Вероятностные модели причинно-следственных связей в сходящихся отображениях

Mark Ikonnikov, Vadim Strijov

19 декабря 2025 г.

1 Abstract

На сегодняшний день работа с мультимодальными данными набирает всё большую популярность: учет взаимосвязей между ними улучшает качество предсказания. В статье мы предлагаем новую формулировку причинно-следсвенных связей применительно к методу сходящихся отображений. Дополнительно мы обсуждаем связь CCA с механизмами внимания и показываем, что attention-подобные конструкции, как и классический CCA, аппроксимируют ассоциативную структуру данных, но не идентифицируют каузальный эффект. Предложенная формулировка закладывает основу для интеграции причинного вывода в методы мультимодального представления, включая динамические системы и текстовые данные.

keywords : CCA, CCM, Attention, CI.

2 Introduction

Современные системы анализа данных всё чаще работают с несколькими разнородными источниками информации, формируя мультимодальные представления, объединяющие текст, изображения, временные ряды и сенсорные данные. Такие постановки возникают в задачах здравоохранения, робототехники, аффективных вычислений и анализа сложных динамических систем [**?**]. Эффективная интеграция модальностей требует методов, способных выявлять согласованную структуру между различными представлениями данных.

Одним из классических инструментов мультимодального анализа является канонический корреляционный анализ (CCA) [**?**], предназначенный для поиска линейных проекций двух наборов переменных, максимизирующих корреляцию в общем латентном пространстве. CCA и его обобщения — kernel CCA, deep CCA и probabilistic CCA — широко применяются в задачах сопоставления модальностей и кроссмодального обучения [**? ?**]. Однако фундаментальным ограничением этих методов является их ассоциативная природа: максимизация корреляции не различает прямые причинные эффекты и зависимости, возникающие вследствие скрытых общих факторов.

С другой стороны, в анализе динамических систем были предложены методы, ориентированные на выявление направленных зависимостей, в частности Convergent Cross Mapping (CCM), основанный на реконструкции аттракторов фазового пространства. CCM способен отличать корреляции, обусловленные общей динамикой, от направленного влияния, однако не имеет явной вероятностной интерпретации и плохо масштабируется на высокоразмерные и мультимодальные данные.

CCA выявляет общую скрытую структуру между переменными и, следовательно, чувствителен к искажающим эффектам, вызванным ненаблюдаемыми общими причинами. В отличие от этого, метод CCM (Convergent Cross Mapping) основан на реконструкции динамических аттракторов и не выявляет чисто ассоциативные зависимости, лишенные направленной динамической связи.

Параллельно с этим в последние годы ключевую роль в моделировании сложных зависимостей играют механизмы внимания (attention), лежащие в основе трансформерных архитектур. Несмотря на выразительность, attention-механизмы также основаны на ассоциативных мерах сходства и не обеспечивают каузальной интерпретации получаемых зависимостей.

В данной работе мы ставим своей целью формально связать канонический корреляционный анализ с аппаратом причинно-следственного вывода. Используя структурные причинные модели и оператор интервенции $\mathrm{do}(X)$, мы показываем, что даже в линейно-гауссовской постановке канонические корреляции в общем случае не совпадают с каузальным эффектом. Это наблюдение мотивирует введение причинной версии probabilistic CCA, в рамках которой интервенционный эффект выражается через явный линейный оператор, независимый от скрытых конфаундеров.

3 Связь CCA - attention

3.1 CCA: Canonical-Correlation analysis - transformer

Мы продолжаем расширять область применимости CCA и представляем способ встраивания его в Attention.
Canonical Correlation Analysis (CCA) - стандартный инструмент для выявления линейных зависимостей между двумя наборами данных .[1] Пусть нам дано множество векторов $X \in \mathbb{R}^{n_1 \times m}$ и $Y \in \mathbb{R}^{n_2 \times m}$, где m - количество векторов. Задача CCA - найти такие афинные преобразования $\mathbf{W}_x, \mathbf{W}_y$, которые максимизируют корреляцию между X, Y в новом пространстве:

$$\begin{aligned}
\mathbf{W}_x^*, \mathbf{W}_y^* &= \arg\max_{\mathbf{W}_x, \mathbf{W}_y} \operatorname{corr}(\mathbf{W}_x^\mathsf{T} X, \mathbf{W}_y^\mathsf{T} Y) \\
&= \arg\max_{\mathbf{W}_x, \mathbf{W}_y} \frac{\mathbf{W}_x^\mathsf{T} \hat{\mathbf{E}}[XY^\mathsf{T}] \mathbf{W}_y}{\sqrt{\mathbf{W}_x^\mathsf{T} \hat{\mathbf{E}}[XX^\mathsf{T}] \mathbf{W}_x \mathbf{W}_y^\mathsf{T} \hat{\mathbf{E}}[YY^\mathsf{T}] \mathbf{W}_y}} \\
&= \arg\max_{\mathbf{W}_x, \mathbf{W}_y} \frac{\mathbf{W}_x^\mathsf{T} C_{12} \mathbf{W}_y}{\sqrt{\mathbf{W}_x^\mathsf{T} C_{11} \mathbf{W}_x \mathbf{W}_y^\mathsf{T} C_{22} \mathbf{W}_y}},
\end{aligned} \tag{1}$$

где $\hat{\mathbf{E}}[f(\mathbf{x}, \mathbf{y})] = \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{x}_i, \mathbf{y}_i)$, матрицы ковариации X и Y есть $C_{11} = \frac{1}{m} XX^\mathsf{T} \in \mathbb{R}^{n_1 \times n_1}$, $C_{22} = \frac{1}{m} YY^\mathsf{T} \in \mathbb{R}^{n_2 \times n_2}$, а матрица кросс-ковариации X, Y есть $C_{12} = \frac{1}{m} XY^\mathsf{T} \in \mathbb{R}^{n_1 \times n_2}$.

Развивая идею [2] для решения воспользуемся методом Singular Value Decomposition (SVD, Martin and Maes 1979) для $Z = C_{11}^{-1/2} C_{12} C_{22}^{-1/2}$ и получим матрицы U, S, V. Тогда

$$\begin{aligned}
\mathbf{W}_x^* &= C_{11}^{-\frac{1}{2}} U = (\frac{1}{m} XX^\mathsf{T})^{-\frac{1}{2}} U \\
\mathbf{W}_y^* &= C_{22}^{-\frac{1}{2}} V = (\frac{1}{m} YY^\mathsf{T})^{-\frac{1}{2}} V \\
\operatorname{corr}(\mathbf{W}_x^{\mathsf{T}*} X, \mathbf{W}_y^{\mathsf{T}*} Y) &= \operatorname{trace}(Z^\mathsf{T} Z)^{\frac{1}{2}}
\end{aligned} \tag{2}$$

Мы рассмотрим механизм внимания, который используется для определения важности разных частей входных данных.

Механизм самовнимания определяется следующим образом:

$$\begin{aligned} &\text{attn} : \mathbb{R}^{m\times d} \times \mathbb{R}^{m\times d} \times \mathbb{R}^{m\times d} \longrightarrow \mathbb{R}^{m\times d} \\ &\text{attn}(Q, K, V) = \varphi\left(\frac{QK^\top}{\sqrt{d}}\right) V \end{aligned} \tag{3}$$

где $Q, K, V \in \mathbb{R}^{m\times d}$ представляют собой запросы, ключи и значения соответственно, а $\varphi : \mathbb{R}^{m\times m} \longrightarrow \mathbb{R}^{m\times m}$ — нелинейная функция, применяемая по строкам (обычно softmax).

Самовнимание, применяемое к входным данным $X \in \mathbb{R}^{m\times n_1}$, вычисляется следующим образом:

$$\begin{aligned} &\text{self-attn} : \mathbb{R}^{m\times n_1} \longrightarrow \mathbb{R}^{m\times d} \\ &\text{self-attn}(X) = \text{attn}(XW_q, XW_k, XW_v) \end{aligned} \tag{4}$$

где $W_q, W_k, W_v \in \mathbb{R}^{n_1\times d}$ — матрицы параметров.

3.2 CCA и механизм внимания

И CCA, и механизмы внимания направлены на выявление взаимосвязей между двумя наборами данных. Однако они существенно различаются по подходам и областям применения:

Аспект	**Механизм внимания**	**Канонический корреляционный анализ (CCA)**
Цель	Выявить релевантные части входных последовательностей	Получить вложения в одном скрытом пространстве + снижение размерности
Мера сходства	$A = \frac{1}{\sqrt{d}} QK^\top$ — матрица внимания	$\text{tr}(A^\top S_{12} B)$, при условии $A^\top S_{11} A = B^\top S_{22} B = I$
Цель оптимизации	Минимизация задачи-специфической функции потерь	$\max_{A,B} \text{corr}(A^\top X, B^\top Y)$

Таблица 1: Сравнение механизмов внимания и CCA

Отметим, что $A^\top S_{12} B = \frac{1}{m} A^\top X Y^\top B = \frac{1}{m} A^\top X \left(B^\top Y\right)^\top = \frac{1}{m} \widehat{Q}\widehat{K}^\top$. Это весьма похоже на формулу матрицы внимания $A = \dfrac{1}{\sqrt{d}} QK^\top$. Особенно в случае кросс-внимания, где Q — линейное преобразование X_1, а K — линейное преобразование X_2.

4 Формулировка CI через CCA

4.1 Вероятностная модель канонического корреляционного анализа

Рассмотрим две случайные величины $X \in \mathbb{R}^{n_1}$ и $Y \in \mathbb{R}^{n_2}$ с совместным распределением P_{XY}. Согласно вероятностной интерпретации канонического корреляционного анализа [3], существует латентная переменная $Z \in \mathbb{R}^k$ ($k \leqslant \min(n_1, n_2)$), такая что:

$$\begin{aligned} &Z \sim \mathcal{N}(0, I_k) \\ &X = AZ + \varepsilon_X, \quad \varepsilon_X \sim \mathcal{N}(0, \Psi_X) \\ &Y = BZ + \varepsilon_Y, \quad \varepsilon_Y \sim \mathcal{N}(0, \Psi_Y) \end{aligned} \tag{5}$$

где $\Psi_X \in \mathbb{R}^{n_1 \times n_1}$ и $\Psi_Y \in \mathbb{R}^{n_2 \times n_2}$ — диагональные матрицы ковариаций шумов, $A \in \mathbb{R}^{n_1 \times k}$ и $B \in \mathbb{R}^{n_2 \times k}$ — матрицы загрузок.

Теорема 4.1 (Эквивалентность вероятностного и классического CCA)**.** *Пусть (U,V) — канонические направления классического CCA. Тогда существует параметризация вероятностной модели, в которой матрицы загрузок имеют вид (A^*, B^*) в вероятностной модели CCA связаны с решениями классического CCA следующим образом:*

$$A^* = C_{11}^{1/2} U \Lambda^{1/2}, \quad B^* = C_{22}^{1/2} V \Lambda^{1/2} \tag{6}$$

где U, V — левые и правые сингулярные векторы матрицы
$Z = C_{11}^{-1/2} C_{12} C_{22}^{-1/2}$, *$\Lambda$ — диагональная матрица сингулярных значений, $C_{11} = Cov(X)$, $C_{22} = Cov(Y)$, $C_{12} = Cov(X,Y)$.*

Доказательство. Из вероятностной модели следует, что ковариационная матрица совместного распределения:

$$\Sigma = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} AA^T + \Psi_X & AB^T \\ BA^T & BB^T + \Psi_Y \end{pmatrix}$$

Максимизация канонической корреляции эквивалентна решению обобщенной проблемы собственных значений:

$$C_{12}C_{22}^{-1}C_{21}u = \lambda C_{11}u$$

Подстановка выражений для ковариаций из вероятностной модели и использование спектрального разложения показывает, что оптимальные матрицы загрузок должны иметь указанную форму. Подробное доказательство эквивалентности следует из анализа максимального правдоподобия вероятностной CCA и его связи с обобщённой задачей собственных значений; см. [3]. □

4.2 Структурная причинная модель с оператором $\mathbf{do}(X)$

Введем структурную причинную модель (SCM) для системы (X,Y), где X является потенциальной причиной Y. Предположим, что существуют скрытые общие причины Z, влияющие как на X, так и на Y.

[Структурная причинная модель для CCA] Структурная причинная модель $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$ определяется как:

$$\begin{aligned} Z &\leftarrow U_Z, \quad U_Z \sim \mathcal{N}(0, I_k) \\ X &\leftarrow f_X(Z, U_X) = AZ + U_X, \quad U_X \sim \mathcal{N}(0, \Psi_X) \\ Y &\leftarrow f_Y(X, Z, U_Y) = CX + BZ + U_Y, \quad U_Y \sim \mathcal{N}(0, \Psi_Y) \end{aligned} \tag{7}$$

где $C \in \mathbb{R}^{n_2 \times n_1}$ — матрица прямого причинного влияния X на Y, U_X, U_Y — независимые шумовые переменные.

Теорема 4.2 (Интервенционное распределение в causal pCCA)**.** *В структурной модели*

$$\begin{aligned} Z &\sim \mathcal{N}(0, I_k), \\ X &= AZ + U_X, \\ Y &= CX + BZ + U_Y, \end{aligned}$$

интервенционное распределение $P(Y \mid do(X = x))$ *имеет вид*

$$P(Y \mid do(X = x)) = \mathcal{N}\big(Cx,\, BB^\top + \Psi_Y\big).$$

Доказательство. Интервенция $do(X = x)$ заменяет структурное уравнение для X на

$$X \leftarrow x,$$

тем самым устраняя зависимость X от Z.

Поскольку Z не является потомком X, его распределение не меняется:

$$P(Z \mid do(X = x)) = P(Z) = \mathcal{N}(0,I_k).$$

Подставляя в уравнение для Y,

$$Y = Cx + BZ + U_Y,$$

и используя независимость Z и U_Y, получаем заявленное гауссовское распределение. □

4.3 Связь с каноническими корреляциями и оператором $\mathbf{do}(X)$

Теперь сформулируем ключевую теорему, связывающую причинный эффект с каноническими корреляциями.

Теорема 4.3 (Проекция каузального оператора на CCA-подпространство)**.** *Пусть* (u_i,v_i) *— канонические направления, нормированные так, что*

$$u_i^\top C_{11} u_i = v_i^\top C_{22} v_i = 1.$$

Тогда проекция интервенционного эффекта

$$\mathcal{T}_{X\to Y}(x) := \mathbb{E}[Y \mid do(X = x)]$$

на канонические координаты имеет вид

$$v_i^\top \mathcal{T}_{X\to Y}(x) = (v_i^\top C u_i)\,(u_i^\top x).$$

Доказательство. Из предыдущей теоремы

$$\mathbb{E}[Y \mid do(X = x)] = Cx.$$

Проецируя на v_i,

$$v_i^\top \mathbb{E}[Y \mid do(X = x)] = v_i^\top Cx = (v_i^\top C u_i)(u_i^\top x),$$

где разложение $x = \sum_i (u_i^\top x) u_i$ понимается как ортогональная проекция x на подпространство, натянутое на $\{u_i\}$. □

Теорема 4.4 (Некаузальность канонических корреляций)**.** *Пусть данные порождены SCM*

$$X = AZ, \qquad Y = CX + BZ,$$

где $Z \sim \mathcal{N}(0,I)$*, и для простоты* $\Psi_X = \Psi_Y = 0$*. Тогда при нормировке*

$$u_i^\top C_{11} u_i = v_i^\top C_{22} v_i = 1$$

каноническая корреляция имеет разложение

$$\rho_i = u_i^\top C_{11} C^\top v_i + u_i^\top A B^\top v_i.$$

В частности, при наличии скрытой переменной Z *каноническая корреляция не совпадает с каузальным вкладом.*

Доказательство. RAW Подставить в $C_{12} = A(CA + B)^\top$, затем $\rho_i = u_i^\top C_{12} v_i$. Первое слагаемое отражает путь через C и confounder, второе — чистый confounded путь □

Следствие 4.4.1.

$$\mathbb{E}[Y \mid do(X = x_1)] - \mathbb{E}[Y \mid do(X = x_0)] = C(x_1 - x_0).$$

4.4 Практическая интерпретация для временных рядов и текстов

Для d-вариантных временных рядов оператор $\mathrm{do}(X_t = x)$ интерпретируется как искусственное форсирование состояния системы X в момент времени t и наблюдение за реакцией $Y_{t+\tau}$. В контексте сходящихся отображений (Convergent Cross-Mapping) это позволяет отделить прямое причинное влияние от корреляций, обусловленных общими динамическими факторами.

Для текстовых документов интервенция $\mathrm{do}(X = x)$ соответствует модификации эмбеддингов определенных лингвистических признаков (например, тональности или тематических маркеров) и измерению изменений в связанных документах. Скрытая переменная Z здесь представляет собой семантические концепции, общие для обоих наборов документов.

5 Experiments

TODO Сказать что будем использовать многомерные временные ряды и тексты

5.1 Dataset Details

- что за датасет
- какая предобработка данных проводилась
- в каком виде данные подавались в модель

5.2 Training Details

TODO

- TODO

5.3 Experimental Results

TODO

Список литературы

[1] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[2] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8992–8999, 2020.

[3] Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[4] Nick Martin and Hermine Maes. Multivariate analysis. *London, UK: Academic*, 1979.

[5] Yu-Ting Lan, Wei Liu, and Bao-Liang Lu. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2020.