировать взаимную информацию между компонентами $\widehat{S}_k(t)$.

$$\mathrm{MI}\big(\widehat{\mathbf{S}}(t)\big) \; \approx \; \sum_{k=1}^{d_S} H\big(S_k\big) - H\Big(\sum_k S_k\Big),$$

Convergent Cross Mapping

Теневое вложение (delay embedding):

$$M_{X,t} = \big(X_t,\, X_{t-\tau},\, \ldots,\, X_{t-(E-1)\tau}\big) \;\in\; \mathbb{R}^E,$$

где E — размерность вложения, τ — временной лаг. Аналогично задаётся $M_{Y,t} = \big(Y_t,\; Y_{t-\tau},\, \ldots\big)$.

Реконструкция:

$$\widehat{Y}_t = \sum_{i=1}^{k} w_i\, Y_{n_i},$$

здесь n_i — индексы ближайших соседей точки $M_{X,t}$ в пространстве M_X, а w_i — веса, зависящие от расстояния до $M_{X,t}$.

Критерий причинности:

$$\rho_{X\to Y} = \mathrm{corr}\Big(\{\widehat{Y}_t\}, \{Y_t\}\Big).$$

Если при увеличении размера “библиотеки” (количества доступных точек) значение $\rho_{X\to Y}$ *возрастает*, считается, что $\mathbf{X}(t)$ действительно влияет на $\mathbf{Y}(t)$.

Probabilistic CCM

Идея: Вместо единственного прогноза $\widehat{Y}_t$ рассматривается *полное условное распределение*

$$p_L\big(Y_t \mid M_{X,t}\big),$$

оценённое по выборке размера L. Ближайшие соседи в пространстве M_X позволяют построить *вероятностную аппроксимацию* (например, ядерным методом):

$$p_L\big(y \mid M_{X,t}\big) \;=\; \frac{1}{Z_t} \sum_{i \in N_L(t)} K\Big(y - Y_{n_i}\Big),$$

где $K(\cdot)$ — ядро (например, гауссово), $N_L(t)$ — множество соседей точки $M_{X,t}$, а Z_t — нормировочная константа.

Оценка причинности как MI:

$$I_L(X \to Y) \;=\; \mathbb{E}_{\hat{M_{X,t}}} D_{\mathrm{KL}}\Big(p_L\big(Y_t \mid \hat{M_{X,t}}\big) \,\big\|\, p\big(Y_t\big)\Big),$$

Сходящееся свойство: При $L \to \infty$ (при достаточно плотном покрытии пространства)

$$p_L\big(Y_t \mid M_{X,t}\big) \;\to\; p\big(Y_t \mid M_{X,t}\big),$$

Предлагаемый метод

1. **Независимый анализ компонент (ICA).**
   Для исходных ЭЭГ-данных $\mathbf{X}_{\mathrm{raw}}(t) \in \mathbb{R}^{d_X}$ получаем независимые компоненты:
   $$\widehat{\mathbf{S}}(t) = \widehat{A}^{-1}\,\mathbf{X}_{\mathrm{raw}}(t).$$

2. **Построение эмбеддингов.**
   Для каждого времени t формируем вектор:
   $$M_{X,t} = \big(\widehat{\mathbf{S}}(t), \widehat{\mathbf{S}}(t-\tau), \ldots, \widehat{\mathbf{S}}(t-(E-1)\tau)\big),$$

3. **Оценка причинно-следственных связей.**
   В полученном пространстве $(M_{X,t},\, M_{Y,t})$ определяем меру влияния $\mathbf{X} \to \mathbf{Y}$, вычисляя:
   $$\gamma(t) = \mathrm{Prob} - \mathrm{CCM}\big(M_{X,t},\, M_{Y,t}\big).$$
   Результат — временной ряд $\{\gamma(t)\}$, отражающий динамику влияния $X \to Y$.

Риманова постановка задачи

Причинно-следственная связь — вероятность push-forward-a или диффеоморфизма.
Основная проблема — построение многообразий
Возможные варианты:

- PINN
- Riemannian space of covariance matrices

Вычислительный эксперимент на данных ЭЭГ - ИИМ

Данные

У 25 участников были записаны показания ЭЭГ, ИИМ, МРТ во время игры в настольный теннис. С каждым участником было сыграно 4 сессии, длительность каждой из них составляет 7-10 минут.