# Transformers and Canonical Correlation Analysis for Causal Inference

IA pour la Science

January 3th, 2026

# Connected and disconnected causality

The main difference between LinGAM, PCMCIA and CCM is that CCM is a non-parametric way to reveal dependence in the variables. LinGAM and PCMCIA assume fixed model structure to investigate and their causality inference depends on the model structure. CCM keeps two spaces disconnected.

To gain benefits of the two approaches we propose the following:

1. Optimize the latent spaces of the causality by reconstruction
2. Assign the model structure to connect the spaces: CCA or Transformers
3. Investigate the causality through the change of distribution in separate spaces
4. Apply the interventional distribution to estimate the causal effect united spaces

# Interpretation for Time Series and CCM

Given two sets of vectors $X \in \mathbb{R}^{n_1 \times m}$ and $Y \in \mathbb{R}^{n_2 \times m}$, where $m$ denotes the number of vectors, CCA learns two linear transformations $A \in \mathbb{R}^{n_1 \times r}$ and $B \in \mathbb{R}^{n_2 \times r}$ such that the correlation between $A^\mathsf{T} X$ and $B^\mathsf{T} Y$ is maximized. Note the covariances of $X$ and $Y$ as $S_{11} = \frac{1}{m} XX^\mathsf{T} \in \mathbb{R}^{n_1 \times n_1}, S_{22} = \frac{1}{m} YY^\mathsf{T} \in \mathbb{R}^{n_2 \times n_2}$, and the cross-covariance of $X, Y$ as $S_{12} = \frac{1}{m} XY^\mathsf{T} \in \mathbb{R}^{n_1 \times n_2}$. The CCA objective is

$$A^*, B^* = \arg\max_{A,B} \operatorname{corr}(A^T X, B^T Y) = \arg\max_{A,B} \frac{A^T S_{12} B}{\sqrt{A^T S_{11} A \cdot B^T S_{22} B}}$$

The solution of the above equation is fixed and can be solved in multiple ways. Let $U, S, V^\mathsf{T}$ be the SVD of the matrix $Z = S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}}$. Then $A^*, B^*$ and the total maximum canonical correlation are

$$A^* = S_{11}^{-\frac{1}{2}} U = \left( \frac{1}{m} XX^T \right)^{-\frac{1}{2}} U$$

$$B^* = S_{22}^{-\frac{1}{2}} V = \left( \frac{1}{m} YY^T \right)^{-\frac{1}{2}} V$$

$$\operatorname{corr}(A^{*T} X, B^{*T} Y) = \operatorname{trace}(Z^T Z)^{\frac{1}{2}}.$$

# Self-attention and cross-attention (1)

Attention mechanisms are used to determine the relevance of different parts of the input data. The self-attention mechanism is defined as follows:

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

where $Q, K, V \in \mathbb{R}^{m\times d}$ represent the queries, keys, and values, respectively, and $\varphi : \mathbb{R}^{m\times m} \longrightarrow \mathbb{R}^{m\times m}$ is row-wise applied nonlinear function, usually softmax. The dot product between $Q$ and $K$ determines the attention weights, which are normalized using the softmax function. The result is then applied to the values $V$ to generate the output.

# Self-attention and cross-attention (2)

Self-attention applied to the input $X \in \mathbb{R}^{m \times n_1}$ is computed as:

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

where $W_q, W_k, W_v \in \mathbb{R}^{n_1 \times d}$ — parameter matrices

In multihead attention, several attention heads are used in parallel, where each head computes its own attention weights and outputs. The outputs are then concatenated and linearly transformed by a weight matrix $W^Q \in \mathbb{R}^{p \cdot d \times d}$:

$$\text{multihead-attn}(Q, K, V) = [\text{head}_1, \ldots, \text{head}_p] W^Q,$$

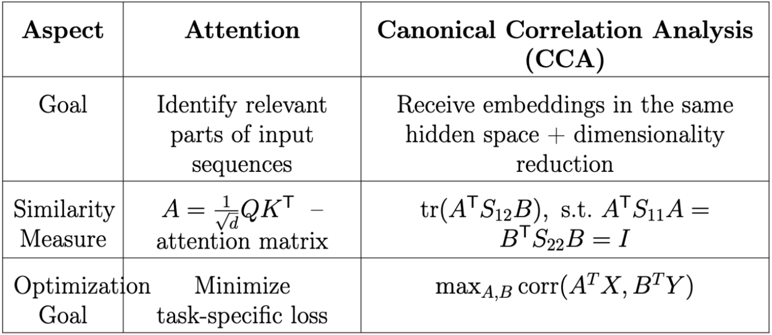
where $\text{head}_i = \text{self-attn}(X)$

Cross-attention, in contrast, involves attention between two different sets of inputs. It computes attention by using one set of inputs for queries $X_1 \in \mathbb{R}^{m \times d_1}$ and another set for keys and values $X_2 \in \mathbb{R}^{m \times d_2}$:

$$\text{cross-attn}(X_1, X_2) = \text{attn}(X_1 W_q, X_2 W_k, X_2 W_v) \tag{1}$$

# Comparison of Attention Mechanisms and CCA

Both CCA and attention mechanisms aim to find relationships between two sets of data. However, they differ significantly in their approach and applications:



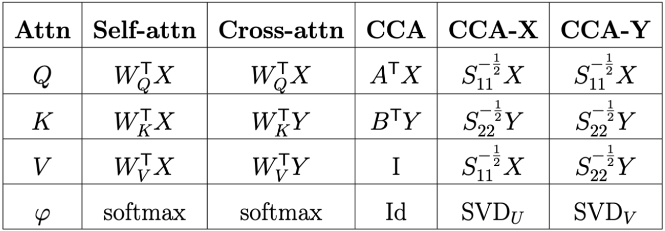
| **Aspect** | **Attention** | **Canonical Correlation Analysis (CCA)** |
|---|---|---|
| Goal | Identify relevant parts of input sequences | Receive embeddings in the same hidden space + dimensionality reduction |
| Similarity Measure | $A = \frac{1}{\sqrt{d}}QK^\top$ – attention matrix | $\mathrm{tr}(A^\top S_{12}B)$, s.t. $A^\top S_{11}A = B^\top S_{22}B = I$ |
| Optimization Goal | Minimize task-specific loss | $\max_{A,B} \mathrm{corr}(A^T X, B^T Y)$ |

# United notation of CCA and attention

Note that $A^{\mathsf{T}} S_{12} B = \frac{1}{m} A^{\mathsf{T}} X Y^{\mathsf{T}} B = \frac{1}{m} A^{\mathsf{T}} X \left(B^{\mathsf{T}} Y\right)^{\mathsf{T}} = \frac{1}{m} \widehat{Q} \widehat{K}^{\mathsf{T}}$.

And it's quite similar to attention matrix formula $A = \dfrac{1}{\sqrt{d}} Q K^{\mathsf{T}}$.

Especially, in cross attention case, where $Q$ is a linear transformation of $X_1$ and $K$ is a linear transformation of $X_2$.



| **Attn** | **Self-attn** | **Cross-attn** | **CCA** | **CCA-X** | **CCA-Y** |
|---|---|---|---|---|---|
| $Q$ | $W_Q^{\mathsf{T}} X$ | $W_Q^{\mathsf{T}} X$ | $A^{\mathsf{T}} X$ | $S_{11}^{-\frac{1}{2}} X$ | $S_{11}^{-\frac{1}{2}} X$ |
| $K$ | $W_K^{\mathsf{T}} X$ | $W_K^{\mathsf{T}} Y$ | $B^{\mathsf{T}} Y$ | $S_{22}^{-\frac{1}{2}} Y$ | $S_{22}^{-\frac{1}{2}} Y$ |
| $V$ | $W_V^{\mathsf{T}} X$ | $W_V^{\mathsf{T}} Y$ | I | $S_{11}^{-\frac{1}{2}} X$ | $S_{22}^{-\frac{1}{2}} Y$ |
| $\varphi$ | softmax | softmax | Id | $\mathrm{SVD}_U$ | $\mathrm{SVD}_V$ |

Details of the CCA projection of $X$ to latent space:

$$\begin{aligned} \mathrm{CCA}_{XY}(X) &= U^{\mathsf{T}} S_{11}^{-\frac{1}{2}} X = U^{\mathsf{T}} X_1 \\ \mathrm{CCA}_{XY}(Y) &= V^{\mathsf{T}} S_{22}^{-\frac{1}{2}} Y = V^{\mathsf{T}} Y_1 \qquad (2) \\ Z &= S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}} = \frac{1}{m} X_1 Y_1^{\mathsf{T}} \end{aligned}$$

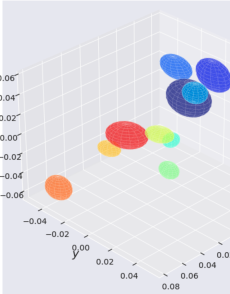
# Estimate empirtical distribution of one neuron

For each neuron from the autoencoder we estimate the empirical distribution through the monte-carlo sampling

$$f = \sigma \circ W^T x, \quad W = [w_1, \dots, w^N], \quad \mathrm{w}_j \sim \mathcal{N}(\hat{\mathrm{w}}_j, \hat{\mathrm{A}}_j).$$

The algorithm:

1. Sample $k$-th batch $\{x\}$ from the data set $D$.
2. For each batch find the parameters $w$.
3. Construct the matrix $\bar{W}$ stacking the parameters.
4. Estimate from $\bar{W}$ the $E(w), A = Cov(w) = \frac{1}{K-1} W^T W$.
5. Do intervention.
6. Estimate the difference between the two distributions (before and after intervention) through KL-divergence.



Hypothesis. The metric tensor $A$ changes with its 1st derivative continuously.

# Probabilistic Model of CCA

## Theorem (Probabilistic CCA)

*Let* $Z \sim \mathcal{N}(0, I_k)$*, then*

$$X = AZ + \varepsilon_X, \quad \varepsilon_X \sim \mathcal{N}(0, \Psi_X), \quad Y = BZ + \varepsilon_Y, \quad \varepsilon_Y \sim \mathcal{N}(0, \Psi_Y).$$

*There exists a parametrization* $(A^*, B^*)$ *such that the classical canonical directions* $(U, V)$ *satisfy*

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

## Insight

pCCA represents CCA as a latent-variable model and relates probabilistic parameters to canonical directions.

# SCM with do($X$) Intervention

Consider the model

$$Z \sim \mathcal{N}(0, I_k), \quad X = AZ + U_X, \quad Y = CX + BZ + U_Y.$$

Then $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$ defines a structural causal model for $X \to Y$ with hidden confounder $Z$ and direct effect $C$.

## Insight

Formalizes causal relationships and sets the basis for interventional analysis via `do(X)`.

# Interventional Distribution

## Theorem (Interventional Distribution)

*For the SCM above,*

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

## Insight

Mean encodes direct causal effect, variance accounts for latent confounders. Allows separation of causation from correlation.

## Corollary (Linear Causal Effect)

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

## Insight

Identifies linear, pure causal effect, independent of hidden confounders.

# Projection onto CCA Subspace

## Theorem (Projection of Causal Effect)

*Let $(u_i, v_i)$ be canonical directions normalized by $u_i^\top C_{11} u_i = v_i^\top C_{22} v_i = 1$. Then*

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

## Insight

Shows how interventional effect projects linearly onto canonical coordinates, facilitating interpretation in CCA space.

## Theorem (Non-Causal Nature of Canonical Correlations)

*For SCM with $X = AZ$, $Y = CX + BZ$ and $\Psi_X = \Psi_Y = 0$, the canonical correlation $\rho_i$ satisfies*

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

# A question to discuss

1. A simple approach to infer causality between two latent spaces applicable to transformer so that we can use both time series and textual data.

- The thing is the literature review shows that Causal Inference is used for time series and spatial time series data (which is great),
- but it seems there is no significant works for textual data.