Index

Introduction

Background

Main body

Problem Formulation

MPDAG : identifiability

Counterfactual fairness in MPDAG

Experiments

Conclusion

Counterfactual Fairness with Partially known Causal graph

Yeonho Jung

Seoul National University

January 29, 2023

Contents

Index

Introduction

Background

Main body

Problem Formulation

MPDAG : identifiability

Counterfactual fairness in MPDAG

Experiments

Conclusion

1. Introduction
2. Background
3. Problem Formulation
4. Identifiability of ancestral relations in MPDAGs
5. Counterfactual fairness in MPDAGs
6. Experiments
7. Conclusion

1. Introduction(1/2)

Fair machine learning...

- To avoid treating samples unfairly based on **sensitive attributes**
- To achieve the notion of **counterfactual fairness** when the true causal graph is unknown

'Counterfactual' :

- a probablilistic answer to a "what would have happened if" question

Interestingly,

Counterfactual fairness can be achieved as if the true causal graph were fully known when specific background knowledge is provided.

1. Introduction(2/2)

Can we learn causal fairenss with a partially known causal graph, MPDAG?

In MPDAG, with respect to a variable S, a variable T can be :

- **a definite descendant of S** : if T is a descendant of S
- **a definite non-descendant of S** : if T is a non-descendant of S
- **a possible descendant of S** : if T is neither a definite descendant nor a definite non-descendant of S

2. Background(1/5)

Structural Causal Model and Causal Graph

- Structural causal model(SCM)
  - A framework to model causal relations b/w variables
  - $V_i = f_i(pa_i, U_i)$
  - The set of equations F induces a causal graph D
- DAG(Directed Acyclic Graph)
  - All edges are directed and no directed cycle in the graph
  - If some edges are undirected, it is a PDAG(partially DAG)

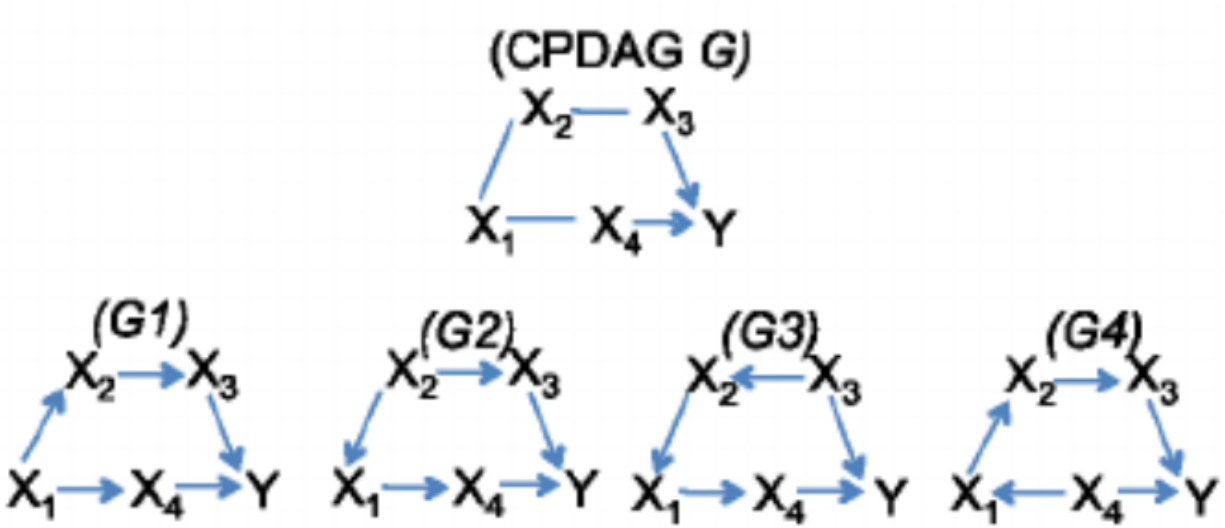
2.Background(2/5)

Structural Causal Model and Causal Graph

- CPDAG(Completed partially DAG, $[g^*]$)
  - $[g^*]$ represents a Markov equaivalence class of a DAG
  - Multiple DAGs are Markov equivalent ; if encoded the same set of conditional independence relations
- MPDAG(Maximally oriented PDAG, $[g]$)
  - CPDAG with background knowledge constraints

2. Background(3/5)

- Marcov equivalence class
  - All DAGs in a Marcov equivalence class have the same skeleton and the same v-structures
  - They can be uniquely represented by a CPDAG



2. Background(4/5)

The Paper
Yeonho Jung
Index
Introduction
Background
Main body
Problem Formulation
MPDAG : identifiability
Counterfactual fairness in MPDAG
Experiments
Conclusion

- Causal Graphs : DAG, CPDAG, MPDAG

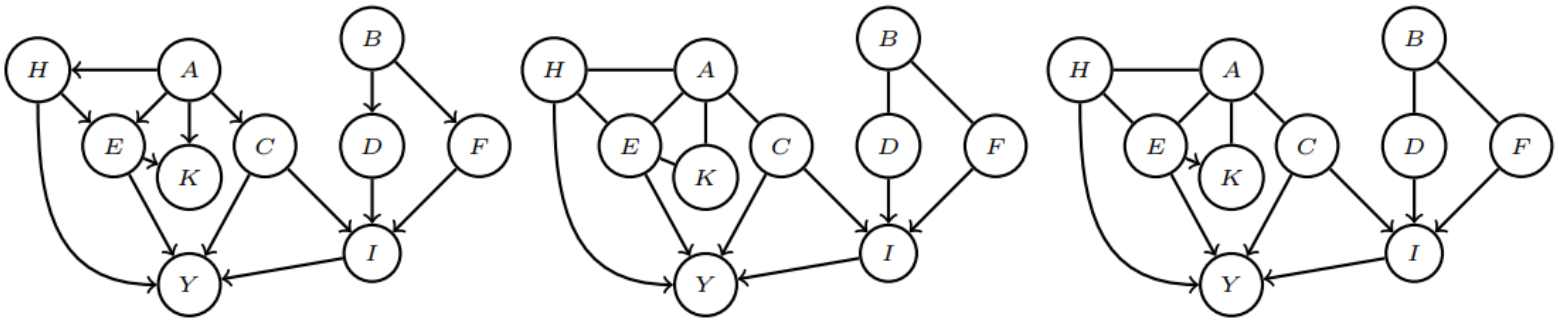
(a) DAG with 10 nodes and 10 directed edges

(b) CPDAG [g*]

(c) MPDAG [g] : CPDAG + with the background knowledge

* E is a direct cause of K

F.1 Causal graphs for one simulation



(a) DAG $\mathcal{D}$ (b) CPDAG $\mathcal{G}^*$ (c) MPDAG $\mathcal{G}$

2. Background(5/5)

The Paper

Yeonho Jung

Index

Introduction

Background

Main body

Problem Formulation

MPDAG : identifiability

Counterfactual fairness in MPDAG

Experiments

Conclusion

Counterfactual Fairness : Fairness criterion based on SCM

Definition (Counterfactual Fairness)

We say the prediction $\widehat{Y}$ is counterfactually fair if under any context $\chi = x$(observable attribute) and $A = a$(sensitive attribute),

$$P\left(\hat{Y}_{A\leftarrow a}(U) = y \mid \mathcal{X} = \mathrm{x}, A = a\right) = P\left(\hat{Y}_{A\leftarrow a'}(U) = y \mid \mathcal{X} = \mathrm{x}, A = a\right)$$

, for all y and any value a' attainable by A

Lemma

Let [g] be the causal graph of the given model (U, V ,F). Then,

$\hat{Y}$ *;* ***counterfactually fair*** *if it's a function of non-descendants of A*

A fair classifier give the same prediction had the person had a different sensitive attribute **to design a counterfactually fair model**

3.Problem Formulation

- Achieving counterfactual fairness given MPDAG
  - Counterfactuall fairness prediction can be framed as **Selecting the non-descendants of A to predict Y**
  - Not all ancestral relations b/w A and attributes in X are identifiable in a CPDAG or MPDAG
- To achieve counterfactually fair prediction
  1. 3 types of descendants
     - definite non-descendants of A
     - definite descendants of A
     - possible descendants of A
  2. Build a counterfactually fair model based on identified ancestral relations

4.Identifiability of ancestral relations in MPDAGs

The Paper
Yeonho Jung
Index
Introduction
Background
Main body
Problem Formulation
MPDAG : identifiability
Counterfactual fairness in MPDAG
Experiments
Conclusion

Theorem

Let A and F be two distinct vertices in an MPDAG g, and S be the b-critical set of A with respect to F in g. Then F is a definite descendant of A iff either A has a definite arrow into S, that is $X \cap ch(A, g) \neq \phi$*, or A does not have a definite arrow into S but S is non-empty and induces an imcomplete subgraph of g*

We identify whether T is a definite descendant of S in MPDAG

1. possible descendants of A : B,C,D,E and H
2. b-critical set of A as to F : B,C,D
3. F : a definite descendant of A

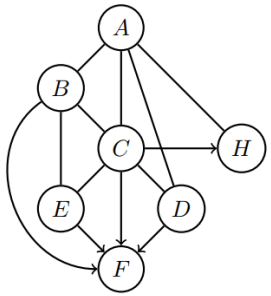


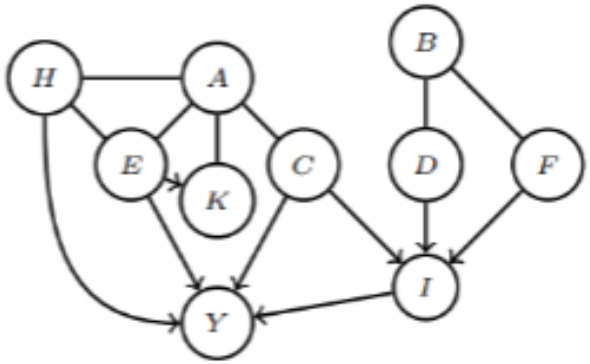
Figure 5: An MPDAG $\mathcal{G}$ for illustrating ancestral relations of the node A with any other nodes. The node B, C, D, E and H are possible descendants of A; node F is a definite descendant of A.

5. Counterfactual fairness in MPDAGs(1/2)

The Paper
Yeonho Jung
Index
Introduction
Background
Main body
Problem Formulation
MPDAG : identifiability
Counterfactual fairness in MPDAG
Experiments
Conclusion

- Propose two methods

1. Fair
   - Making predictions using all definite non-descendatns of the sensitive attribute in MPDAG
   - the number of definite non-descendants in an MPDAG is too small : low prediction accuracy
2. FairRelax
   - Making predictions using all definite non-descendants and possible descendants of sensitive attribute in MPDAG



(c) MPDAG $\mathcal{G}$

5. Counterfactual fairness in MPDAGs(2/2)

- Assumption 5.1.
  - **Sensitive attribute can only be a root node in MPDAG**
  - Ex) gender cannot be caused by 'education or salary'
- Proposition 5.2.
  - In an MPDAG with sensitive attribute A, if 'Assumption 5.1' holds, **any other attribute is either a definite descendant or a definite non-descendant of A**
- Fitting a model with the definite non-descendants of A is same thing as fitting a model with the non-descendants of A in the true DAG.
  - **Thus, counterfactual fairness can be achieved as if true causal DAG is fully known**

Experiments(1/3)

Proposed methods : Fair and FairRelax

- Baselines

1) Full : standard model using all attributes
2) Unaware : Using all attributes except sensitive attributes
3) Oracle : Using all attributes that are non-descendants of the sensitive attribute **given the groundtruth DAG**
4) Fair : using all definite non-descendants of the sensitive attribute in **MPDAG**
5) FairRelax : Using all definite non-descendants and possible descendants of sensitive attribute in **MPDAG**

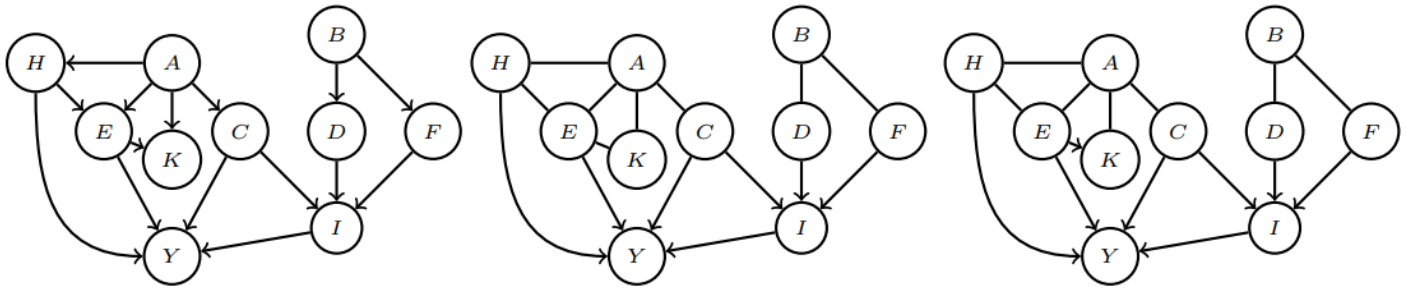
Experiments(2/3)

The Paper
Yeonho Jung
Index
Introduction
Background
Main body
Problem Formulation
MPDAG : identifiability
Counterfactual fairness in MPDAG
Experiments
Conclusion

Dataset

- 1. synthetic data
  - Simulated DAG is known and randomly generate 100 DAGs
  - True (simulated) DAG $\rightarrow$ CPDAG $\rightarrow$ MPDAG
  - Sensitive attribute: 2 or 3 values from Binomial(Multinomial) distribution
- 2. real data
  - 395 students with 32 attributes and 'sex' as the sensitivity attribute -> predict 'Grade'

F.1 Causal graphs for one simulation



(a) DAG $\mathcal{D}$ (b) CPDAG $\mathcal{G}^*$ (c) MPDAG $\mathcal{G}$

Experiments(2/3)

The Paper
Yeonho Jung
Index
Introduction
Background
Main body
Problem Formulation
MPDAG : identifiability
Counterfactual fairness in MPDAG
Experiments
Conclusion

2. Counterfactual fairness
   - Evaluate the counterfactual fairness by the absolute difference of two values : $\hat{Y}_{A\leftarrow a}(U)$, $\hat{Y}_{A\leftarrow a'}(U)$

F.2 Density plot for simulated data

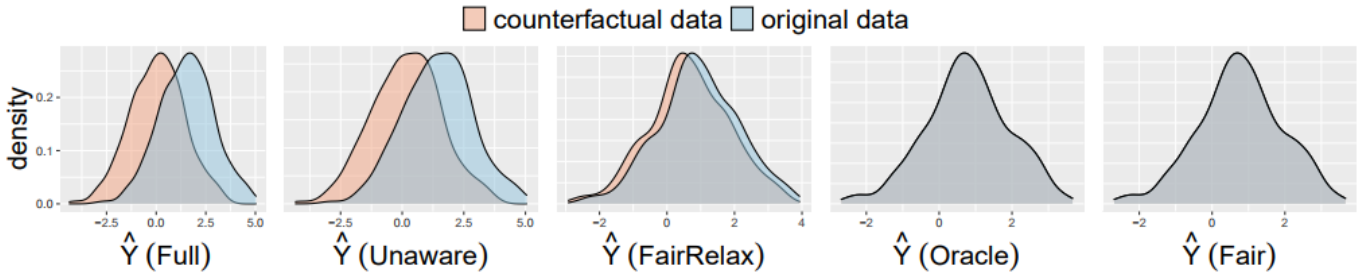


Figure 7: Density plot of the predicted $Y_{A\leftarrow a}(u)$ and $Y_{A\leftarrow a'}(u)$ in synthetic data.

Experiments(3/3)

3. Accuracy
   1. 'Full' model : the lowest RMSE (not surprising)
   2. 'FairRelax' : better accuracy than 'Fair' and 'Oracle' model

Table 1: Average unfairness and RMSE for synthetic datasets on held-out test set. For each graph setting, the unfairness gets decreasing from left to right and the RMSE gets increasing from left to right.

	Node	Edge	Full	Unaware	FairRelax	Oracle	Fair
Unfairness	10	20	0.288 ± 0.363	0.200 ± 0.322	0.023 ± 0.123	0.000 ± 0.000	0.000 ± 0.000
	20	40	0.203 ± 0.341	0.165 ± 0.312	0.019 ± 0.145	0.000 ± 0.000	0.000 ± 0.000
	30	60	0.155 ± 0.304	0.143 ± 0.312	0.020 ± 0.123	0.000 ± 0.000	0.000 ± 0.000
	40	80	0.095 ± 0.189	0.075 ± 0.182	0.009 ± 0.055	0.000 ± 0.000	0.000 ± 0.000
RMSE	10	20	0.621 ± 0.251	0.637 ± 0.261	1.031 ± 0.751	1.065 ± 0.751	1.137 ± 0.824
	20	40	0.595 ± 0.255	0.599 ± 0.253	0.818 ± 0.488	0.847 ± 0.55	0.952 ± 0.645
	30	60	0.597 ± 0.24	0.601 ± 0.242	0.797 ± 0.489	0.849 ± 0.644	1.024 ± 0.908
	40	80	0.600 ± 0.273	0.601 ± 0.272	0.755 ± 0.441	0.766 ± 0.452	0.800 ± 0.480

Conclusion

We treated a general approach to achieve counterfactual fairness using MPDAGs when true DAG is unknown

- We can achieve counterfactual fairness without a true causal DAG to be specified.

- Finished -