Counterfactual Fairness under Hidden Confounding

Marc Braun MARC.BRAUN@LIU.SE
Department of Computer and Information Science, Linköping University
Jose M. Peña JOSE.M.PENA@LIU.SE
Department of Computer and Information Science, Linköping University
Adel Daoud ADEL.DAOUD@LIU.SE
Institute for Analytical Sociology, Linköping University
Chalmers University of Technology

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

Designing algorithms that are non-discriminatory has attracted major attention in recent years. It has been shown that a causal perspective is necessary to mitigate unfair discrimination. However, little attention has been paid to how algorithms can make fair decisions under hidden confounding. In this paper, we study counterfactual fairness in the presence of hidden confounding and make two contributions. First, we show that under commonly adopted assumptions, counterfactual fairness is invariant across all causal models consistent with the observed data and causal ordering of the variables. In particular, even in the presence of hidden confounding and non-identifiability of the underlying causal model, all compatible models yield the same counterfactual fairness conclusions. Second, we show how instrumental variables, a tool commonly used in causal inference to address hidden confounding, can help to design fair predictors when the protected attribute is unobserved. Together, our results extend the scope of causal fairness methods to more realistic settings where both confounders and sensitive attributes may be partially or entirely unobserved.

Keywords: Counterfactual Fairness, Hidden Confounding, Instrumental Variables

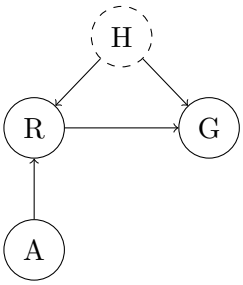
1. Introduction

Machine Learning (ML) methods are increasingly deployed in sensitive domains such as insurance or criminal justice. In these settings, ensuring that predictive models do not discriminate against protected groups is of central importance.

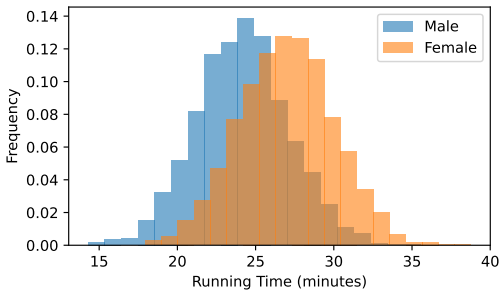
To illustrate this, consider the following toy example. Let A be the sex of a person and let R be the time it took them to run a distance of 10 km. Additionally, we assume that the physical-education (PE) grade G they got in school and their insurance rate Y are known. However, assume we do not have access to the underlying physical health H, which we assume is a confounder for the running time R and PE grade G as shown in the causal graph in Figure 1(a). The insurance company may want to train a ML model that predicts the health-insurance rate Y[1] for new customers without discriminating against individuals based on sex. Sex is therefore referred to as the *protected attribute*. A predictor $\hat{Y}(R, G)$ trained to minimize prediction error may nevertheless discriminate based on sex because running time partly reflects sex differences. In contrast, a predictor depending only on the latent health status H would remain invariant under interventions on A.

1. Y is not included the causal graphs as it cannot be used as input to the predictor since it is unobserved at inference time. The causal relationship to other variables is irrelevant for the prediction task.

© 2026 M. Braun, J.M. Peña & A. Daoud.



(a) Causal graph with health status H as hidden confounder.



(b) Histogram of running times R.

Figure 1: Relationship between sex A, running time R, and PE grade G, along with the observed distribution of R in our toy example.

While different definitions of fairness exist (Makhlouf et al., 2024), we will focus on the notion of *counterfactual fairness* (Kusner et al., 2017), which defines a predictor as fair if it produces the same output under an intervention on the protected attribute. In our example, changing an individuals sex in a counterfactual world (resulting in different values of R and G), should not change the prediction (more detail in Section 2).

A predictor depending only on non-descendants of the protected attribute is trivially counterfactually fair. In many applications, however, all observed variables are descendants of the protected attribute. Kusner et al. (2017) propose to base predictions on latent exogenous variables. Under assumptions such as strictly monotone structural equations and no hidden confounding, these variables can be identified up to counterfactual equivalence (Braun et al., 2026). In the presence of hidden confounding, however, multiple latent variable distributions can induce the same observed data distribution while leading to different counterfactuals, raising concerns about identifiability.

To avoid this non-identifiability problem, existing literature on counterfactual fairness has therefore mostly assumed that there are no hidden confounders (Kusner et al., 2017; Chiappa, 2019; Zuo et al., 2022). Other literature has proposed a method aiming at evaluating the extent of unfairness that is caused by hidden confounding (Kilbertus et al., 2017). This leaves open the question of whether counterfactual fairness is fundamentally ill-posed in such settings. In this work, we show that this is not necessarily the case and make two main contributions.

1. We show that counterfactual fairness can remain well-defined despite hidden confounding. In particular, under standard structural assumptions, all causal models consistent with the observed data and causal ordering are counterfactually equivalent with respect to the protected attribute. As a result, predictors learned under unconfoundedness assumptions yield the same fairness conclusions even when these assumptions are violated.

2. We propose a method based on instrumental variables (IVs) to construct fair predictors in settings where the protected attribute is unobserved and hidden confounding is present.

The remainder of this paper is structured as follows. We formally introduce the concepts like counterfactual fairness and IVs used in this paper in Section 2. We then review related work in Section 3 and present our own contribution in Section 4. Lastly, we show simulated and real-world experiments in Section 5 and close with a conclusion in Section 6.

2. Background

In this section, we review the causal inference framework and fairness concepts required to formalize our problem and results. We start by formally introducing structural causal models (SCMs).

Definition 1 (SCM) *A* SCM *is a tuple $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U}))$, where $\mathcal{U}$ is a set of exogenous variables and $\mathcal{X}$ is a set of endogenous variables. $\mathcal{F}$ is a set of functions, one for each $X \in \mathcal{X}$, that define structural equations of the form $X = f_X(U_X, \mathrm{Pa}_X)$, where $U_X \in \mathcal{U}$ and $\mathrm{Pa}_X \subseteq \mathcal{X} \setminus \{X\}$ are the direct causes of X. $P(\mathcal{U})$ is a joint probability distribution over the exogenous variables.*

When some or all of the variables in $\mathcal{X}$ are indexed, we may write Pa_i, f_i, U_i for $X_i \in \mathcal{X}$ instead of $\mathrm{Pa}_{X_i}, f_{X_i}, U_{X_i}$. Non-independence of variables U_i and $U_j \in \mathcal{U}$ indicates exogenous confounding between the variables X_i and $X_j \in \mathcal{X}$.

A SCM can be thought of as a model of reality, that describes how observations of $\mathcal{X}$ are generated through the causal mappings in $\mathcal{F}$ from background variables $\mathcal{U}$. Each SCM induces a causal graph with a directed edge from X_i to $X_j \in \mathcal{X}$ if $X_i \in \mathrm{Pa}_j$ and hence we call X_i a parent of X_j and X_j a child of X_i. The set of all children and children's children etc. is called the descendants. From now on, we will consider only SCMs that induce directed acyclic graphs (DAGs). We furthermore assume that the variables in $\mathcal{U}$ are unobserved.

Definition 2 (Interventional SCM) *Let $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U}))$ be an SCM. The* interventional SCM *$\mathcal{M}_{X \leftarrow x'}$ is defined by the tuple $(\mathcal{U}, \mathcal{X}, \mathcal{F}_{X \leftarrow x'}, P(\mathcal{U}))$ where $\mathcal{F}_{X \leftarrow x'}$ is obtained by replacing the function $f_X(U_X, \mathrm{Pa}_X) \in \mathcal{F}$ with $\tilde{f}_X(U_X, \mathrm{Pa}_X) \coloneqq x'$ where x' is a fixed value. The remaining structural equations are unchanged.*

Interventions help us to capture causal effects, because they break all incoming arrows in the intervened variable and therefore ensure that any association between the intervened variable and any other variable is due to direct or indirect causal effects.

Definition 3 (Counterfactual) *Let $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U}))$ be an SCM, and let $X, Y \in \mathcal{X}$. For a given realization $U = u$ of the exogenous variables, the* counterfactual variable *$Y_{X \leftarrow x'}(u)$ is defined as the value of Y obtained by evaluating the structural equations of the intervened model $\mathcal{M}_{X \leftarrow x'}$ with $U = u$.*

The corresponding counterfactual distribution *is given by*

$$P\big(Y_{X \leftarrow x'} = y \mid E = e\big) = \int \mathbf{1}\{Y_{X \leftarrow x'}(u) = y\} \, \mathrm{d}P(u \mid E = e),$$

where $E \subseteq \mathcal{X}$ denotes the set of observed variables and e their observed realisations.

Counterfactuals compare outcomes across hypothetical worlds that share the same latent realization while differing through an intervention. This requires a three-step procedure: (i) abduction, where u is inferred from evidence, (ii) action, where an intervention is applied, and (iii) prediction, where the outcome is computed under the modified model.

Definition 4 (Counterfactual Equivalence) *Let* $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U}))$ *and* $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{U}}, \mathcal{X}, \widetilde{\mathcal{F}}, P(\widetilde{\mathcal{U}}))$ *be two SCMs over the same set of endogenous variables* $\mathcal{X}$*. The two SCMs are said to be* counterfactually equivalent with respect to interventions on $X \in \mathcal{X}$ *if they induce the same counterfactual distribution* $P(Y_{X \leftarrow x'} \mid E = e)$ *for all* $Y \in \mathcal{X}$ *and* $E \subseteq \mathcal{X}$*.*

If the two SCMs are counterfactually equivalent with respect to interventions on all $X \in \mathcal{X}$ *we say the two SCMs are* counterfactually equivalent*.*

It has been argued that designing fair algorithms for prediction tasks requires reasoning about causality (Loftus et al., 2018). While different notions of fairness exist (Dwork et al., 2012; Hardt et al., 2016; Zafar et al., 2017), we will use the concept of counterfactual fairness that was introduced by Kusner et al. (2017).

Definition 5 (Counterfactual Fairness) *Let* $\mathcal{M}$ *be an SCM, let* $A \in \mathcal{X}$ *be a protected attribute, and let* $\hat{Y}$ *be a predictor. The predictor* $\hat{Y}$ *is said to be* counterfactually fair *if under any evidence* $E = e$ *and* $A = a$ *with* $E \subseteq \mathcal{X} \setminus \{A\}$ *we have*

$$P\big(\hat{Y}_{A \leftarrow a} = y \mid E = e, A = a\big) = P\big(\hat{Y}_{A \leftarrow a'} = y \mid E = e, A = a\big).$$

for all a' *and all* y*.*

According to this definition, a predictor is fair if its output is the same in a counterfactual world where the individual belonged to a different protected group.

Causal Graphical Normalizing Flows A causal graphical normalizing flow (CGNF) is special normalizing flow generating the observed joint distribution of the variables $\mathcal{X}$ in an SCM $\mathcal{M}$ from a latent distribution $P(\mathcal{V})$. It can be used for causal and counterfactual inference (Balgi et al., 2024). The normalizing flow consists of a conditional normalizing flow h_X for each $X \in \mathcal{X}$ where the conditioning is on $\mathrm{Pa}_X \subseteq \mathcal{X} \setminus \{X\}$.

The parameters of each h_X are then chosen such that $X \overset{d}{=} h_X(\mathrm{Pa}_X, V_X)$ where $\overset{d}{=}$ denotes equality in distribution and each h_X is strictly monotonic with respect to $V_X \in \mathcal{V}$. The latent joint distribution over the variables in $\mathcal{V} = \{V_X \mid X \in \mathcal{X}\}$ that generate the observed variables is typically chosen to be the standard multivariate normal distribution $\mathcal{N}(\mathbf{0}, I_n)$. The parameters of the functions h_X are optimized via maximum-likelihood.

When each function f_X in $\mathcal{F}$ of the SCM $\mathcal{M}$ is strictly monotonic and continuous with respect to U_X and when the distribution of the exogenous variables in $\mathcal{M}$, $P(\mathcal{U})$, is equal to the distribution of the latent variables in the CGNF, $P(\mathcal{V})$, then the structural equations in the SCM are identified as $f_X = h_X$. Thus, the SCM induced by the CGNF $\widetilde{\mathcal{M}} = (\mathcal{V}, \mathcal{X}, \mathcal{H}, P(\mathcal{V}))$ with $\mathcal{H} = \{h_X \mid X \in \mathcal{X}\}$ is counterfactually equivalent to $\mathcal{M}$.

Instrumental Variables IVs can be used to mitigate bias from hidden confounding when estimating causal effects between a treatment variable T and an outcome variable Y (Wald, 1940; Angrist et al., 1996). A variable Z is called IV if it satisfies three conditions.

1. The *relevance condition* $Z \not\perp\!\!\!\perp T$ requires that the instrument must be associated with the treatment.

2. The *exclusion restriction* requires that the instrument Z can affect the outcome Y only through the treatment T.

3. *Marginal exchangeability* requires that the instrument and outcome do not share common causes.

In the example in the introduction, when R is the treatment and G the outcome, a reasonable instrument Z could be the temperature on the day of the run, because it affects the running time R but only affects the PE grade G through R.

Many different methods that utilize the existence of IVs to handle hidden confounding have been proposed in the past (Imbens and Angrist, 1994; Robins, 1994; Guo and Small, 2016; Hahn and Ridder, 2017; Puli and Ranganath, 2020; Hartford et al., 2017).

In an SCM with an instrument Z, treatment T, and outcome Y, Braun et al. (2026) have shown that the outcome function f_Y is identifiable up to counterfactual equivalence by training a CGNF where the joint distribution of (V_T, V_Y) itself is learned via another normalizing flow. The authors call such model *Flow IV*. When the variables are one-dimensional and continuous, the only requirement for the result to hold is that the outcome function is strictly monotonic with respect to U_Y and that the instrument is a so called strong IV (i.e. $P(T = t \mid Z = z) > 0$). No assumption about the absence of hidden confounders between T and Y is required.

3. Related Work

Algorithmic fairness has been studied extensively, with early work focusing on statistical criteria such as fairness through awareness (Dwork et al., 2012), equality of opportunity (Hardt et al., 2016), and related notions (Zafar et al., 2017). More recent work has emphasized the importance of causal reasoning for fairness, arguing that purely observational criteria may fail to capture discriminatory mechanisms (Kilbertus et al., 2017; Loftus et al., 2018; Makhlouf et al., 2024). Within this line of work, counterfactual fairness (Kusner et al., 2017) has emerged as a prominent framework.

A central challenge in counterfactual fairness is that it relies on a structural causal model, which is typically not identifiable from observational data alone. Existing approaches therefore impose assumptions such as the absence of hidden confounders to enable the identification of the relevant counterfactual quantities. Kusner et al. (2017) construct fair predictors based on latent background variables, while Chiappa (2019) extend this idea to path-specific counterfactual fairness. Other work considers settings where the causal graph is only partially known, but still assumes the absence of hidden confounding (Zuo et al., 2022; Ma et al., 2022).

Several works have highlighted the limitations of these assumptions. Wu et al. (2019) study the non-identifiability of counterfactual fairness and propose bounds and algorithms under model uncertainty. Kilbertus et al. (2020) analyse the sensitivity of counterfactual fairness to unmeasured confounding in additive noise models. In contrast, we show that

under mild structural assumptions, all causal models consistent with the observed distribution are counterfactually equivalent with respect to interventions on the protected attribute, implying that fairness conclusions remain invariant despite hidden confounding.

Only few works address counterfactual fairness under hidden confounding directly. Ling et al. (2025) consider graph-structured data and propose identifiable latent variable models learned via variational inference to capture unobserved confounders. In contrast, we focus on standard structural causal models and derive conditions under which counterfactual fairness remains well-defined even when the underlying causal model is not identifiable.

IVs are a classical tool in causal inference for addressing hidden confounding. While their use in fairness has been limited, Centorrino et al. (2022) study fairness constraints in structural econometric models and propose IV-based estimation procedures to address endogeneity when the protected attribute is observed. Recent work has also shown how IVs can be used for counterfactual inference in nonseparable models (Braun et al., 2026). In contrast, we demonstrate how IVs can be leveraged to construct counterfactually fair predictors in the presence of hidden confounding when the protected attribute itself is unobserved.

4. Counterfactual Fairness under Hidden Confounding

Most existing work on counterfactual fairness assumes the absence of hidden confounding. In this section, we study when counterfactual fairness remains well-defined despite such confounding. We first show that, under mild structural assumptions, all causal models consistent with the observed data are counterfactually equivalent with respect to interventions on the protected attribute. Consequently, fairness conclusions are invariant to hidden confounding even when the underlying causal model is not identifiable. We then show how IVs can be used to construct counterfactually fair predictors when the protected attribute is unobserved.

4.1. Counterfactual Fairness Under Hidden Confounding when the Protected Attribute is Observed

Assume that SCM $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U}))$ is the SCM that generated a given data set, and SCM $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{U}}, \mathcal{X}, \widetilde{\mathcal{F}}, P(\widetilde{\mathcal{U}}))$ is an alternative SCM that induces the same observed data distribution $P(\mathcal{X})$ and causal ordering of the variables in $\mathcal{X}$. In other words, $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are observationally equivalent and we can therefore not decide from observed data alone which of the two is a better model for the data generating process (DGP). Let $A \in \mathcal{X}$ be a protected attribute and let $\mathcal{X} \setminus \{A\} = \{X_1, \ldots, X_n\}$ be the remaining observed variables that can be used as input to a predictor for the prediction target Y. We introduce two assumptions.

Assumption 6 (Strict Monotonicity) *Let $\mathcal{S}_i$ denote the support of the joint distribution of the parent variables Pa_i. For each $i \in \{1, \ldots, n\}$, assume that U_i and $\tilde{U}_i$ are one-dimensional and absolutely continuous and define*

$$f_i(\cdot, p) \colon \mathbb{R} \longrightarrow \mathbb{R}, \quad u \longmapsto f_i(u, p),$$

for each $p \in \mathcal{S}_i$. Define $\tilde{f}_i(\cdot, p)$ analogously.

Assume that, for each $i \in \{1, \ldots, n\}$, the functions $f_i(\cdot, p)$ and $\tilde{f}_i(\cdot, p)$ are continuous and either strictly increasing for every $p \in \mathcal{S}_i$, or strictly decreasing for every $p \in \mathcal{S}_i$.

Assumption 6 is weaker than the additive noise assumptions commonly used in prior work (Kilbertus et al., 2020). It implies that the variables X_i are continuous and one-dimensonal, while the protected attribute A and prediction target Y are unrestricted and may be arbitrary random variables (e.g. continuous, discrete, high-dimensional). Intuitively, the monotonicity assumption requires that counterfactual interventions preserve the conditional relative rank induced by the latent noise variables. In the example from the introduction, a counterfactually fair predictor would then produce the same output for the fastest female with the highest PE grade among all females and the fastest male with the best PE grade among males. Under Assumption 6, the structural equations become invertible with respect to the latent variables given the parent variables. Consequently, counterfactual outcomes become deterministic.

Assumption 7 *We assume that the protected attribute A does not have any parents in the causal graph of $\mathcal{M}$ and in the causal graph of $\widetilde{\mathcal{M}}$, i.e. $\mathrm{Pa}_A = \emptyset$. We furthermore assume that $U_A \perp\!\!\!\perp (U_1, \ldots, U_n)$ and $U_A \perp\!\!\!\perp (\tilde{U}_1, \ldots, \tilde{U}_n)$.*

Assumption 7 implies that there are no observed or unobserved confounders that affect A and any other observed variable X_i. Assumption 7 is common in the fairness literature, and while some authors have criticized it for being too restrictive (Plečko and Bareinboim, 2024), others have argued that it holds in many real-world settings, because typically we want to ensure fairness with respect to attributes that are in some sense randomly assigned by nature, such as sex or age (Kusner et al., 2017). Under Assumptions 6 and 7, we obtain the following result.

Theorem 8 *The two SCMs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are counterfactually equivalent with respect to interventions on the protected attribute A.*

Theorem 8 implies that, under the given assumptions, counterfactual fairness is invariant across all observationally equivalent SCMs that induce the same causal ordering of the variables in $\mathcal{X}$.

Designing a Fair Predictor in Practice Given a dataset generated from SCM $\mathcal{M}$, and given that Assumptions 6 and 7 are true, we can train a CGNF to generate the observed distribution. In reality, the causal graph is often provided by experts a priori and hidden confounding is not identifiable from observed data. Then, it suffices to use the given causal graph with independent latent terms $V_1, \ldots, V_n$ when training the CGNF. According to Theorem 8, the SCM $\widetilde{\mathcal{M}}$ induced by the CGNF is counterfactually equivalent to the underlying SCM $\mathcal{M}$ with respect to interventions on the protected attribute A. Whenever a predictor $\hat{Y}$ only takes the latent variables $V_1, \ldots V_n$ of the CGNF as input, it will be counterfactually fair with respect to $\mathcal{M}$, because those variables will be unaltered in the counterfactual world.

If the causal graph provided by the domain expert is not compatible with the observed data under the no-hidden-confounding assumption, then either 1) the latent distribution

used in the CGNF needs to be adapted with methods such as Flow IV to ensure compatibility with observed data, or 2) the DAG used in the CGNF needs to be adapted such that using independent latent variables is compatible with the observed data. For example, a fully connected DAG (see Figure 7) and independent latent variables are compatible with *any* observed distribution, but the model may have a higher variance compared to a model under the true DAG.

4.2. Counterfactual Fairness Under Hidden Confounding when the Protected Attribute is Unobserved

When the protected attribute is unobserved, Theorem 8 no longer applies. However, IVs can still enable identification of latent representations that exclude information about the protected attribute while retaining predictive information about downstream variables.

In the example in the introduction, our data might not contain information about the sex of individuals. However, we may use the temperature on the day that the running examination was performed as an instrument Z for the running time R. It is not unreasonable to assume that the weather affects how fast an individual runs, but does not affect the PE grade other than through the running time.

We then have a setup like shown in Figure 2, where we summarize the effect of A and H in the variable U_R.

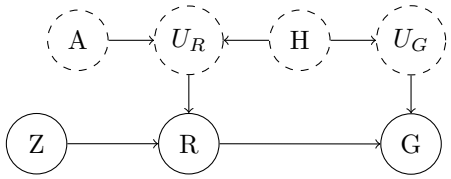


Figure 2: *Running Time* example extended by an IV capturing the temperature Z on the day of the run. The IV affects G only through R and is not confounded with G.

The only observed variable that is not a descendent of the protected attribute A is the instrument Z, which means that, without further preprocessing, this is the only variable that a counterfactually fair baseline predictor could use to predict Y. However, according to Braun et al. (2026), under Assumption 6 and the assumption that the IV is a strong IV, we can identify U_R and U_G up to counterfactual equivalence despite the presence of the hidden confounder H. Since U_R is a descendant of the protected attribute A, it cannot be used in the predictor. However, U_G does not contain any information about A and can therefore be used. U_G contains information about the overall physical health status of individuals H which might be strongly associated with the insurance rate. Therefore, adding U_G as input to the predictor can be beneficial compared to only relying on the instrument Z.

Proposition 9 *Given a causal graph as shown in Figure 2 and under Assumption 6, then (i) U_G is identifiable up to counterfactual equivalence (Braun et al., 2026); (ii) a predictor $\hat{Y}$ that is a function of Z and U_G is counterfactually fair.*

Proof Z and U_G are non-descendants of A, thus invariant under interventions on A. ■

Remark 10 *The usefulness of IVs for constructing counterfactually fair predictors depends on the causal role of the hidden protected attribute. If the protected attribute affects only G but not R, then R is not a descendant of the protected attribute and can therefore be used directly as input to a fair predictor. Because there exists a bijection between (Z, R) and (Z, U_R), both are equally informative.*

Conversely, if the protected attribute affects both R and G, then both U_R and U_G are descendants of the protected attribute and therefore cannot be used in a counterfactually fair predictor. Consequently, the setting shown in Figure 2, where the protected attribute affects R but not G, is the only configuration in which the IV enables identification of a latent representation that is both informative and counterfactually fair.

5. Experiments

We evaluate the theoretical results from Section 4 on both synthetic and real-world datasets. The synthetic experiments are designed to validate the invariance properties established in Theorem 8 and the usefulness of IVs when the protected attribute is unobserved. The real-world experiment studies whether the same empirical behaviour persists when the true data-generating process is unknown.

5.1. Synthetic Experiments

For the synthetic experiments, we train a CGNF as a proxy for the true SCM. Unless stated otherwise, we use the true causal graph and the latent variables $\mathcal{V}$ of the CGNF are assumed to be independent. This corresponds to the standard assumption of no hidden confounding. The learned latent representations are then used as input to a predictor.

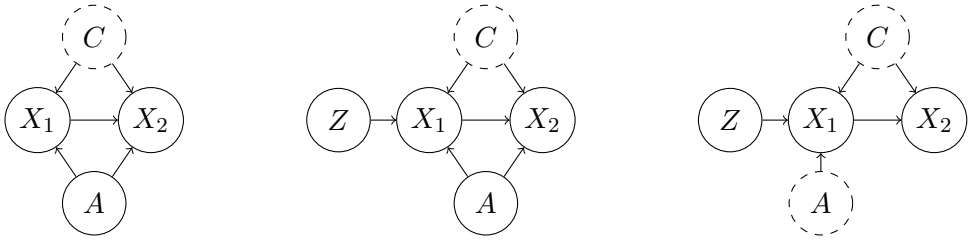
We compare the resulting predictors against several baselines. As an unfair baseline, we train a predictor that uses all observed variables except for the protected attribute A. Whenever observed non-descendants of A exist, we additionally include a fair baseline predictor that uses only those variables as input. In experiments where an IV is available, we also train a Flow IV model (Braun et al., 2026). Flow IV extends the CGNF framework by learning the latent distribution under hidden confounding up to counterfactual equivalence. The latent variables learned by the Flow IV model are then used as input to the predictor. For all methods, predictors are implemented as multi-layer perceptrons (MLPs) with identical architectures.

To evaluate fairness, we generate factual and counterfactual sample pairs from the underlying SCM. Following a notion of unfairness similar to that proposed by Kilbertus et al. (2020), we define the unfairness score as

$$\mathrm{UF} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{Y}(u_i) - \hat{Y}_{A \leftarrow a'}(u_i) \right)^2,$$

where $\hat{Y}_{A \leftarrow a'}(u_i)$ denotes the prediction for the corresponding counterfactual data point under an intervention on the protected attribute A. The value for the intervention a' is drawn randomly from the marginal distribution of the protected attribute. Lower values of UF indicate greater counterfactual fairness. We additionally report the mean squared prediction error (MSE) on the target variable.

We consider three synthetic settings, illustrated in Figure 3 and all DGPs can be found in Appendix B.1.



(a) Setup without IV. (b) Setup with IV and observed A. (c) Setup with IV and unobserved A.

Figure 3: DAGs of the synthetic experiments. Dashed nodes indicate unobserved variables.

Synthetic Experiment 1: Observed A The first experiment evaluates whether counterfactual fairness remains invariant under hidden confounding when the protected attribute is observed as claimed in Theorem 8. The causal graph is shown in Figure 3(a).

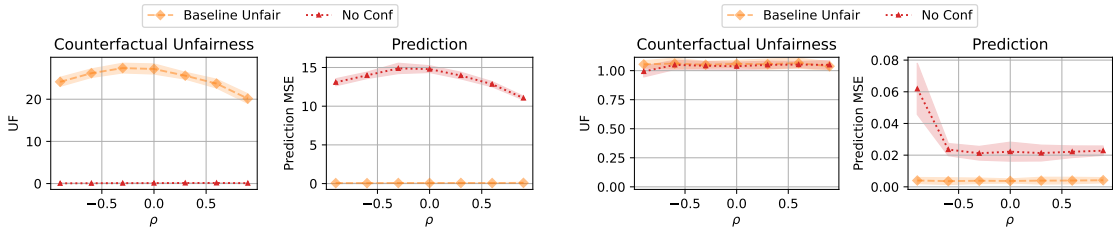
In the first part of the experiment, the DGP satisfies Assumptions 6 and 7. Figure 4(a) shows that the predictor based on the no-confounding CGNF remains almost perfectly fair across all confounding strengths, empirically illustrating Theorem 8. This demonstrates that fairness conclusions remain stable even when the latent confounding structure is misspecified. The unfair baseline achieves lower prediction error by exploiting information about the protected attribute encoded in X_1 and X_2.

In a variation of the experiment, we replace the structural equations of the DGP with equations that do not satisfy the strict-monotonicity assumption. Figure 4(b) shows that the unfairness is notably larger than zero and thereby illustrates that Assumption 6 appears to be necessary for Theorem 8 to hold.

Synthetic Experiment 2: Observed A with IV In the second experiment, we introduce an IV as shown in Figure 3(b). Because of the presence of the IV, the independent noise terms of the no-confounding CGNF and the true causal graph are not compatible with the observed data. We therefore use a fully connected DAG in the no-confounding CGNF. The experiment studies whether relying on the true DAG and explicitly modelling the latent confounding structure using Flow IV improves prediction quality over using a fully connected DAG in the no-confounding CGNF.

The results are shown in Figure 5(a). Consistent with Theorem 8, both approaches yield predictors that remain almost perfectly fair across all levels of confounding. In this experiment, the Flow IV predictor does not substantially improve prediction performance compared to the no-confounding CGNF, despite correctly recovering the latent confounding structure up to counterfactual equivalence. The unfair baseline achieves the lowest prediction error at the cost of substantial unfairness, whereas, the fair baseline predictor remains fair but discards all information in X_1 and X_2.

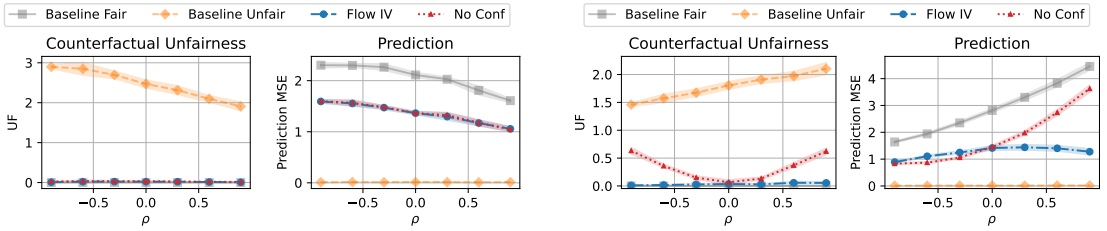
Synthetic Experiment 3: Unobserved A with IV The final synthetic experiment evaluates the setting where the protected attribute is unobserved (Figure 3(c)). As discussed



(a) The DGP satisfies the strict monotonicity assumption.

(b) The DGP violates the strict monotonicity assumption.

Figure 4: Prediction error and unfairness for the synthetic experiments with observed protected attribute A. $|\rho| \in [0, 1]$ controls the strength of hidden confounding.



(a) The protected attribute A is observed.

(b) The protected attribute A is unobserved.

Figure 5: Synthetic experiments with instrumental variable.

in Remark 10, this is the setting in which IVs become essential for constructing informative and counterfactually fair predictors. Just like before, we use a fully connected DAG in the no-confounding CGNF.

For both the Flow IV and the no-confounding CGNFs, predictors are constructed using the latent variables V_Z and V_2. The latent variable V_1 is excluded because it contains information about the unobserved protected attribute A.

Figure 5(b) shows that the Flow IV model substantially outperforms the fair baseline predictor while remaining almost perfectly fair across all levels of confounding. In contrast, the no-confounding CGNF becomes unfair despite using the same subset of latent variables. This demonstrates that simply removing latent variables causally downstream of the protected attribute is insufficient when the latent distribution itself is misspecified. In this setting, Theorem 8 does not apply because the protected attribute is unobserved. Therefore, correctly modelling the latent confounding structure via Flow IV is necessary to obtain a fair representation that remains informative for prediction.

5.2. Law Schools Dataset

We additionally evaluate our approach on the law schools dataset (Wightman, 1998), which has been widely used in the counterfactual fairness literature (Kusner et al., 2017; Kilbertus

et al., 2017). The prediction task is to estimate the first-year average Y from the GPA X_1, LSAT score X_2, and the protected attribute A, which summarizes race and sex. Following prior work, we assume the causal graph $A \to X_1 \to X_2$ and $A \to X_2$.

Because the true underlying SCM is unknown, counterfactual fairness cannot be evaluated directly. Instead, we study the invariance predicted by Theorem 8 in a semi-synthetic setting. To this end, we train CGNFs under different assumptions about latent confounding by varying the correlation between the latent variables V_1 and V_2. Those CGNFs serve as surrogate SCMs that could hypothetically have generated the observed data. We then evaluate the fairness under these surrogate SCMs just like in the synthetic experiments.

Figure 6 shows that the resulting unfairness scores of the no-confounding CGNF remain close to zero across all confounding levels. This empirical stability is consistent with our theoretical result that counterfactual fairness remains invariant across observationally equivalent models under the assumptions of Theorem 8.

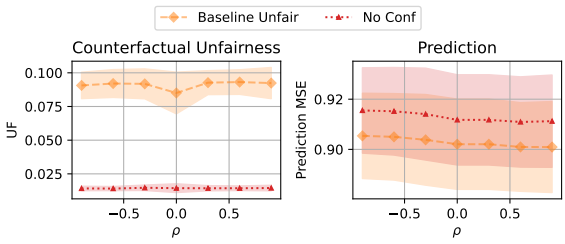


Figure 6: Unfairness under surrogate SCMs and prediction MSE on the law schools dataset.

6. Conclusion

In this paper, we studied counterfactual fairness under hidden confounding, including settings where the protected attribute is unobserved. We showed that, under mild structural assumptions, counterfactual fairness remains identified even under hidden confounding. In particular, all causal models consistent with the observed data and causal ordering remain counterfactually equivalent with respect to interventions on the protected attribute. Consequently, predictors learned under no-confounding assumptions can remain counterfactually fair even when hidden confounding is present. Our synthetic and real-world experiments empirically support this invariance property. We further demonstrated that IVs can be used to construct informative and counterfactually fair representations when the protected attribute itself is unobserved. In this setting, correctly modelling the latent confounding structure becomes necessary to avoid leakage of protected information into the learned representation. A direction for future work is to study the robustness of counterfactual fairness under approximate violations of the assumptions we make in this work and to study the scalability of our results by performing experiments with many variables and high-dimensional data.

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291629.

Sourabh Balgi, Adel Daoud, José M. Peña, Geoffrey Wodtke, and Jesse Zhou. Deep learning with dags. *Sociological Methods & Research*, 2024. doi: 10.1177/00491241251319291. URL https://doi.org/10.1177/00491241251319291.

Marc Braun, Jose Peña, and Adel Daoud. Flow IV: Counterfactual inference in nonseparable outcome models using instrumental variables. In *Fifth Conference on Causal Learning and Reasoning*, 2026. URL https://openreview.net/forum?id=EinusCeXQp.

Samuele Centorrino, Jean-Pierre Florens, and Jean-Michel Loubes. Fairness constraint in structural econometrics and application to fair estimation using instrumental variables. *arXiv preprint arXiv:2202.08977*, 2022.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA, 2019. AAAI Press. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33017801. URL https://doi.org/10.1609/aaai.v33i01.33017801.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Zijian Guo and Dylan S. Small. Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 17(100):1–35, 2016. URL http://jmlr.org/papers/v17/14-379.html.

Jinyong Hahn and Geert Ridder. Instrumental variable estimation of nonlinear models with nonclassical measurement error using control variables. *Journal of Econometrics*, 200(2):238–250, 2017. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2017.06.008. URL https://www.sciencedirect.com/science/article/pii/S0304407617300945. Measurement Error Models.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: a flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1414–1423, Sydney, NSW, Australia, 2017. JMLR.org.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf.

Niki Kilbertus, Philip J. Ball, Matt J. Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 616–626. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/kilbertus20a.html.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4069–4079, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Hongyi Ling, Zhimeng Jiang, Na Zou, and Shuiwang Ji. Counterfactual fairness on graphs: Augmentations, hidden confounders, and identifiability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=hJHf7PCuVt.

Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *CoRR*, abs/1805.05859, 2018. URL http://arxiv.org/abs/1805.05859.

Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 695–703, 2022.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. When causality meets fairness: A survey. *Journal of Logical and Algebraic Methods in Programming*, 141:101000, 2024. ISSN 2352-2208. doi: https://doi.org/10.1016/j.jlamp.2024.101000. URL https://www.sciencedirect.com/science/article/pii/S2352220824000543.

Drago Plečko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends in Machine Learning*, 17(3):304–589, 01 2024. ISSN 1935-8237. doi: 10.1561/2200000106. URL https://doi.org/10.1561/2200000106.

Aahlad Puli and Rajesh Ranganath. General control functions for causal effect estimation from ivs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors,

Advances in Neural Information Processing Systems, volume 33, pages 8440–8451. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/604f2c31e67034642b288d76a8df11d5-Paper.pdf.

James M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8): 2379–2412, 1994. doi: 10.1080/03610929408831393. URL https://doi.org/10.1080/03610929408831393.

Abraham Wald. The Fitting of Straight Lines if Both Variables are Subject to Error. *The Annals of Mathematical Statistics*, 11(3):284 – 300, 1940. doi: 10.1214/aoms/1177731868. URL https://doi.org/10.1214/aoms/1177731868.

Linda F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. Technical report, Law School Admission Council, 1998. URL https://api.semanticscholar.org/CorpusID:151073942.

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1438–1444. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/199. URL https://doi.org/10.24963/ijcai.2019/199.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. Counterfactual fairness with partially known causal graph. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1238–1252. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/08887999616116910fccec17a63584b5-Paper-Conference.pdf.

Appendix A. Technical Proofs

SCM $\mathcal{M}$ Let $\mathcal{M} = (\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U}))$ be an SCM. We are given the observed joint distribution of the variables $X_n, X_{n-1}, \ldots, X_1$, and A. We introduce the shorthand notation G_i for the conditional cumulative distribution function (CDF) $G_i(x_i; x_{i-1} \ldots, x_1, a) \coloneqq \Pr(X_i \leq x_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1, A = a)$ for all $i \in \{2, \ldots, n\}$ and $G_1(x_1; a) \coloneqq \Pr(X_1 \leq x_1 \mid A = a)$. We write H_i for the unobserved conditional CDF $H_i(u_i; x_{i-1}, \ldots, x_1, a) \coloneqq \Pr(U_i \leq u_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1, A = a)$ for all $i \in \{2, \ldots, n\}$ and $H_1(u_1; a) \coloneqq \Pr(U_1 \leq u_1 \mid A = a)$. The CDFs exist because for all $i \in \{1, \ldots, n\}$, U_i is absolutely continuous and $f_i \in \mathcal{F}$ is continuous w.r.t. U_i. The causal graph belonging to $\mathcal{M}$ is depicted in Figure 7.

The following structural equations describe the data generating process.

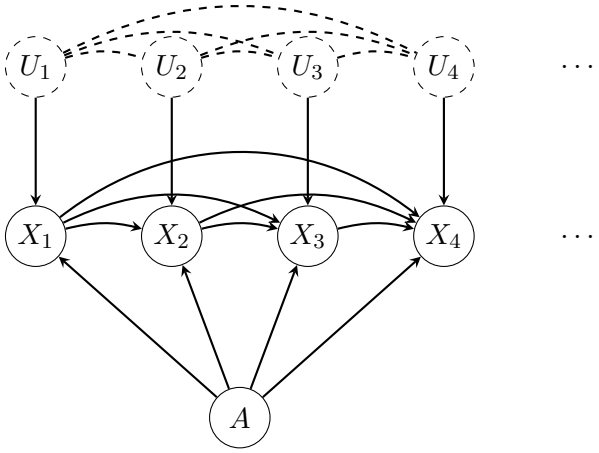


Figure 7: A general causal graph that can be transformed into any DAG where A does not have parents in the graph by removing directed edges. The variable A is a protected attribute and there may be latent confounding between the variables $X_i,\ i \in \{1, \ldots, n\}$.

$$
\begin{aligned}
X_1 &= f_1(U_1, A) \\
X_2 &= f_2(U_2, X_1, A) \\
X_3 &= f_3(U_3, X_2, X_1, A) \\
&\vdots
\end{aligned}
$$

Note that we can assume this structure without loss of generality (w.l.o.g.) because any SCM with acyclic causal graph can be written in this form.

SCM $\mathcal{M}^\star$ Let $\mathcal{M}^\star$ be defined by the SCM $(\widetilde{\mathcal{U}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{F}}, P(\widetilde{\mathcal{U}}))$ with $\widetilde{\mathcal{X}} = \{\tilde{X}_1, \ldots, \tilde{X}_n, A\}$ and structural equations

$$
\begin{aligned}
\tilde{X}_1 &= \tilde{f}_1(\tilde{U}_1, A) \\
\tilde{X}_2 &= \tilde{f}_2(\tilde{U}_2, \tilde{X}_1, A) \\
\tilde{X}_3 &= \tilde{f}_3(\tilde{U}_3, \tilde{X}_2, \tilde{X}_1, A) \\
&\vdots
\end{aligned}
$$

and, similarly to G_i and H_i, we define $\tilde{G}_i$ and $\tilde{H}_i$ as the conditional CDFs of $\mathcal{M}^\star$. Let the two models share the same joint observed distribution, i.e.

$$
\tilde{G}_i(x_i; x_{i-1}, \ldots, x_1, a) = G_i(x_i; x_{i-1}, \ldots, x_1, a) \quad \forall i \in \{1, \ldots, n\}, \tag{1}
$$

which implies that we associate the variable X_i with the variable $\tilde{X}_i$. According to this correspondence, the structural equations of $\mathcal{M}$ and $\mathcal{M}^\star$ induce the same causal ordering of the variables.

Assumption 11 *We assume that* $P(\widetilde{\mathcal{U}}) = P(\tilde{U}_A) \cdot \prod_{i=1}^{n} P(\tilde{U}_i)$*, i.e. the noise terms are unconditionally independent in model* $\mathcal{M}^\star$*.*

This corresponds to the case where there is no hidden confounding.

We assume that Assumptions 6 and 7 are satisfied, i.e. f_i and $\tilde{f}_i$ are strictly monotonic with respect to U_i and $\tilde{U}_i$ respectively for all $i \in \{1, \dots n\}$ and there is no observed or hidden confounder between A and X_i (resp. $\tilde{X}_i$) for all $i \in \{1, \dots, n\}$.

Without loss of generality we furthermore assume that the marginal distributions of U_i and $\tilde{U}_i$ is Unif$(0, 1)$ for all $i \in \{1, \dots, n\}$.

Because the variables $\tilde{U}_1, \dots, \tilde{U}_n$ are independent of each other, this then implies

$$\tilde{H}_i(\tilde{u}_i; \tilde{x}_{i-1}, \dots \tilde{x}_1, a) = \tilde{u}_i \qquad \text{for all } i \in \{2, \dots, n\} \tag{2}$$

$$\tilde{H}_1(\tilde{u}_1; a) = \tilde{u}_1 \tag{3}$$

$$H_1(u_1; a) = u_1. \tag{4}$$

We introduce the following short-hand notation. Let $x' = (x'_n, x'_{n-1}, \dots, x'_1)$ be defined as the (deterministic) counterfactual of an individual with observation $(x_n, x_{n-1}, \dots, x_1, a)$ under intervention $A \leftarrow a'$ in $\mathcal{M}$. This means $x'_i \in \mathbb{R}$ is equal to

$$x'_i = f_i(u_i, x'_{i-1}, \dots, x'_1, a')$$

with $u_i = f_i^{-1}(x_i, x_{i-1}, \dots, x_1, a)$. Let $\tilde{x}' = (\tilde{x}'_n, \tilde{x}'_{n-1}, \dots, \tilde{x}_1)$ be similarly defined as the counterfactual of an individual with the same observations under intervention $A \leftarrow a'$ in $\mathcal{M}^\star$.

Lemma 12 *The noise term* $u'_i = f_i^{-1}(x'_i, x'_{i-1}, \dots, x'_1, a')$ *that generates the counterfactual outcome* x' *is the same as the noise term* $u_i = f_i^{-1}(x_i, x_{i-1}, \dots, x_1, a)$ *that generates the factual outcome* x *for all* $i \in \{1, \dots, n\}$*.*

Proof The lemma follows immediately from the definition of counterfactuals as follows. For $i \in \{1, \dots, n\}$ we derive the counterfactual x'_i.

$$\begin{aligned} u_i &= f_i^{-1}(x_i, x_{i-1}, \dots, x_1, a) && \text{(abduction)} \\ x'_i &= f_i\left(u_i, x'_{i-1}, \dots, x'_1, a'\right) && \text{(prediction)} \\ \implies u'_i &\coloneqq f_i^{-1}(x'_i, x'_{i-1}, \dots, x'_1, a') = u_i \end{aligned}$$

We conclude that $u'_i = u_i$ as claimed by the lemma. ■

Lemma 13 *For any* $i \in \{1, \dots, n\}$ *it holds that*

$$H_i(u_i; x'_{i-1}, \dots, x'_1, a') = H_i(u_i; x_{i-1}, \dots, x_1, a)$$

Proof We use the following properties of our problem setup. Note that for each $i-1$ we have that U_{i-1} is uniquely determined by $X_{i-1}, \ldots, X_1, A$ through the function f_{i-1}^{-1}. Therefore, when we condition U_i on $X_{i-1}, \ldots, X_1, A$, we can add $U_{i-1}, \ldots, U_1$ to the conditioning set as we do not add or remove any information. But according to the causal graph in Figure 7 we have that $U_i \perp\!\!\!\perp \{X_{i-1}, \ldots, X_1, A\} \mid \{U_{i-1}, \ldots, U_1\}$. We can therefore write

$$\begin{aligned} H_i(u_i; x'_{i-1}, \ldots, x'_1, a') &= F_{U_i|U_{i-1}\ldots,U_1}\left(u_i; u'_{i-1}, \ldots, u'_1\right) \\ &= F_{U_i|U_{i-1}\ldots,U_1}\left(u_i; u_{i-1}, \ldots, u_1\right) && \text{(Lemma 12)} \\ &= H_i(u_i; x_{i-1}, \ldots, x_1, a) \end{aligned}$$

where $F_{U_i|U_{i-1},\ldots,U_1}$ is the conditional CDF of U_i conditioned on $U_{i-1}, \ldots, U_1$. ■

Lemma 14 *The functions f_i and $\tilde{f}_i$ satisfy the following equalities*

$$f_i(u_i, x_{i-1}, \ldots, x_1, a) = G_i^{-1}\left(H_i(u_i; x_{i-1}, \ldots, x_1, a); x_{i-1}, \ldots, x_1, a\right) \quad \text{(i)}$$

$$f_i^{-1}(x_i, x_{i-1}, \ldots, x_1, a) = H_i^{-1}\left(G_i(x_i; x_{i-1}, \ldots, x_1, a); x_{i-1}, \ldots, x_1, a\right) \quad \text{(ii)}$$

$$\tilde{f}_i(\tilde{u}_i, \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a) = G_i^{-1}(\tilde{u}_i; \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a) \quad \text{(iii)}$$

Proof We first show Equation ii. Equation i then follows immediately from Equation ii. We have that

$$\begin{aligned} &\Pr(X_i \leq x_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1, A = a) \\ &\quad = \Pr(f_i(U_i, x_{i-1}, \ldots, x_1, a) \leq x_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1, A = a) \\ &\quad = \Pr(U_i \leq f_i^{-1}(x_i, x_{i-1}, \ldots, x_1, a) \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1, A = a) \\ &\quad = H_i\left(f_i^{-1}(x_i, x_{i-1}, \ldots, x_1, a); x_{i-1}, \ldots, x_1, a\right) \end{aligned}$$

and at the same time

$$\Pr(X_i \leq x_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1, A = a) = G_i(x_i; x_{i-1}, \ldots, x_1, a)$$

which together implies that

$$f_i^{-1}(x_i, x_{i-1}, \ldots, x_1, a) = H_i^{-1}\big(G_i(x_i; x_{i-1}, \ldots, x_1, a); x_{i-1}, \ldots, x_1, a\big).$$

We can do the exact same derivations to show that

$$\tilde{f}_i^{-1}(\tilde{x}_i, \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a) = \tilde{H}_i^{-1}\big(\tilde{G}_i(\tilde{x}_i; \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a); \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a\big)$$

which, using Equation 1 and Equation 2, implies that

$$\tilde{f}_i^{-1}(\tilde{x}_i, \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a) = G_i(\tilde{x}_i; \tilde{x}_{i-1}, \ldots, \tilde{x}_1, a).$$

This immediately implies Equation iii. ■

Proposition 15 *Given an observation* $(x_n, \ldots, x_1, a)$*, the counterfactual outcome* x' *under intervention* $A = a'$ *in* $\mathcal{M}$ *is equal to the counterfactual outcome* $\tilde{x}'$ *under intervention* $A = a'$ *in* $\mathcal{M}^\star$*, i.e.* $\mathcal{M}$ *and* $\mathcal{M}^\star$ *are counterfactually equivalent with respect to interventions on* A*.*

$$x' = \tilde{x}'$$

Proof We start by deriving the counterfactual outcome under $\mathcal{M}$

$$\begin{aligned}
u_i &= f_i^{-1}(x_i, x_{i-1}, \ldots, x_1, a) && \text{(abduction)} && (6)\\
&= H_i^{-1}\left(G_i(x_i; x_{i-1}, \ldots, x_1, a); x_{i-1}, \ldots, x_1, a\right) && \text{(Lemma 14 ii)} && (7)\\
x_i' &= f_i(u_i, x_{i-1}', \ldots, x_1', a') && \text{(action)} && (8)\\
&= G_i^{-1}\left(H_i(u_i; x_{i-1}', \ldots, x_1', a'); x_{i-1}', \ldots, x_1', a'\right) && \text{(Lemma 14 i)} && (9)\\
&= G_i^{-1}\left(H_i(u_i; x_{i-1}, \ldots, x_1, a); x_{i-1}', \ldots, x_1', a'\right) && \text{(Lemma 13)} && (10)\\
&= G_i^{-1}\left(G_i(x_i; x_{i-1}, \ldots, x_1, a); x_{i-1}', \ldots, x_1', a'\right) && && (11)
\end{aligned}$$

where in the last equality we have plugged-in the expression from Equation 7 for u_i.

Next, we derive the counterfactual outcome under $\mathcal{M}^\star$

$$\begin{aligned}
\tilde{u}_i &= \tilde{f}_i^{-1}(x_i, x_{i-1}, \ldots, x_1, a) && \text{(abduction)} && (12)\\
&= G_i(x_i; x_{i-1}, \ldots, x_1, a) && \text{(Lemma 14 iii)} && (13)\\
\tilde{x}_i' &= \tilde{f}_i(\tilde{u}_i, x_{i-1}', \ldots, x_1', a') && \text{(action)} && (14)\\
&= G_i^{-1}(\tilde{u}_i; x_{i-1}', \ldots, x_1', a') && \text{(Lemma 14 iii)} && (15)\\
&= G_i^{-1}(G_i(x_i; x_{i-1}, \ldots, x_1, a); x_{i-1}', \ldots, x_1', a') && && (16)
\end{aligned}$$

Comparing the expressions in Equation 11 and Equation 16, it can be seen that $x_i' = \tilde{x}_i'$ for all $i \in \{1, \ldots, n\}$ as claimed by Proposition 15. ■

It now only remains to prove Theorem 8.

Proof Let the SCM $\widetilde{\mathcal{M}}$ now be defined as in Section 4. It follows immediately from Proposition 15 that $\mathcal{M}^\star$ is both counterfactually equivalent to $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ with respect to interventions on A. Therefore, due to the transitivity of the equivalence relation, $\mathcal{M}$ is counterfactually equivalent to $\widetilde{\mathcal{M}}$ with respect to interventions on A as claimed by Theorem 8. ■

Appendix B. Experimental Details

The replication code and data for all our experiments can be found on GitHub (https://github.com/Marbr987/counterfactual-fairness-with-ivs).

B.1. Data Generating Processes Used in the Synthetic Experiments

DGP of synthetic experiment 1: Observed A The data for the first part of this experiment, where the structural equations are invertible with respect to U, is generated via the following DGP

$$A \sim \mathcal{N}(0,1), \qquad X_1 = A + U_1, \qquad X_2 = \sin(X_1) + A + U_2, \qquad Y = X_1 + X_2^2,$$

where the exogenous variables (U_1, U_2) follow a bivariate Gaussian distribution

$$(U_1, U_2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

The data for the second part of the experiment is generated via the following DGP

$$A \sim \mathcal{N}(0,1), \quad X_1 = \sin(2\,(A + U_1)), \quad X_2 = \sin(2\,(X_1 + A + U_2)), \quad Y = X_1 + X_2^2,$$

where the exogenous variables (U_1, U_2) follow a bivariate Gaussian distribution

$$(U_1, U_2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

DGP of synthetic experiment 2: Observed A with IV We generate data from the following DGP

$$\begin{aligned} &A \sim \mathcal{N}(0,1), && Z \sim \mathcal{N}(0,1) && X_1 = (Z + A + U_1)/\sqrt{3},\\ &X_2 = \sin(\frac{1}{3}\,X_1 + \frac{\pi}{2}) \cdot \frac{A + U_2}{\sqrt{2}}, && Y = X_1 + X_2^2, \end{aligned}$$

where the exogenous variables (U_1, U_2) follow a bivariate Gaussian distribution

$$(U_1, U_2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

DGP of synthetic experiment 3: Unobserved A with IV In the last experiment, we generate data from the following DGP

$$\begin{aligned} &A \sim \mathcal{N}(0,1), && Z \sim \mathcal{N}(0,1) && X_1 = \sqrt{0.2}\,Z + \sqrt{0.8}\,(\sqrt{07}\,A + \sqrt{0.3}\,U_1),\\ &X_2 = \sqrt{0.2}\,X_1 + \sqrt{0.8}\,U_2, && Y = X_1 + X_2^2, \end{aligned}$$

where the exogenous variables (U_1, U_2) follow a bivariate Gaussian distribution

$$(U_1, U_2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$