Robustness against Distribution Shift using Causal Models

Table of Contents

Causality

- Causality is the study of the underlying structure of a system;
- Through studying causality and causal models, we can:
  - Answer questions about how a system responds to **changes** in its mechanisms;
  - Specify causal relationships that are more **stable** and more fundamental than generic statistical relationships.
- A model assigns truth values to statements about the system; a causal model assigns truth values to causal statements/queries

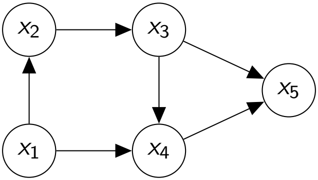
Causal Bayesian Networks

Causal Bayesian Networks are a means for representing a set of interventional distributions, consisting of:

Causal Bayesian Networks

Causal Bayesian Networks are a means for representing a set of interventional distributions, consisting of:

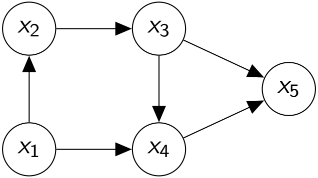
- A directed acyclic graph (DAG) G:



Causal Bayesian Networks

Causal Bayesian Networks are a means for representing a set of interventional distributions, consisting of:

- A directed acyclic graph (DAG) G:

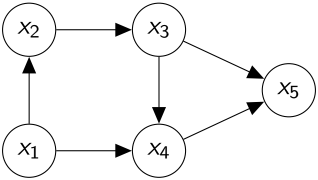


- A probability distribution $P(x|pa(x))$ for every node x in the DAG

Causal Bayesian Networks

Causal Bayesian Networks are a means for representing a set of interventional distributions, consisting of:

- A directed acyclic graph (DAG) G:



- A probability distribution $P(x|pa(x))$ for every node x in the DAG

The joint probability distribution is then $P(x_1, ..., x_N) = \prod_i P(x_i|pa(x_i))$.

Principle of Independent Causal Mechanisms

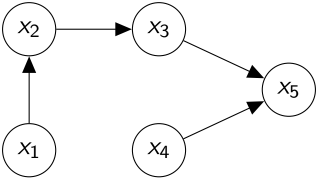
The **ICM principle** states that the individual causal mechanisms of a systems' causal generative process do not:

Principle of Independent Causal Mechanisms

The **ICM principle** states that the individual causal mechanisms of a systems' causal generative process do not:

- *inform* each other
- *influence* each other

To mimic this in our causal model, we derive interventional distributions by making the necessary intervention and leaving all other mechanisms.
Example: $P_{X_4=x'}(x_1, ..., x_5)$



$$P_{X_4=x'}(x_1, ..., x_5) = P(x_2|x_1)P(x_3|x_2)\mathbb{1}_{x_4=x'}P(x_5|x_3, x_4)$$

Table of Contents

Problem Setting

Problem: Machine learning models are not robust to changes in the data generating process (DGP) from train to test time (*distribution shift*).

- Deployment of medical treatments in different hospitals;
- Autonomous vehicles in harsh weather conditions (snow, rain, etc.);
- Variants of environments in RL;
- Different dialects in language;

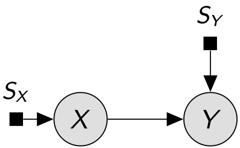
How do we define what kinds of changes our machine learning models should be robust to?

How do we actually achieve this robustness?

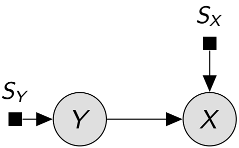
Table of Contents

What might vary?

In a typical machine learning problem, we wish to predict a label or target Y, given covariates X. There are two basic possible causal graphs:



(a) $p_X, p_{Y|X}$



(b) $p_Y, p_{X|Y}$

and 4 possible types of distribution shift:

1. **Covariate shift**: $X \to Y$ with p_X changing and $p_{Y|X}$ fixed.
2. $X \to Y$ with $p_{Y|X}$ changing.
3. **Target/label shift**: $Y \to X$ with p_Y changing and $p_{X|Y}$ fixed.
4. $Y \to X$ with $p_{X|Y}$ changing

How much does it vary by?

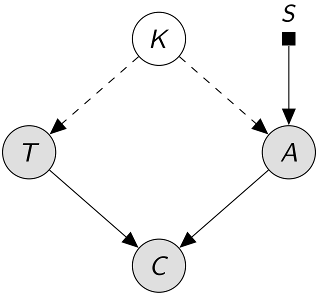
We need additional assumptions, or restrictions on how much these probability distributions can change.

Approaches to obtaining robust models rely on these assumptions:

- Assume only covariate shift/only label shift occurs
- Limit change in $p(Y|X)$ by KL-divergence [Duchi and Namkoong, 2018]
- Assume that target environment is "sufficiently" similar to multiple source environments [Zhang et al., 2015]

Distribution shift as interventions in DGP

We reason about changes in the underlying DGP that generates the data.



Here grey nodes are **observed**, blank nodes are **unobserved/latent** and black squares are **selection variables**. $X = \{A, C\}, Y = \{T\}$.

Selection variables represent potential *interventions* on variables which might have their mechanisms changed in the target environment, in this case $p_{A|K}$.

Table of Contents

Reactive vs Proactive approaches

Reactive approaches use some (usually unlabelled) data from the target environment in order to optimize the machine learning model for that domain.

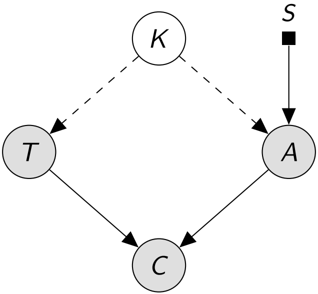
Proactive approaches use machine learning models which are agnostic to the specific target environment (within the class of environments satisfying assumptions).

We focus on a **proactive** approach, which has the following advantages:

- Do not require any data from the target environment
- "One-shot": Achieves performance immediately on first test from target environment
- Closer to how humans adapt

Table of Contents

Motivating example



We wish to predict in hospitalized patients whether or not they have **lung cancer** (T), on the basis of **chest pain symptoms** C and whether or not they take **aspirin** A.

Smoking K is an *unobserved confounder* which affects the chance of lung cancer and heart disease. Aspirin is taken for the latter.

The policy for prescribing aspirin will differ across hospitals, so we add a selection variable S to represent that this mechanism is **unreliable**.

Task and Assumptions

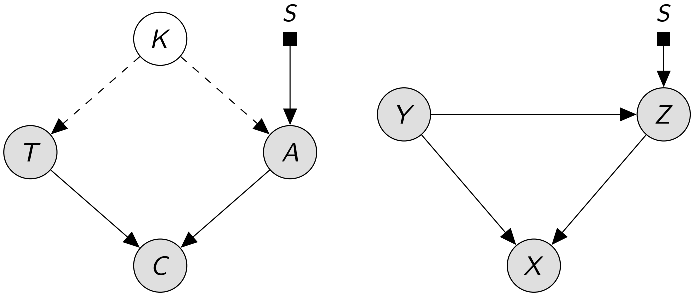
We want an **estimator** $\hat{P}(T|C, A)$ which performs well under all environments, as specified by the selection variables.

We assume that:

- The causal graph is known (but not the conditional distributions $p(x|pa(x))$;
- We know which mechanisms are unreliable;

Stability

Definition An estimator $\hat{P}$ is *stable* if it is conditionally independent of all selection variables given its inputs.



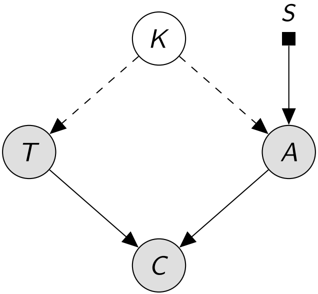
Intuitively, this means given the same input, the predictions $\hat{P}$ are independent of which environment we are in.

Graph Surgery

Idea: Remove dependence on environment-varying mechanisms by performing graph surgery.

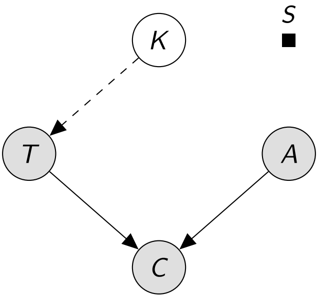
Graph Surgery

Idea: Remove dependence on environment-varying mechanisms by performing graph surgery.



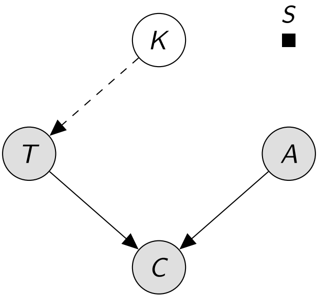
Graph Surgery

Idea: Remove dependence on environment-varying mechanisms by performing graph surgery.



Graph Surgery

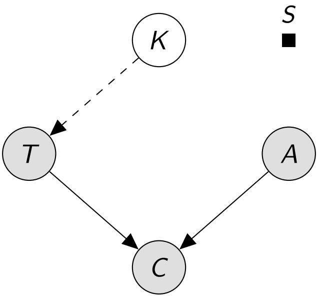
Idea: Remove dependence on environment-varying mechanisms by performing graph surgery.



$$p(T|C, do(A)) \propto p(T, C|do(A)) \propto p(T)p(C|T, A)$$

Graph Surgery

Idea: Remove dependence on environment-varying mechanisms by performing graph surgery.



$$p(T|C, do(A)) \propto p(T, C|do(A)) \propto p(T)p(C|T, A)$$

Theorem: Any estimator of the form $P(T|Z, do(X))$ where $X \supseteq M$ is stable.

Table of Contents

Semi-Markovian Models

In general, we are concerned with **Semi-Markovian models**, that is, models where there are unobserved/latent nodes which cause confounding.

The previous example suggested that we should use the interventional distribution:

$$p(T|do(M), O \setminus (M \cup T))$$

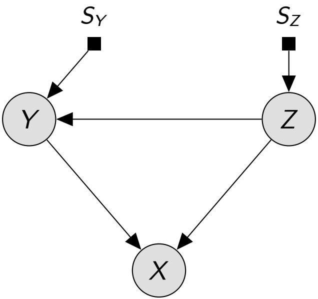
where T are the target variable(s), M are the mutable variables, and O are the observable variables.

However, this runs into two issues:

Semi-Markovian Models

1. Mutable Target

In some cases, the target itself is mutable:

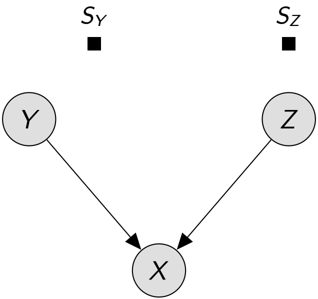


Here $T = \{Y\}$ and $M = \{Y, Z\}$ are not disjoint.

Semi-Markovian Models

1. Mutable Target

In some cases, the target itself is mutable:



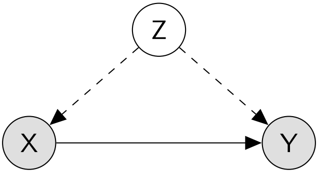
Here $T = \{Y\}$ and $M = \{Y, Z\}$ are not disjoint.

We use $p_{G_{\bar{T}}}(T|do(M \setminus T), O \setminus M) = p_{G_{\bar{Y}}}(Y|do(Z), X) = p(X|Y, Z)$.

Semi-Markovian Models

2. Non-identifiability

Sometimes it's not possible to evaluate an interventional quantity from observational data.



$p(Y|do(X))$ is non-identifiable as we don't know how much of the correlation in $p(Y|X)$ is due to the unobserved confounder.

For more complex DAGs we can test for identifiability using the ID algorithm. The idea is that if $p(T|do(M), O \setminus (M \cup T))$ is not identifiable, we try conditioning on a smaller set, i.e. $p(T|do(M), W)$.

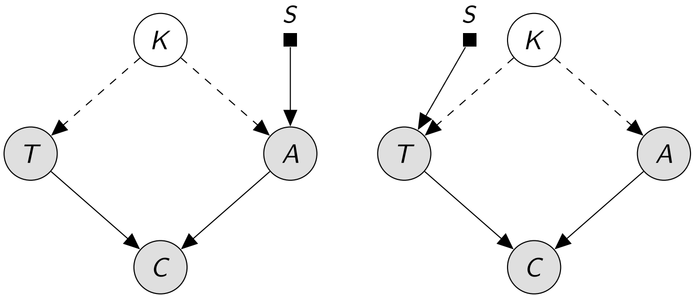
Overall Algorithm

```
Algorithm 2: Graph Surgery Estimator
input  : ADMG 𝒢, mutable variables M, target T
output: Expression for the surgery estimator or
        FAIL if there is no stable estimator.
Let S_ID = ∅; Let Loss = ∅;
for Z ∈ 𝒫(O \ (M ∪ {T})) do
    if T ∉ M then
        Let X, Y = UQ(M, {T}, Z; 𝒢);
        try
            P = ID(X, Y; 𝒢);
            P_s = P / Σ_T P;
            Compute validation loss ℓ(P_s);
            S_ID.append(P_s); Loss.append(ℓ(P_s));
        catch
            pass;
    Let X, Y = UQ(M, {T}, Z; 𝒢_{\bar{T}});
    X = X ∪ {T}; Y = Y \ {T};
    if Y ∩ (T ∪ ch(T)) = ∅ then
        continue;
    try
        P = ID(X, Y; 𝒢);
        P_s = P / Σ_T P;
        Compute validation loss ℓ(P_s);
        S_ID.append(P_s); Loss.append(ℓ(P_s));
    catch
        continue;
if S_ID = ∅ then
    return FAIL;
return P_s ∈ S_ID with lowest corresponding Loss;
```

Table of Contents

Simulated Data

Data is simulated according to the following two models, where the conditional probability distributions are linear Gaussian systems.



The environments differ in the coefficients of K in the structural equation for A, T respectively.

Simulated Data

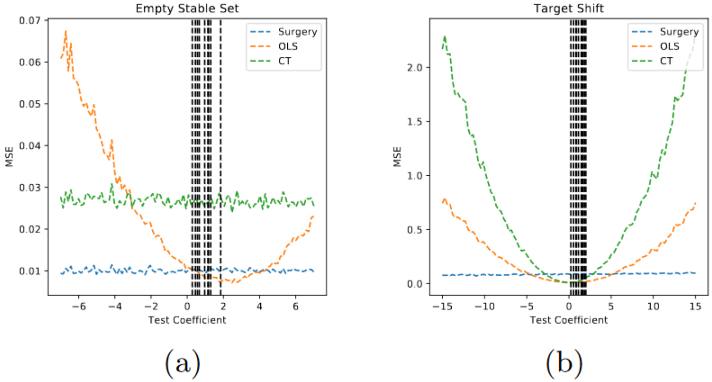
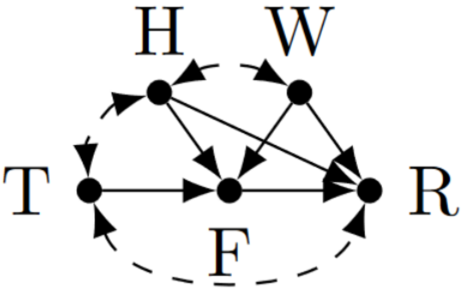


Figure 3: (a) MSE in test environments for the Fig $\boxed{1}$a scenario. (b) MSE in test environment for target shift scenario. Vertical lines denote training environments.

Bike Rentals

We wish to predict hourly bike rentals (R), on the basis of temperature T, feeling temperature F, wind speed W, and humidity H.



The data (over 2 years) is partitioned by season and year to create environments with different mechanisms. The mutable mechanisms are assumed to be $M = \{H, T, W\}$.

Bike Rentals

Table 1: MSE on the Bike Sharing Dataset

Test Data	OLS	AR	CT	Surgery
(Y1) Season 1	20.8±0.10	**20.5**±0.10	42.2±2.04	20.7±0.36
Season 2	**23.2**±0.05	**23.2**±0.05	29.9±0.09	23.8±0.09
Season 3	32.2±0.14	31.4±0.13	32.2±0.14	**29.9**±0.26
Season 4	29.2±0.08	29.1±0.08	29.1±0.08	**28.2**±0.07
(Y2) Season 1	32.5±0.11	**32.2**±0.11	32.6±0.15	36.1±0.37
Season 2	39.3±0.11	**39.2**±0.11	46.1±0.12	39.5±0.13
Season 3	47.7±0.17	**46.7**±0.16	48.2±0.22	54.8±0.73
Season 4	46.2±0.16	46.0±0.16	46.1±0.16	**44.4**±0.16

Table of Contents

Proactive robustness, formalized

Recall that we wish to predict Y from X, in a way that achieves good performance over all environments. One possible way to formulate this is to be adversarial over environments:

$$\min_f \max_e \mathbb{E}_e \left[L(f(X), Y) \right]$$

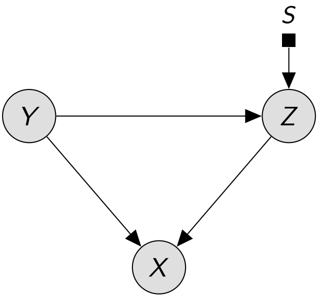
In the context of causal diagrams, we can write this as a maximum over settings of the mutable variables M:

$$\min_f \max_m \mathbb{E}_{p(\cdot | do(M=m))} \left[L(f(X), Y) \right]$$

This means that the optimal predictor f should achieve **uniform risk** over the possible settings.

Proactive robustness, formalized

Example



$$\min_f \max_z \int \int L(f(X,Z),Y)p(Y)p(X|Y,Z)dxdy$$

Stable estimators are not necessarily optimal...

Conclusion

- In many scenarios, it is desirable to obtain a classifier which is **proactively** robust against distribution shift;
- Causal approaches which take into can more accurately model unreliability in the DGP;
- There are still interesting questions regarding the tradeoff between performance/stability, and whether there might be other causal methods which can improve upon graph surgery

Duchi, J. and Namkoong, H. (2018).
Learning Models with Uniform Performance via Distributionally Robust Optimization.
arXiv e-prints, page arXiv:1810.08750.

Zhang, K., Gong, M., and Scholkopf, B. (2015).
Multi-source domain adaptation: A causal view.
In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3150–3157. AAAI Press.