Week 13

Wentao Gao

Content

- Applications of Causality in Science
- More specific challenge in our problem
- ICA based method

Applications of Causality in Science

- What it can do in practice,
- How to solve the actual problem in real life
- What problem we wish to solve

Applied in a range of area

- In climate prediction.(Inferring causation from time series in Earth system sciences)
- Evaluate climate model.(Causal networks for climate model evaluation and constrained projections)
- Ecosystem (Detecting Causality in Complex Ecosystems CCM)
- Extreme climate event prediction(The missing risks of climate change)
- Stock prediction (A causal feature selection algorithm for stock prediction modeling)
- Trajectory prediction in robotics or self-driving car(Causal Temporal–Spatial Pedestrian Trajectory Prediction With Goal Point Estimation and Contextual Interaction)
- …Complex dynamic system.

In climate prediction

- We normally have two ways:

One is to use the Physics model to generate prediction data.

One is to use the observational data to train a DL prediction model to do the prediction, however, it's quite hard for us to know actually the mechanism of the climate change.

So an important research area is to find the mechanism or the dynamic system for the climate change to help us understand the nature deeply.

Understand the mechanism will help us build an interpretable model which represent what nature really happened

How can we achieve the goal

- One way is to use causality to help understand such complex dynamic system.
- The causal graph in causality will show a real causal relationship between variables. This will indicate the mechanism of the system.
- And causal effect will show to what extent one variable affect another variable.
- However, How to find the causal graph and measure the causal effect is still a challenge for us.

More specific challenge in our problem

- Causality problem
- Causal discovery
- Or causal inference

Challenges

Process:

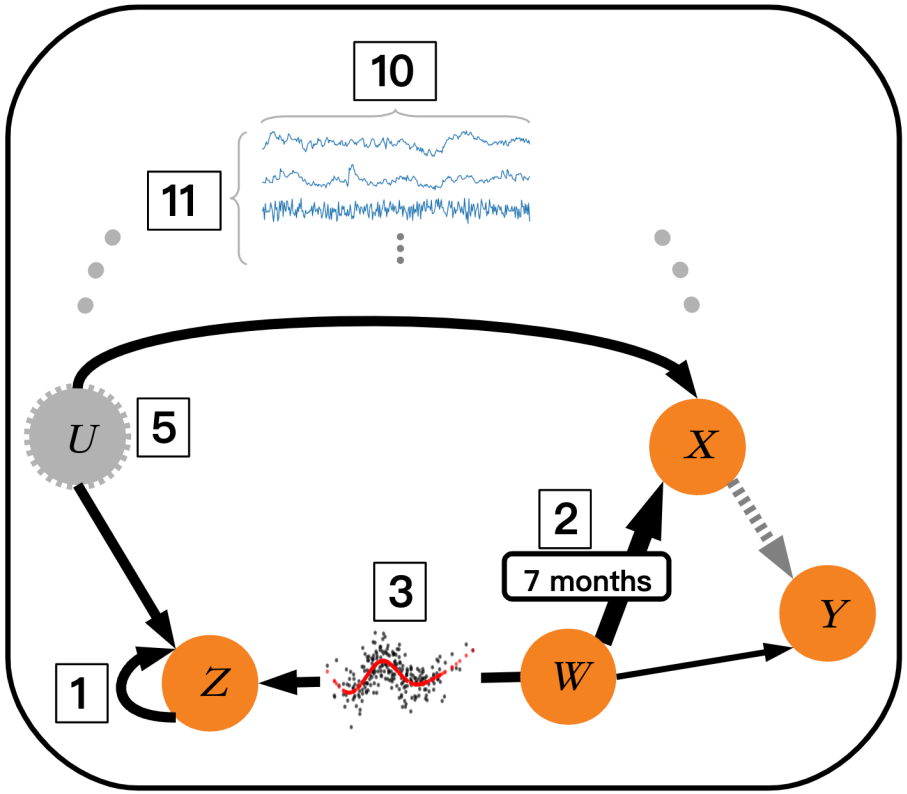
1 Autocorrelation
2 Time delays
3 Nonlinear dependencies
4 Non–gaussian noise

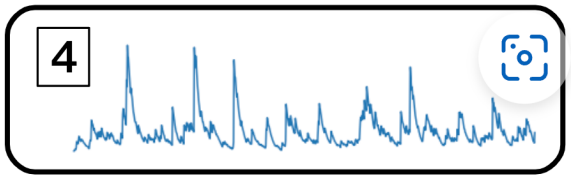
Data:

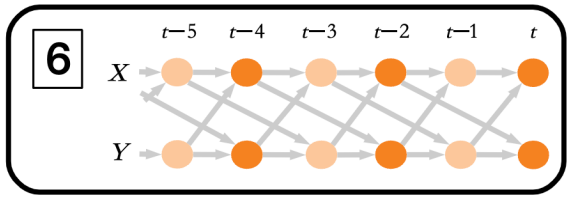
5 Non–stationarity due to unobserved drivers
6 Time subsampling
7 Time aggregation
8 Observational noise
9 Selection bias

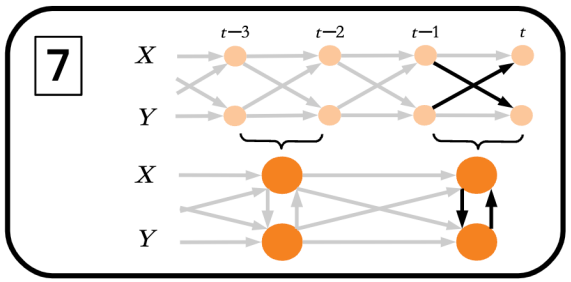
Computational / statistical:

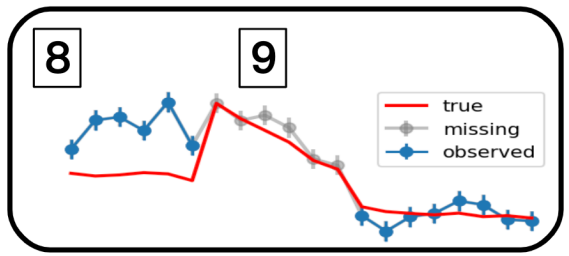
10 Sample size
11 High dimensionality











Observational variables

True variables

Real causal relationship

Independent Conditional Analysis

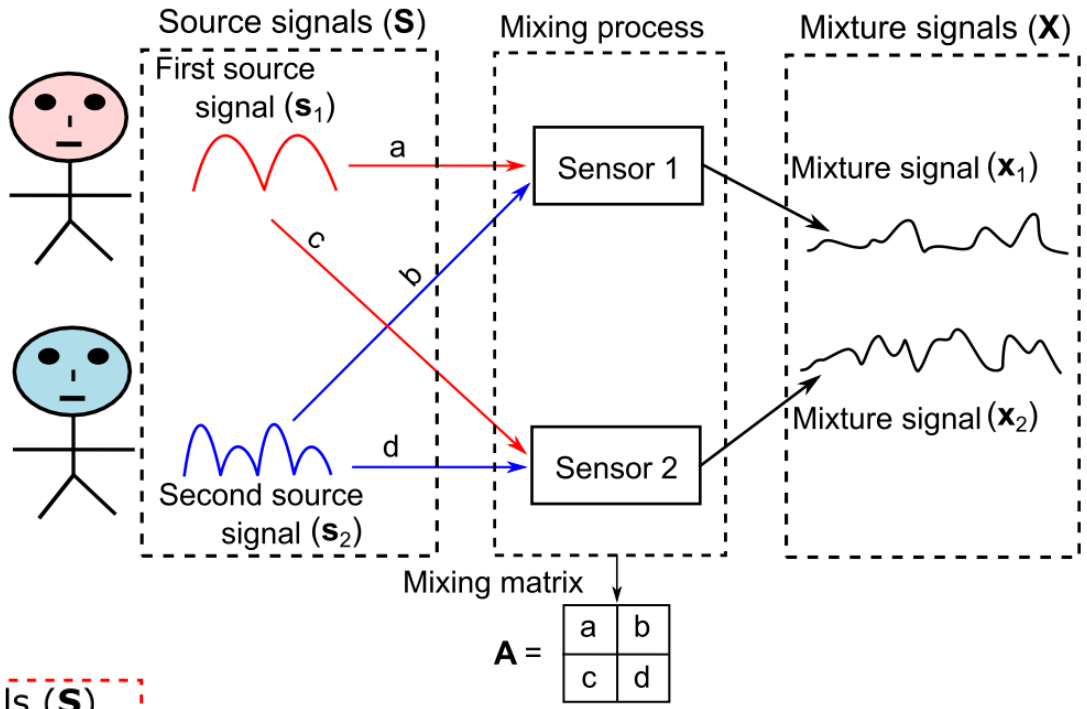
- ICA Intro
- ICA identifiability
- Nonlinear ICA
- TCL，PCL
- iVAE

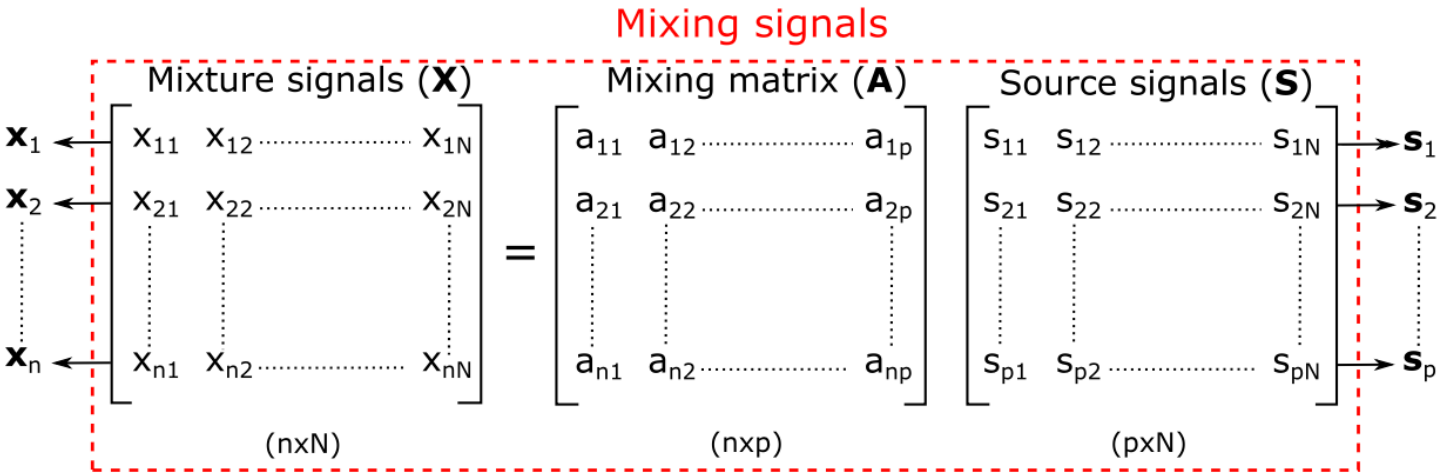
ICA Intro

$$\mathbf{S} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix} = \begin{pmatrix} (s_{11}, s_{12}, \ldots, s_{1N}) \\ (s_{21}, s_{22}, \ldots, s_{2N}) \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} a\mathbf{s}_1 + b\mathbf{s}_2 \\ c\mathbf{s}_1 + d\mathbf{s}_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix} = \mathbf{As}$$

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$





ICA identifiability

Objective: Blindly separate observed mixed signals into independent source signals.

- **Key Points on Identifiability**:

1. **Requirement of Non-Gaussianity**: This is a fundamental requirement for the identifiability in ICA. If all sources were Gaussian, ICA would be unidentifiable. This is due to the rotational symmetry of the Gaussian distribution: any linear combination of multiple Gaussian sources remains Gaussian, hence the original mixing matrix and source signals cannot be uniquely determined.
2. **Independence**: ICA assumes that source signals are statistically independent. This means there's no correlation or dependence between the sources.
3. **Matching in Number**: The number of observed signals should match the number of independent sources. In other words, if we have n observed signals, then we assume there are n independent source signals.
4. **Conditions on the Mixing Matrix**: The mixing matrix should be full-rank, implying it's square and invertible. This ensures that recovering the original sources from the mixed signals is feasible.
5. **Ambiguities in ICA**: While ICA can recover independent source signals, it cannot determine the order or scaling of the sources. This means that ICA can give multiple valid interpretations of the original signals, differing only in scaling and permutation.

Nonlinear ICA

General nonlinear ICA model:

$$[x_1(t), \ldots, x_n(t)] = \mathbf{f}([s_1(t), \ldots, s_n(t)])$$

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)).$$

So, the Goal of Nonear ICA is to reconstruct f^{-1}.

Then, according to the observational data, we can get its independent latent variables.

Existing paper show that if the observational data is e independent and identically distributed*i.i.d* ,which there is no time structure or similar,
it is unidentifiable *(*Nonlinear independent component analysis: Existence and uniqueness results
Aapo Hyvärinen*)*

Output(Observation - x)

Transformation(Function - f)

Input(Source - s)

Existing method mainly using two ways to import new assumption to get indentifiability an its proof

Autocorrelation of data: the main assumption is that there is a temporal dependency between the data (frame data), which intuitively determines the order of occurrence.

Nonstationary of the data: capitalizes on the differences that exist between the data (more on this later)

Time contrastive learning

(nonstationary)

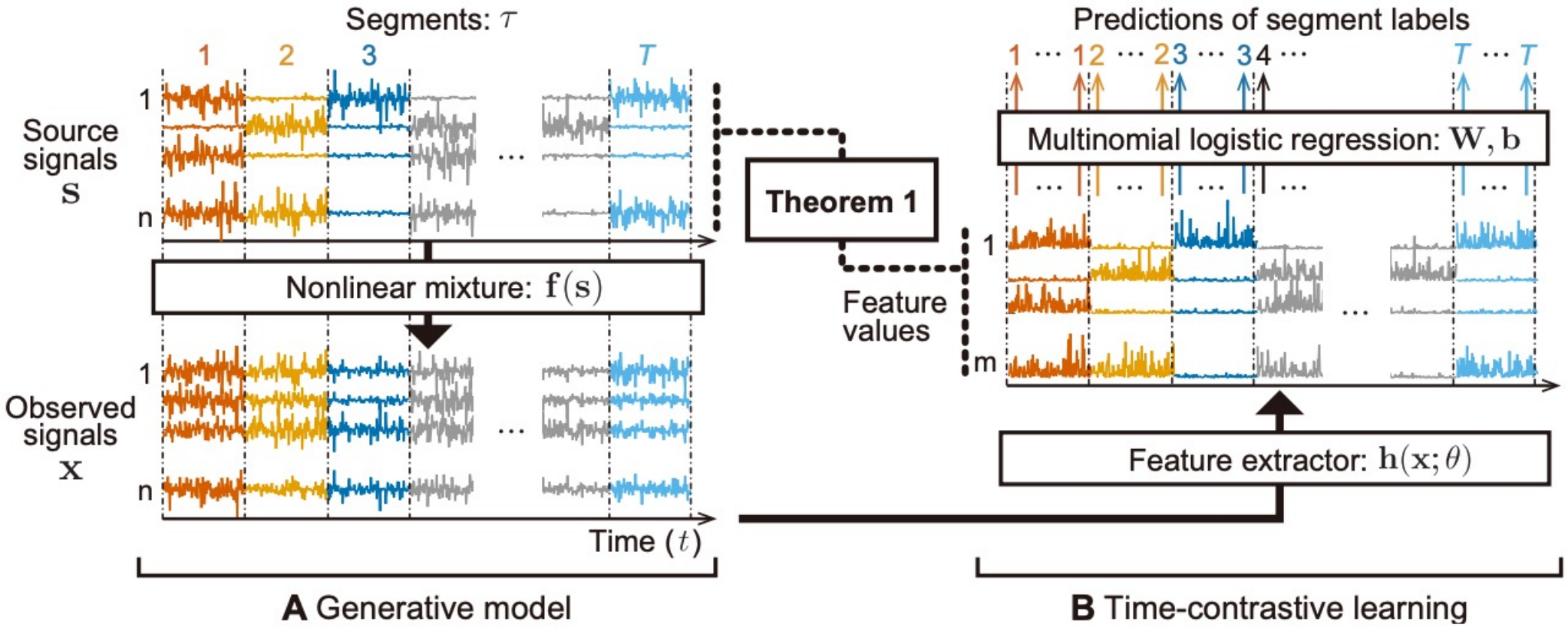
Paper: Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA

Definition: A method to extract independent temporal sources from multi-channel time series data.

How it Works: Time partitioning, contrastive learning, and source signal recovery.

Advantages: Nonlinear separation, robustness, etc.

Applications: Neural data analysis, audio source separation.



Observe a sequence data $x(t) \sim R^n$, separate the x(t) into equal length T segment.

Training MLP, give a datapoint , do a T classification problem to determine which segment the data point belongs to.

In the last hidden layer h of MLP, NN should be able to learn the non-stationarity between data. (Difference of signal)

It is quite hard to recover the source signal from the observational signal directly, Using feature Extractor to get a better representation. With this representation signal, using ICA to recover the source signal will be easier.

TCL introduces a new assumption: the latent variables *s* are nonstationary, meaning their statistical properties change over time. This nonstationarity might manifest as a large variance difference.

It's proven that TCL essentially performs a specific nonlinear mix, given by

$x=(Ws)^2$,

where *W* is a weight matrix for linear mixing.

This means that TCL decomposes the nonlinear ICA problem into a combination of linear mixing and a squaring operation.

Permutation-contrastive learning(PCL)

(Temporal dependencies)

Model and assumption

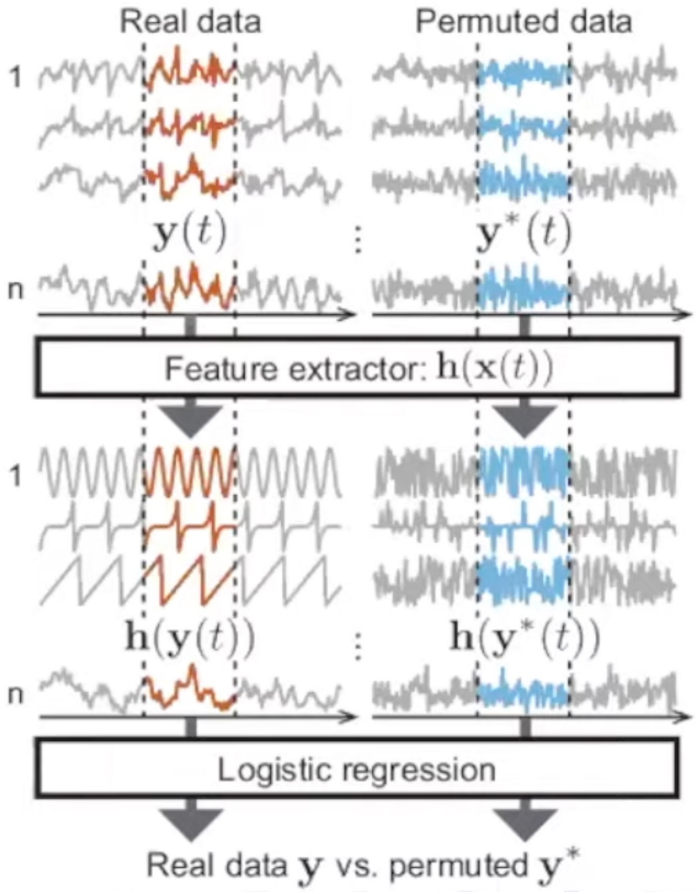
We assume the n observed signals (i.e. time series or stochastic processes) $x_1(t), \ldots, x_n(t)$ are generated as a nonlinear transformation $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ of n latent signals $s_1(t), \ldots, s_n(t)$:

$$[x_1(t), \ldots, x_n(t)] = \mathbf{f}([s_1(t), \ldots, s_n(t)]) \qquad (1)$$

Denoting by $\mathbf{x}(t)$ the vector $[x_1(t), \ldots, x_n(t)]$, and likewise for $\mathbf{s}$, this can be expressed simply as

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)). \qquad (2)$$

We assume the function $\mathbf{f}$ is invertible (bijective) and sufficiently smooth but we do not constrain it in any particular way.



Take short time windows as new data

$$\mathbf{y}(t) = (\mathbf{x}(t), \mathbf{x}(t-1))$$

Create randomly time-permuted data

$$\mathbf{y}^*(t) = (\mathbf{x}(t), \mathbf{x}(t^*))$$

with t^* a random time point.

Using logistic regression to classify the positive sample or negative sample.

The hidden layer of logistic regression using mlp is a feature extractor that can discriminate y from y*, which can learn the time dependence and structure feature.

Performs Nonlinear ICA with the extracted feature to get temporally dependent components

A more general ICA framework

$$p_\theta(\mathbf{x}, \mathbf{s}) = p_\theta(\mathbf{x}|\mathbf{s})p_\theta(\mathbf{s}),\ p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{s})d\mathbf{s}$$

VAE is unidentifiable which is not able to recover the original source.

One solution is we further assume every s_i conditionally dependent on auxiliary variable u, But conditional independent on other s_j, for the situation of non-stationarity, the u means Segment information t-[0, T], for correlation is the last time information x(t-1), besides, u can be other auxiliary variable depends on the different situation.

$$\log p(\mathbf{s}|\mathbf{u}) = \sum_{i=1}^{n} q_i(s_i, \mathbf{u})$$

Some thinking

- In the nonlinear ICA, we use the time structure to help indentifiability.

- So we can get the source variables from the observational variables.
- But our observational variables not only included X_t^i but also X_{t-n}^i.
- How can we add the time lag variables into the observational data to consider its source variable.

Another problem is that how do we know the meaning of each source. If we can get the causal structure using source variables. How do we transfer it to observational variable version.

- LR
- Drought prediction
- MLDL not enough
- Causality based

- RG
- Precision

- Feature selection
- ICA
- Causal transfer learning