LEARNING COUNTERFACTUALLY INVARIANT PREDICTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a method to learn predictors that are invariant under counterfactual changes of certain covariates. This method is useful when the prediction target is causally influenced by covariates that should not affect the predictor output. For instance, this could prevent an object recognition model from being influenced by position, orientation, or scale of the object itself. We propose a model-agnostic regularization term based on conditional kernel mean embeddings to enforce *counterfactual invariance* during training. We prove the soundness of our method, which can handle mixed categorical and continuous multivariate attributes. Empirical results on synthetic and real-world data demonstrate the efficacy of our method in a variety of settings.

1 INTRODUCTION AND RELATED WORK

Invariance, or equivariance to certain transformations of data, has proven essential in numerous applications of machine learning (ML), since it can lead to better generalization capabilities Arjovsky et al. (2019); Chen et al. (2020); Bloem-Reddy & Teh (2020). For instance, in image recognition, predictions ought to remain unchanged under scaling, translation, or rotation of the input image. Data augmentation is one of the earliest heuristics developed to promote this kind of invariance, that has become indispensable for training successful models like deep neural networks (DNNs) Shorten & Khoshgoftaar (2019); Xie et al. (2020). Well-known examples of certain types of "invariance by design" include convolutional neural networks (CNNs) for translation invariance Krizhevsky et al. (2012), group equivariant CNNs for other group transformations Cohen & Welling (2016), recurrent neural networks (RNNs) and transformers for sequential data Vaswani et al. (2017), DeepSet Zaheer et al. (2017) for sets, and graph neural networks (GNNs) for different types of geometric structures Battaglia et al. (2018).

Many real-world applications in modern ML, however, call for an arguably stronger notion of invariance based on causality, called *counterfactual invariance*. This case has been made for image classification, algorithmic fairness Hardt et al. (2016); Mitchell et al. (2021), robustness Bühlmann (2020), and out-of-distribution generalization Lu et al. (2021). These applications require predictors to exhibit invariance with respect to hypothetical manipulations of the data generating process (DGP) Peters et al. (2016); Heinze-Deml et al. (2018); Rojas-Carulla et al. (2018); Arjovsky et al. (2019); Bühlmann (2020). In image classification, for instance, we want a model that "would have made the same prediction, if the object position had been different with everything else being equal". Similarly, in algorithmic fairness Kilbertus et al. (2017); Kusner et al. (2017) introduce notions of interventional and counterfactual fairness, based on certain invariances in the DGP of the causal relationships between observed variables.

Counterfactual invariance has the significant advantage that it incorporates structural knowledge of the DGP. However, enforcing this notion in practice is very challenging, since it is untestable in real-world observational settings, unless strong prior knowledge of the DGP is available. Inspired by problems in natural language processing (NLP), Veitch et al. (2021) provide a method to achieve counterfactual invariance based on distribution matching via the maximum mean discrepancy (MMD). This method enforces a *necessary, but not sufficient* condition of counterfactual invariance during training. Consequently, it is unclear whether this method achieves actual invariance in practice, or just an arguably weaker proxy. Furthermore, the work by Veitch et al. (2021) only considers discrete random variables when enforcing counterfactual invariance, and it only applies to

specific, selected causal graphs. To overcome the aforementioned problems, we propose a general definition of counterfactual invariance and a novel method to enforce it. Our main contributions can be summarized as follows:

- Based on a structural causal model (SCM), we provide a new definition of counterfactual invariance (cf. Definition 2.2) that is more general than that of Veitch et al. (2021).
- We establish a connection between counterfactual invariance and conditional independence that is provably *sufficient* for counterfactual invariance (cf. Theorem 3.2).
- We propose a new objective function that is composed of the loss function and on the flexible Hilbert-Schmidt Conditional Independence Criterion (HSCIC) Park & Muandet (2020), to enforce counterfactual invariance in practice. Our method works well for both categorical and continuous covariates and outcomes, as well as in multivariate settings.

2 Preliminaries and Background

Counterfactual invariance. We introduce structural causal models as in Pearl (2000).

Definition 2.1 (Structural causal model (SCM))**.** *A structural causal model is a tuple* $(\mathbf{U}, \mathbf{V}, F, \mathbb{P}_{\mathbf{U}})$ *such that* $\mathbf{U}$ *is a set of background variables that are exogenous to the model;* $\mathbf{V}$ *is a set of observable (endogenous) variables;* $F = \{f_V\}_{V \in \mathbf{V}}$ *is a set of functions from (the domains of)* $\mathsf{pa}(V) \cup U_V$ *to (the domain of)* V*, where* $U_V \subset \mathbf{U}$ *and* $\mathsf{pa}(V) \subseteq \mathbf{V} \setminus \{V\}$ *such that* $V = f_V(\mathsf{pa}(V), U_V)$*; (iv)* $\mathbb{P}_{\mathbf{U}}$ *is a probability distribution over the domain of* $\mathbf{U}$*. Further, the subsets* $\mathsf{pa}(V) \subseteq \mathbf{V} \setminus \{V\}$ *are chosen such that the graph* $\mathcal{G}$ *over* $\mathbf{V}$ *where the edge* $V' \to V$ *is in* $\mathcal{G}$ *if and only if* $V' \in \mathsf{pa}(V)$ *is a directed acyclic graph (DAG).*

We always denote with $\mathbf{Y} \subset \mathbf{V}$ the outcome (or prediction target), and with $\hat{\mathbf{Y}}$ a predictor for that target. The predictor $\hat{\mathbf{Y}}$ is not strictly part of the SCM, because we get to tune $f_{\hat{\mathbf{Y}}}$. Since it takes inputs from $\mathbf{V}$, we often treat it as an observed variable in the SCM. As such, it also "derives its randomness from the exogenous variables", i.e., is defined on the same probability space. Each SCM implies a unique *observational distribution* over $\mathbf{V}$ (Pearl, 2000), but it also entails interventional distributions. Given a variable $A \in \mathbf{V}$, an intervention $A \leftarrow a$ amounts to replacing f_A in F with the constant function $A = a$. This yields a new SCM, which induces the *interventional distribution* under intervention $A \leftarrow a$. Similarly, we can intervene on multiple variables $\mathbf{V} \supseteq \mathbf{A} \leftarrow \mathbf{a}$. We then write $\mathbf{Y}^*_{\mathbf{a}}$ for the outcome in the intervened SCM, also called *potential outcome*. Note that the interventional distribution $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y})$ differs in general from the conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{A}}(\mathbf{y} \mid \mathbf{a})$, due to possible unobserved confounding effects.[1] We can also condition on a set of variables $\mathbf{W} \subseteq \mathbf{V}$ in the (observational distribution of the) original SCM before performing an intervention, which we denote by $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w})$. This is a *counterfactual distribution*: "Given that we have observed $\mathbf{W} = \mathbf{w}$, what would $\mathbf{Y}$ have been had we set $\mathbf{A} \leftarrow \mathbf{a}$, instead of the value $\mathbf{A}$ had actually taken?" Note that the sets $\mathbf{A}$ and $\mathbf{W}$ need not be disjoint. We can now define counterfactual invariance.

Definition 2.2 (Counterfactual invariance)**.** *Let* $\mathbf{A}$*,* $\mathbf{W}$ *be (not necessarily disjoint) sets of nodes in a given SCM. A predictor* $\hat{\mathbf{Y}}$ *is* counterfactually invariant *in* $\mathbf{A}$ *with respect to* $\mathbf{W}$*, if* $\mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w}) = \mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}|\mathbf{W}}(\mathbf{y} \mid \mathbf{w})$ *almost surely, for all* $\mathbf{a}, \mathbf{a}'$ *in the domain of* $\mathbf{A}$ *and all* $\mathbf{w}$ *in the domain of* $\mathbf{W}$*.*[2]

A counterfactually invariant predictor can be viewed as robust to changes of $\mathbf{A}$, in the sense that the (conditional) post-interventional distribution of $\hat{\mathbf{Y}}$ does not change for different values of the intervention. Our Definition 2.2 is more general than previously considered notions of counterfactual invariance. For instance, the invariance in Definition 1.1 by Veitch et al. (2021) requires $\hat{\mathbf{Y}}^*_{\mathbf{a}} = \hat{\mathbf{Y}}^*_{\mathbf{a}'}$ almost surely for all $\mathbf{a}, \mathbf{a}'$ in the domain of $\mathbf{A}$. First, it does not allow to condition on observed evidence, i.e., it cannot consider "true counterfactuals" and is thus unable to promote—for example—counterfactual fairness, see Definition 3.6 (Kusner et al., 2017). Second, it may appear

[1] We use $\mathbb{P}$ for distributions as is common in the kernel literature (Muandet et al., 2021) and the potential outcome notation $\mathbf{Y}^*_{\mathbf{a}}$ instead of $\mathbf{Y} \mid do(\mathbf{a})$ for conciseness when mixing conditioning with interventions.

[2] With a mild abuse of notation, if $\mathbf{W} = \emptyset$ then the requirement of conditional counterfactual invariance becomes $\mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}}}(\mathbf{y}) = \mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}}(\mathbf{y})$ almost surely, for all $\mathbf{a}, \mathbf{a}'$ in the domain of $\mathbf{A}$.

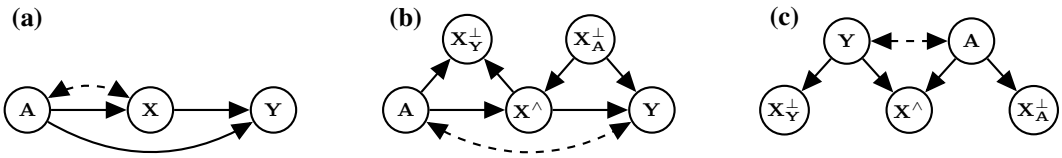


Figure 1: **(a) An example for Theorem 3.2**, where $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y})$ is unidentifiable, due to the confounding effect denoted by the dashed double arrow. However, by Theorem 3.2 any predictor $\hat{\mathbf{Y}}$ such that $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$ with $\mathbf{Z} = \{\mathbf{X}\}$ is counterfactually invariant in $\mathbf{A}$. Hence, the criterion in Theorem 3.2 is weaker than identifiability. **(b)-(c) Causal and anti-causal structure** as in Veitch et al. (2021). The variable $\mathbf{X}$ is decomposed in three parts. $\mathbf{X}_{\mathbf{A}}^{\perp}$ the part of $\mathbf{X}$ that is not causally influenced by $\mathbf{A}$, $\mathbf{X}_{\mathbf{Y}}^{\perp}$ is the part that does not causally influence $\mathbf{Y}$, and $\mathbf{X}^{\wedge}$ is the remaining part, that is both influenced by $\mathbf{A}$ and that influences $\mathbf{Y}$.

stronger in that it asks for equality of random variables instead of equality of distributions. However, (a) contrary to their definition, Veitch et al. (2021) only enforce equality of distributions in practice (via MMD), and (b) since $\hat{\mathbf{Y}}^*_{\mathbf{a}}$, $\hat{\mathbf{Y}}^*_{\mathbf{a}'}$ are (deterministic) functions of the *same* exogenous (unobserved) random variables, distributional equality is a natural choice for counterfactual invariance.

Kernel mean embeddings and conditional measures. Our new objective function heavily relies on kernel mean embeddings (KMEs). We now highlight the main concepts pertaining KMEs and refer the reader to Smola et al. (2007); Schölkopf et al. (2002); Berlinet & Thomas-Agnan (2011); Muandet et al. (2017) for more details. Fix a measurable space $\mathscr{Y}$ with respect to a σ-algebra $\mathcal{F}_{\mathscr{Y}}$, and consider a probability measure $\mathbb{P}$ on the space $(\mathscr{Y}, \mathcal{F}_{\mathscr{Y}})$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) with a bounded kernel $k_{\mathbf{Y}} \colon \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$, i.e., $k_{\mathbf{Y}}$ is such that $\sup_{\mathbf{y} \in \mathscr{Y}} k(\mathbf{y}, \mathbf{y}) < +\infty$. The kernel mean embedding $\mu_{\mathbb{P}}$ of $\mathbb{P}$ is defined as the expected value of the function $k(\,\cdot\,, \mathbf{y})$ with respect to $\mathbf{y}$, i.e., $\mu_{\mathbb{P}} \coloneqq \mathbb{E}\left[k(\,\cdot\,, \mathbf{y})\right]$. The definition of KMEs can be extended to conditional distributions (Fukumizu et al., 2013; Grünewälder et al., 2012; Song et al., 2009; 2013). Consider two random variables $\mathbf{Y}$, $\mathbf{Z}$, and denote with $(\Omega_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}})$ and $(\Omega_{\mathbf{Z}}, \mathcal{F}_{\mathbf{Z}})$ the respective measurable spaces. These random variables induce a probability measure $\mathbb{P}_{\mathbf{Y},\mathbf{Z}}$ in the product space $\Omega_{\mathbf{Y}} \times \Omega_{\mathbf{Z}}$. Let $\mathcal{H}_{\mathbf{Y}}$ be a RKHS with a bounded kernel $k_{\mathbf{Y}}(\cdot, \cdot)$ on $\Omega_{\mathbf{Y}}$. We define the KME of a conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{Z}}(\cdot \mid \mathbf{z})$ via $\mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}} \coloneqq \mathbb{E}\left[k_{\mathbf{Y}}(\,\cdot\,, \mathbf{y}) \mid \mathbf{Z} = \mathbf{z}\right]$. Here, the expected value is taken over $\mathbf{y}$. KMEs of conditional measures can be estimated from samples. To illustrate this, consider i.i.d. samples $(\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n)$. Denote with $\hat{K}_{\mathbf{Y}}$ the kernel matrix with entries $[\hat{K}_{\mathbf{Y}}]_{i,j} \coloneqq k_{\mathbf{Y}}(\mathbf{y}_i, \mathbf{y}_j)$. Furthermore, let $k_{\mathbf{Z}}$ be a bounded kernel on $\Omega_{\mathbf{Z}}$. Then, $\mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}}$ can be estimated as

$$\hat{\mu}_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}} \coloneqq \sum_{i=1}^{n} \hat{w}^{(i)}_{\mathbf{Y}|\mathbf{Z}}(\mathbf{z}) k_{\mathbf{Y}}(\cdot, \mathbf{y}_i)\,, \quad \hat{w}_{\mathbf{Y}|\mathbf{Z}}(\cdot) \coloneqq (\hat{K}_{\mathbf{Z}} - n\lambda I)^{-1} \left[k_{\mathbf{Z}}(\cdot, \mathbf{z}_1), \cdots, k_{\mathbf{Z}}(\cdot, \mathbf{z}_n)\right]^T \quad (1)$$

where, I is the identity matrix and λ is a regularization parameter. Here, $\hat{w}^{(i)}_{\mathbf{X}|\mathbf{A}}(\cdot)$, the i-th entry of $\hat{w}_{\mathbf{X}|\mathbf{A}}(\cdot)$, are the coefficients of kernel ridge regression (Grünewälder et al., 2012).

3 Counterfactually Invariant Predictors

We now establish a simple graphical criterion to express counterfactual invariance as a conditional independence in the observational distribution, rendering it estimable from i.i.d. data. We first repeat the notion of *blocked paths* (Pearl, 2000).

Definition 3.1. *Consider a path π of causal graph $\mathcal{G}$. A set of nodes $\mathbf{Z}$ blocks π, if π contains a triple of consecutive nodes connected in one of the following ways: $N_i \to Z \to N_j$, $N_i \leftarrow Z \to N_j$, or $N_i \to M \leftarrow N_j$, with $N_i, N_j \notin \mathbf{Z}$, $Z \in \mathbf{Z}$, and neither M nor any descendent of M is in $\mathbf{Z}$.*

Theorem 3.2. *Let $\mathcal{G}$ be a causal diagram, and let $\mathbf{A}$, $\mathbf{W}$ be two (not necessarily disjoint) sets of nodes in $\mathcal{G}$. Let $\mathbf{Z}$ be a set of nodes that blocks all non-directed paths from $\mathbf{A} \cup \mathbf{W}$ to $\mathbf{Y}$. Then, for any SCM compatible with $\mathcal{G}$, any predictor $\hat{\mathbf{Y}}$ that satisfies $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$ is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W}$.*

The proof is deferred to Appendix A. Our key observation is that the set $\mathbf{Z}$ as in Theorem 3.2 acts as a valid adjustment set for certain random variables in the joint pre- and post-interventional counterfactual graph as used in Shpitser & Pearl (2008). We can then use similar techniques to covariate

adjustment. However, our proof does *not* rely on identification of the counterfactual distributions (e.g., by simply applying the do-calculus (Pearl, 2000)). Crucially, *the existence of a measurable set* $\mathbf{Z}$ *as in Theorem 3.2 does not imply the identifiability of counterfactual distributions* $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y})$ (see Figure 1(a) for a counterexample). In particular, the assumptions do not rule out hidden confounding in the model. Hence, our method is applicable even when (certain parameters of) $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}}(\mathbf{y})$ or $\mathbb{P}_{\mathbf{Y}^*_{\mathbf{a}}|\mathbf{W}}(\mathbf{y})$ cannot be learned from observational data.

Counterfactually invariant predictors. We propose to use the $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z})$, to promote counterfactual invariance by encouraging the sufficient conditional independence $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$. Given a loss function $\mathcal{L}(\hat{\mathbf{Y}})$, we propose to minimize the following loss

$$\mathcal{L}_{\text{TOTAL}}(\hat{\mathbf{Y}}) = \mathcal{L}(\hat{\mathbf{Y}}) + \gamma \cdot \text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z}) \,, \tag{2}$$

with a parameter $\gamma \geq 0$, regulating the trade-off between accuracy and counterfactual invariance. There is no principled way to choose the "right" trade-off parameter. However, in practice we can use a standard approach, which consists of two steps: (i) learn a collection of models on the Pareto frontier using different values of γ; (ii) from a collection of models choose the best one following, i.e., Muandet et al. (2021). Next, we develop and justify using the HSCIC, (see Corollary 3.5) for which $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z}) = 0$ if and only if $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$.

Hilbert-Schmidt Conditional Independence Criterion (HSCIC). Consider two random variables $\mathbf{Y}$ and $\mathbf{A}$, and denote with $(\Omega_{\mathbf{Y}}, \mathcal{F}_{\mathbf{Y}})$ and $(\Omega_{\mathbf{A}}, \mathcal{F}_{\mathbf{A}})$ the respective measurable spaces. Suppose that we are given two RKHSs $\mathcal{H}_{\mathbf{Y}}$, $\mathcal{H}_{\mathbf{A}}$ over the support of $\mathbf{Y}$ and $\mathbf{A}$ respectively. The tensor product space $\mathcal{H}_{\mathbf{Y}} \otimes \mathcal{H}_{\mathbf{A}}$ is defined as the space of functions of the form $(f \otimes g)(\mathbf{y}, \mathbf{a}) := f(\mathbf{y})g(\mathbf{a})$, for all $f \in \mathcal{H}_{\mathbf{Y}}$ and $g \in \mathcal{H}_{\mathbf{A}}$. The tensor product space yields a natural RKHS structure, with kernel k defined by $k(\mathbf{y} \otimes \mathbf{a}, \mathbf{y}' \otimes \mathbf{a}') := k_{\mathbf{Y}}(\mathbf{y}, \mathbf{y}')k_{\mathbf{A}}(\mathbf{a}, \mathbf{a}')$. We refer the reader, i.e., to Szabó & Sriperumbudur (2017) for more details on tensor product spaces. With this notation we define:

Definition 3.3 (HSCIC, following Definition 5.3 by Park & Muandet (2020))**.** *For (sets of) random variables* $\mathbf{Y}$*,* $\mathbf{A}$*,* $\mathbf{Z}$ *the* HSCIC *between* $\mathbf{Y}$ *and* $\mathbf{A}$ *given* $\mathbf{Z}$ *is defined as the real-valued random variable* $\text{HSCIC}(\mathbf{Y}, \mathbf{A} \mid \mathbf{Z}) = H_{\mathbf{Y},\mathbf{A}|\mathbf{Z}} \circ \mathbf{Z}$*. Here,* $H_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}$ *is a real-valued deterministic function, defined as* $H_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\mathbf{z}) := \left\|\mu_{\mathbf{Y},\mathbf{A}|\mathbf{Z}=\mathbf{z}} - \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}} \otimes \mu_{\mathbf{A}|\mathbf{Z}=\mathbf{z}}\right\|$*, with* $\|\cdot\|$ *the norm induced by the inner product of the tensor product space* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}}$*.*

Our Definition 3.3 differs slightly from Park & Muandet (2020), who define HSCIC using the Bochner conditional expected value. While it is functionally equivalent (with the same implementation, see eq. (3)), it has the benefit of bypassing some technical assumptions required by Park & Muandet (2020). We refer readers to Appendix C and D for a comparison with previous approaches. The HSCIC has the following important property.

Theorem 3.4 (following Theorem 5.4 by Park & Muandet (2020))**.** *If the kernel* k *of* $\mathcal{H}_{\mathbf{X}} \otimes \mathcal{H}_{\mathbf{A}}$ *is characteristic*[3]*,* $\text{HSCIC}(\mathbf{Y}, \mathbf{A} \mid \mathbf{Z}) = 0$ *almost surely if and only if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{A} \mid \mathbf{Z}$*.*

A proof is in Appendix B. We remark that "most interesting" kernels such as the Gaussian and Laplacian kernel are characteristic. Furthermore, if kernels are translation-invariant and characteristic, then their tensor product is also a characteristic kernel (Szabó & Sriperumbudur, 2017). Hence, this natural assumption is non-restrictive in practice. Combining Theorems 3.2 and 3.4, we can now use HSCIC to promote counterfactual invariance.

Corollary 3.5. *Consider an SCM with causal diagram* $\mathcal{G}$ *and fix two (not necessarily disjoint) sets of nodes* $\mathbf{A}$*,* $\mathbf{W}$*. Let* $\mathbf{Z}$ *be a set of nodes in* $\mathcal{G}$ *that blocks all non-directed paths from* $\mathbf{A} \cup \mathbf{W}$ *to* $\mathbf{Y}$*. Then, any predictor* $\hat{\mathbf{Y}}$ *that satisfies* $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z}) = 0$ *almost surely is counterfactually invariant in* $\mathbf{A}$ *with respect to* $\mathbf{W}$*.*

We do *not* require $\mathbf{A}$, $\mathbf{W}$, $\mathbf{Z}$ or $\mathbf{Y}$ to be binary or categorical, a major improvement over existing methods (Chiappa, 2019; Xu et al., 2020), which cannot handle continuous multi-variate attributes.

[3]The tensor product kernel k is characteristic if the mapping $\mathbb{P}_{\mathbf{Y},\mathbf{A}} \mapsto \mathbb{E}_{\mathbf{y},\mathbf{a}}\left[k(\,\cdot\,, \mathbf{y} \otimes \mathbf{a})\right]$ is injective.

Estimating the HSCIC from samples. Given n samples $\{(\mathbf{y}_i, \mathbf{a}_i, \mathbf{z}_i)\}_{i=1}^n$, denote by $\hat{K}_\mathbf{Y}$ and $\hat{K}_\mathbf{A}$ the corresponding kernel matrices for $\mathbf{Y}$ and $\mathbf{A}$ (see eq. (1)). We then estimate the $H_{\mathbf{Y},\mathbf{A}|\mathbf{X}}$ as

$$\hat{H}^2_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot) = \hat{w}^T_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot)\left(\hat{K}_\mathbf{Y} \odot \hat{K}_\mathbf{A}\right)\hat{w}_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot) - 2\left(\hat{w}^T_{\mathbf{Y}|\mathbf{Z}}(\cdot)\hat{K}_\mathbf{Y}\hat{w}_{\mathbf{Y},\mathbf{A}|\mathbf{X}}(\cdot)\right) \cdot \left(\hat{w}^T_{\mathbf{A}|\mathbf{X}}(\cdot)\hat{K}_\mathbf{A}\hat{w}_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot)\right) + \left(\hat{w}^T_{\mathbf{Y}|\mathbf{Z}}(\cdot)\hat{K}_\mathbf{Y}\hat{w}_{\mathbf{Y}|\mathbf{Z}}(\cdot)\right)\left(\hat{w}^T_{\mathbf{A}|\mathbf{Z}}(\cdot)\hat{K}_\mathbf{A}\hat{w}_{\mathbf{A}|\mathbf{Z}}(\cdot)\right) , \tag{3}$$

where $\odot$ is element-wise multiplication. The vectors $\hat{w}_{\mathbf{Y}|\mathbf{Z}}(\cdot)$, $\hat{w}_{\mathbf{A}|\mathbf{Z}}(\cdot)$, and $\hat{w}_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot)$ are found via kernel ridge regression. Caponnetto & Vito (2007) the convergence (and rates) of $\hat{H}^2_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot)$ to $H^2_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot)$ under mild conditions. In practice, computing the HSCIC approximation $\hat{H}^2_{\mathbf{Y},\mathbf{A}|\mathbf{Z}}(\cdot)$ may be time-consuming. To speed it up, we can use random Fourier features to approximate the matrices $\hat{K}_\mathbf{Y}$ and $\hat{K}_\mathbf{A}$ (Rahimi & Recht, 2007; Avron et al., 2017). We emphasize that eq. (3) allows us to consistently estimate the HSCIC *from observational i.i.d. samples, without prior knowledge of the counterfactual distributions*.

Measuring counterfactual invariance. Besides predictive performance, e.g., mean squared error (**MSE**) for regression, our key metric of interest is the level of counterfactual invariance achieved by the predictor $\hat{\mathbf{Y}}$, i.e., a measure of how the distribution of $\hat{\mathbf{Y}}^*_\mathbf{a}$ changes for different values of $\mathbf{a}$ and across all conditioning values $\mathbf{w}$. We quantify the overall counterfactual variance as a single scalar, the **V**ariance of **C**ounter**F**actuals (**VCF**):

$$\mathbf{VCF}(\hat{\mathbf{Y}}) = \mathbb{E}_{\mathbf{w}\sim\mathbb{P}_\mathbf{W}}\left[\mathrm{var}_{\mathbf{a}'\sim\mathbb{P}_\mathbf{A}}[\mathbb{E}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}|\mathbf{W}=\mathbf{w}}(\hat{\mathbf{y}} \mid \mathbf{w})]\right] . \tag{4}$$

That is, we look at how the average outcome varies with the interventional value $\mathbf{a}$ at conditioning value $\mathbf{w}$ and average this variance over $\mathbf{w}$. For deterministic predictors (i.e., point estimators), which we use in all our experiments, the variance term in eq. (4) is zero if and only if counterfactual invariance holds at $\mathbf{w}$ for (almost) all $\mathbf{a}'$ in the support of $\mathbb{P}_\mathbf{A}$ (i.e., the prediction remains constant). Since the variance is non-negative, **VCF** (we often drop the argument) is then zero if and only if counterfactual invariance holds almost surely. To estimate **VCF** in practice, we pick d datapoints $(\mathbf{w}_i)_{i=1}^d$ from the observed data, and for each compute the counterfactual outcomes $\hat{\mathbf{Y}}^*_{\mathbf{a}'} \mid \mathbf{w}_i$ for k different values of $\mathbf{a}'$. The inner expectation is simply the predictor output. We use empirical variances with k examples for each of the d chosen datapoints, and the empirical mean of the d variances for the outer expectation. Crucially, **VCF** requires access to ground-truth counterfactual distributions, which by their very nature are unavailable in practice (neither for training nor at test time). Hence, we can only assess **VCF**, as a direct measure of counterfactual invariance, in synthetic scenarios. We demonstrate in our experiments, that HSCIC (estimable from the observed data) empirically serves as a proxy for **VCF**.

Applications of counterfactual invariance. We briefly outline potential applications of counterfactual invariance, which we will subsequently study empirically on real-world data.

Image classification. Counterfactual invariance serves as a strong notion of robustness in high-dimensional settings, such as image classification: "Would the truck have been classified correctly had it been winter in this exact situation instead of summer?" For concrete demonstration, we will use the dSprites dataset (Matthey et al., 2017), which consists of relatively simple, yet high-dimensional, square black and white images of different shapes (squares, ellipses, etc.), sizes, and orientations (rotation) in different xy-positions.

Counterfactual fairness. The popular definition of counterfactual fairness (Kusner et al., 2017) is an incarnation of our counterfactual invariance (see Definition 2.2). It is captured informally by the following question after receiving a consequential decision: "Would I have gotten the same outcome had I been a different gender, race, or age with all else being equal?". Again, we denote the outcome by $\mathbf{Y} \subset \mathbf{V}$ and so-called *protected attributes* by $\mathbf{A} \subseteq \mathbf{V} \setminus \mathbf{Y}$ such as gender, race, or age protected under anti-discrimination laws (see, e.g., Barocas & Selbst (2016)). Collecting all remaining observed covariates into $\mathbf{W} := \mathbf{V} \setminus \mathbf{Y}$, the definition of counterfactual fairness clearly becomes a special case of our counterfactual invariance.

Definition 3.6 (Counterfactual Fairness, Definition 5 by Kusner et al. (2017))**.** *A predictor* $\hat{\mathbf{Y}}$ *is* counterfactually fair *if under any context* $\mathbf{W} = \mathbf{w}$ *and* $\mathbf{A} = \mathbf{a}$*, it holds* $\mathbb{P}_{\hat{\mathbf{Y}}^*_\mathbf{a}|\mathbf{W},\mathbf{A}}(\mathbf{y} \mid \mathbf{w}, \mathbf{a}) = \mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{a}'}|\mathbf{W},\mathbf{A}}(\mathbf{y} \mid \mathbf{w}, \mathbf{a})$*, for all* $\mathbf{y}$ *and for any value* $\mathbf{a}'$ *attainable by* $\mathbf{A}$*.*

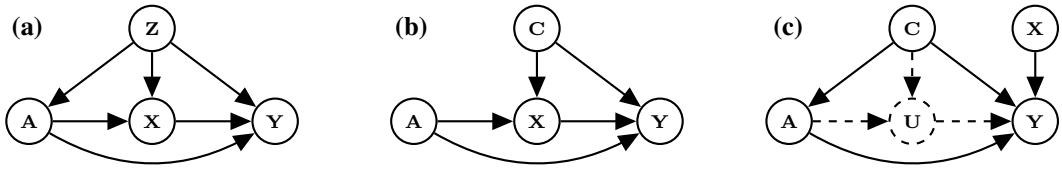


Figure 2: **(a) Assumed causal structure for the synthetic experiments** (see Section 4.1). The precise corresponding generative random variables are described in Appendix E. **(b) Assumed causal structure for the Adult dataset**, where $\mathbf{A}$ consists of the protected attributes *gender* and *age*.**(c) Causal structure for the constructed dSprites ground truth**, where $\mathbf{A} = \{\text{Pos.X}\}$, $\mathbf{U} = \{\text{Scale}\}$, $\mathbf{C} = \{\text{Shape}, \text{Pos.Y}\}$, $\mathbf{X} = \{\text{Color}, \text{Orientation}\}$, and $\mathbf{Y} = \{\text{Outcome}\}$. $\mathbf{U}$ is unobserved.

Text classification. Veitch et al. (2021) motivate the importance of counterfactual invariance in text classification tasks. We now provide a detailed comparison with our method on the (limited) causal models studied in their work, emphasizing that our definition and method applies to a much broader class of causal models. Specifically, we consider the causal and anti-causal structures (Figure 1 in Veitch et al. (2021)). Both diagrams consist of a protected attribute $\mathbf{Z}$, an observed covariate $\mathbf{X}$, and the outcome $\mathbf{Y}$. Veitch et al. (2021) provide *necessary* conditions for counterfactual invariance. They prove that if $\hat{\mathbf{Y}}^*_{\mathbf{z}} = \hat{\mathbf{Y}}^*_{\mathbf{z}'}$ almost surely, then *assuming their proposed causal and anti-causal structures* it holds $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{Z}$ and $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{Y}$, respectively. Our work extends and complements their results, showing that *in their setting*, conditional independence is a *sufficient* condition for a slightly weaker notion of counterfactual invariance, if there is no confounding between $\mathbf{Y}$ and $\mathbf{Z}$. The following result is a direct consequence of Theorem 3.2.

Corollary 3.7. *Under the causal and anti-causal graph, suppose that* $\mathbf{Z}$ *and* $\mathbf{Y}$ *are not confounded. If* $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{Z}$*, it holds* $\mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{z}}}(\mathbf{y}) = \mathbb{P}_{\hat{\mathbf{Y}}^*_{\mathbf{z}'}}(\mathbf{y})$ *almost surely, for all* $\mathbf{z}, \mathbf{z}'$ *in the domain of* $\mathbf{Z}$.

We provide experiments in Appendix E.6 to show that our method, based on Corollary 3.7, is suitable to enforce counterfactual invariance in the limited setting studied by Veitch et al. (2021).

4 Experiments

4.1 Synthetic Experiments

We begin our empirical assessment of HSCIC on synthetic datasets, where ground truth is known and systematically study how performance is influenced by the dimensionality of the variables. We simulate different datasets following to the causal graphical structure in Figure 2(a). The datasets are composed of four sets of observed continuous variables: (i) the prediction target $\mathbf{Y}$, (ii) the variable(s) we want to be counterfactually invariant in $\mathbf{A}$, (iii) observed covariates that mediate effects from $\mathbf{A}$ on $\mathbf{Y}$, and (iv) observed confounding variables $\mathbf{Z}$. The goal is to learn a predictor $\hat{\mathbf{Y}}$ that is counterfactually invariant in $\mathbf{A}$ with respect to $\mathbf{W} := \mathbf{A} \cup \mathbf{X} \cup \mathbf{Z}$. We consider various artificially generated datasets for this example, which mainly differ in the dimension of the observed variables and their correlations. A detailed description of each dataset is deferred to Appendix E.

Model choices and parameters. For all synthetic experiments, we train the model using fully connected networks (MLPs). We use the **MSE** loss $\mathcal{L}_{\text{MSE}}(\hat{\mathbf{Y}})$ as the predictive loss $\mathcal{L}$ in eq. (2) for continuous outcomes $\mathbf{Y}$. We generate 4k samples from the observational distribution in each setting and use an 80 to 20 train to test split. All metrics reported are on the test set. We perform hyperparameter tuning for MLP hyperparameters based on a random strategy (see Appendix E for details). The $\text{HSCIC}(\hat{\mathbf{Y}}, \mathbf{A} \mid \mathbf{Z})$ term is computed as in eq. (3) using a Gaussian kernel with amplitude 1.0 and length scale 0.1. The regularization parameter λ for the ridge regression coefficients in eq. (1) is set to $\lambda = 0.01$. We set $d = 1000$ and $k = 500$ in the estimation of **VCF**.

Model performance. We first perform a set of experiments to study the effect of the HSCIC, and to highlight the trade-off between accuracy and counterfactual invariance. For this set of experiments, we generate a dataset as described in Appendix E.1. Figure 3 (left) shows the variance term of **VCF** for different regularization parameters γ, as a function of the values of $\mathbf{Z}$ (i.e., before taking the outer expectation). We observe that increasing values of γ lead to a consistent decrease of the variances w.r.t. the interventional value $\mathbf{a}$. Figure 3 (center) shows the values attained by the HSCIC and **MSE** for increasing γ, demonstrating the expected trade-off in raw predictive performance and enforcing counterfactual invariance. Finally, Figure 3 (right) highlights the usefulness of HSCIC as a measure of counterfactual invariance, being in strong agreement with **VCF** (see discussion after eq. (4)).

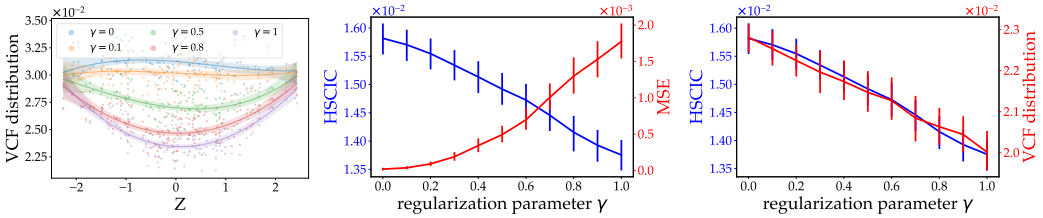


Figure 3: **(Left) variance of the counterfactual distributions** for 100 random datapoints with lines representing 3rd-order polynomial regression and shaded areas being 95% confidence intervals. **(Center) trade-off between the accuracy and counterfactual invariance.** We observe that the HSCIC decreases, as the **MSE** increases. Vertical bars denote standard errors over 15 different random seeds. **(Right) Correspondence between the HSCIC and the VCF**, for increasing γ. Again, vertical bars denote standard errors over 15 different random seeds.

Table 1: **Performance of the HSCIC against baselines** CF1 and CF2 on two synthetic datasets (see Appendix E.2). Notably, for $\gamma = 5$ in Scenario 1 and $\gamma = 13$ in Scenario 2 we outperform CF2 in **MSE** and **VCF** simultaneously.

	Scenario 1			Scenario 2		
	MSE $\times 10^3$	**HSCIC** $\times 10^3$	**VCF** $\times 10^3$	**MSE** $\times 10^3$	**HSCIC** $\times 10^3$	**VCF** $\times 10^3$
$\gamma = 0$	0.36 ± 0.5	16 ± 0.06	20.33 ± 1	2 ± 0.1	17 ± 0.2	239 ± 0.4
$\gamma = 5$	17 ± 0.03	8.77 ± 0.03	6.84 ± 1	43 ± 8	14.1 ± 0.8	230 ± 6
$\gamma = 10$	19.80 ± 2	8.16 ± 0.04	5.89 ± 0.6	50 ± 8	12.1 ± 0.5	190 ± 3
$\gamma = 13$	22 ± 1	7.9 ± 0.08	5.78 ± 0.5	133 ± 6	9.9 ± 0.3	157 ± 7
CF1	24.44 ± 3	7.9 ± 0.05	0	218 ± 10	0.8 ± 0.05	0
CF2	19.50 ± 2	14.0 ± 0.1	7.50 ± 1	137 ± 4	17.7 ± 0.3	167 ± 2

Comparison with baselines. We now compare our method against baselines in two settings, which we refer to as Scenario 1 and Scenario 2. These two settings differ in how the conditioning set $\mathbf{Z}$ affects the outcome $\mathbf{Y}$, with Scenario 1 exhibiting higher correlation of $\mathbf{Z}$ with both the mediator $\mathbf{X}$ and the outcome $\mathbf{Y}$. We refer the reader to Appendix E.2 for a precise description of Scenario 1 and Scenario 2. Since counterfactually invariant training has not received much attention yet, our our choice of baselines for experimental comparison is highly limited. Corollary 3.7 together with the fact that Veitch et al. (2021) in practice only enforce distributional equality implies that our method subsumes theirs in the causal setting they have proposed. While a direct comparison thus can only yield limited insights, we still benchmarked our method against Veitch et al. (2021) in the limited causal and anti-causal settings of Figure 6(b-c) in Appendix E.6, showing that even there our method performs on par or better than theirs. However, because counterfactual fairness is a special case of counterfactual invariance, we can compare to two methods proposed by Kusner et al. (2017) (in applicable settings). We compare to the *Level 1* (only use non-descendants of $\mathbf{A}$ as inputs to $\hat{\mathbf{Y}}$) and the *Level 3* (assume an additive noise model and in addition to non-descendants, only use the residuals of descendants of $\mathbf{A}$ after regression on $\mathbf{A}$ as inputs to $\hat{\mathbf{Y}}$) approaches of Kusner et al. (2017). We refer to these two baselines as CF1 and CF2 respectively. We summarize the results on the two scenarios in Table 1. For a suitable choice of γ, our method outperforms the baseline CF2 in both **MSE** and **VCF** simultaneously. While CF1 satisfies counterfactual invariance perfectly by construction (**VCF** $= 0$), its **MSE** is substantially higher than for all other methods. Our method provides to flexibly trade predictive performance for counterfactual invariance via a single tuning knob λ and pareto-dominates existing methods.

Multi-dimensional variables. We perform a third set of experiments to assess HSCIC's performance in higher dimensions. We consider simulated datasets (described in Appendix E.3), where we independently increase the dimension of $\mathbf{Z}$ and $\mathbf{A}$ in two different simulations, leaving the rest of the variables unchanged. The results in Tables 2 for different regularization coefficients γ and different dimensions of $\mathbf{A}$ and $\mathbf{Z}$ demonstrate that HSCIC can handle multi-dimensional variables while maintaining performance, as counterfactual invariance is approached when γ increases. We provide results where both $\mathbf{A}$ and $\mathbf{Z}$ are multivariate in Appendix E.3.

Table 2: **Results of the MSE ($\times 10^5$ for readability), HSCIC, VCF for increasing dimension of A (top) and Z (bottom)**, on synthetic datasets as in Appendix E.3. All other variables are one-dimensional in both cases.

	dimA=2			dimA=5			dimA=10			dimA=20		
γ	**MSE**	**HSCIC**	**VCF**	**MSE**	**HSCIC**	**VCF**	**MSE**	**HSCIC**	**VCF**	**MSE**	**HSCIC**	**VCF**
0	35	0.0230	0.025	1.53	0.0177	0.013	10	0.0150	0.004	1.45	0.0126	0.0010
$\frac{1}{2}$	86	0.0214	0.022	1580	0.0157	0.010	40	0.0135	0.002	33	0.0115	0.0004
1	210	0.0200	0.220	1450	0.0144	0.008	100	0.0127	0.001	60	0.0111	0.0003

	dimZ=2			dimZ=3			dimZ=5		
γ	**MSE**	**HSCIC**	**VCF**	**MSE**	**HSCIC**	**VCF**	**MSE**	**HSCIC**	**VCF**
0	20	0.002024	6.15×10^{-5}	11	0.03407	0.023	53.9	0.033180	0.0062
$\frac{1}{2}$	20	0.002023	1.29×10^{-5}	49	0.03406	0.016	21.9	0.033178	0.0061
1	20	0.002019	4.40×10^{-6}	57	0.03403	0.009	18	0.033176	0.0039

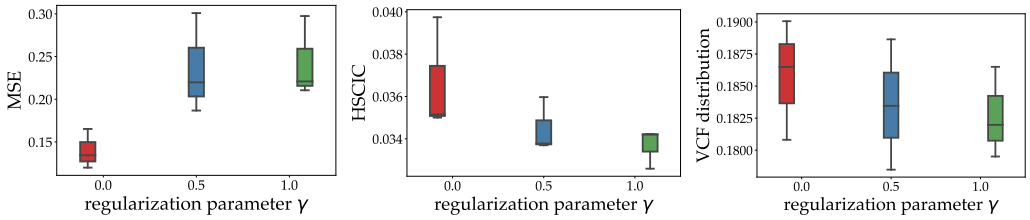


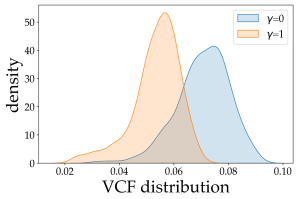
Figure 4: **Results of MSE, HSCIC operator and VCF for the dSprites image dataset experiment.** The HSCIC operator decreases steadily with higher values of γ. Similarly, a necessary increase of **MSE** can be observed. For both $\gamma = 0.5$ and $\gamma = 1$ an overall decrease of **VCF** is observed compared to the not-regularized setting.

4.2 High-dimensional image experiments

We now perform a set of experiments on a high-dimensional real-world dataset. We consider the image classification task on the dSprites dataset (Matthey et al., 2017). Since this dataset is fully synthetic and labelled (we know all factors for each image), we consider a causal model as depicted in Figure 2(c). The full structural equations are provided in Appendix E.4, where we assume a causal graph over the determining factors of the image, and essentially look up the corresponding image in the simulated dataset. This experiment is particularly challenging due to the mixed categorical and continuous variables in $\mathbf{C}$ (`shape`, `y-pos`) and $\mathbf{X}$ (`color`, `orientation`), continuous $\mathbf{A}$ (`x-pos`). All variables except for scale are assumed to be observed, and all variables jointly with the actual high-dimensional image determine the outcome $\mathbf{Y}$. Our goal is to learn a predictor $\hat{\mathbf{Y}}$ that is counterfactually invariant in the x-position with respect to all other observed variables. In the chosen causal structure, $\{\texttt{shape}, \texttt{y-pos}\} \in \mathbf{C}$ block all non-directed paths from $\mathbf{A} \cup \mathbf{C} \cup \mathbf{U}$ to $\mathbf{Y}$. Hence, we seek to achieve $\hat{\mathbf{Y}} \perp\!\!\!\perp \texttt{x-pos} \mid \{\texttt{shape}, \texttt{y-pos}\}$ via the HSCIC operator. To accommodate the mixed input types, $\hat{\mathbf{Y}}$ puts an MLP on top of features extracted from the images via convolutional layers concatenated with features extracted from the remaining inputs via an MLP. Figure 4 demonstrates that HSCIC achieves improved **VCF** as γ increases up to a certain point while affecting **MSE**, an inevitable trade-off. Adjusting our method to achieve higher stability is an interesting direction for future work.

4.3 Fairness with continuous protected attributes

In this experiment we apply our method to the popular UCI Adult dataset (Kohavi & Becker, 1996). Our goal is to predict whether an individual's income is above a certain threshold $\mathbf{Y} \in \{0, 1\}$ based on a collection of (demographic) information including protected attributes such as gender and age. We follow Nabi & Shpitser (2018); Chiappa (2019), where a subset of variables are selected from the dataset and a causal structure is assumed as in Figure 2(b) (see Appendix E.5 and Figure 7 for details). We choose gender (considered binary in this dataset) and age (considered continuous) as



	Accuracy	HSCIC	VCF
$\gamma = 0$	$(81.725 \pm 0.0481)\%$	$0.01067 \pm 3.4 \times 10^{-5}$	0.06685 ± 0.006
$\gamma = 1$	$(81.547 \pm 0.0745)\%$	$0.00739 \pm 6.8 \times 10^{-5}$	0.05156 ± 0.003

Figure 5: **(Left) Distribution of VCF values** (unnormalized) for different choices of γ. We observe less variance and more mass near zero for regularization parameter $\gamma = 1$. **(Right) Results on accuracy, HSCIC and VCF**, showing a considerable improvement of HSCIC and **VCF** for $\gamma = 1$.

the protected attributes $\mathbf{A}$. We denote the marital status, level of education, occupation, working hours per week, and work class jointly by $\mathbf{X}$ and combine the remaining observed attributes in $\mathbf{C}$. With this notation, our goal is to learn a predictor $\hat{\mathbf{Y}}$ that is counterfactually invariant (or fair) in $\mathbf{A}$ with respect to $\mathbf{W} = \mathbf{C} \cup \mathbf{X}$. This amounts to learning a classifier for the income, which is counterfactually invariant in both age and gender with respect to the other attributes. We remark that achieving fairness for continuous or even mixed categorical and continuous protected attributes is an ongoing line of research (even for non-causal fairness notions) (Mary et al., 2019; Chiappa & Pacchiano, 2021), but directly supported by HSCIC.

In this experiment we use an MLP with binary cross-entropy loss for $\hat{\mathbf{Y}}$. Since this experiment is based on real data, the true counterfactual distribution cannot be known. Hence, we follow Chiappa & Pacchiano (2021) and estimate a possible true SCM by inferring the posterior distribution over the unobserved variables using variational autoencoders (Kingma & Welling, 2014). Even though this only provides approximate VCF, Figure 5 (left) shows that HSCIC achieves more counterfactually fair outcome distributions (more mass near zero) than an unconstrained classifier ($\gamma = 0$). Figure 5 (right) highlights once more that the HSCIC operator is in agreement with the VCF, again trading off accuracy.

5 Discussion and Future Work

We studied the problem of learning predictors $\hat{\mathbf{Y}}$ that are counterfactually invariant in changes of certain covariates. We put forth a formal definition of counterfactual invariance and described how it generalizes existing notions. Next, we provided a novel sufficient graphical criterion to characterize counterfactual invariance and reduce it to a conditional independence statement in the observational distribution. Our method does not require identifiability of the counterfactual distribution or exclude the possibility of unobserved confounders. Finally, we propose an efficiently estimable, model-agnostic regularization term to enforce this conditional independence (and thus counterfactual invariance) based on kernel mean embeddings of conditional measures, which works for mixed continuous/categorical, multi-dimensional variables. We demonstrate the efficacy of our method in regression and classification tasks involving controlled detailed simulation studies, high-dimensional images, and in a fairness application, where it outperforms existing baselines.

The main limitation of our work, shared by all studies in this domain, is the assumption that the causal graph is known. Another limitation is that our methodology is applicable only when our graphical criterion is satisfied, requiring a certain set of variables to be observed (albeit unobserved confounders are not generally excluded). From an ethics perspective, the increased robustness of counterfactually invariant (or societal benefits of counterfactually fair) predictors are certainly desirable. However, this presupposes that all the often untestable assumptions are valid. Overall, causal methodology should not be applied lightly, especially in high-stakes and consequential decisions. A critical analysis of the broader context or systemic factors may hold more promise for societal benefits, than a well-crafted algorithmic predictor.

In light of these limitations, an interesting and important direction for future work is to assess the sensitivity of our method to misspecifications of the causal graph or insufficient knowledge of the required blocking set. Finally, we believe our graphical criterion and KME-based regularization can also be useful for causal representation learning, where one aims to isolate causally relevant, autonomous factors underlying the data generating process of high-dimensional data.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, volume 70, pp. 253–262, 2017.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *The Journal of Machine Learning Research*, 21:90–1, 2020.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Silvia Chiappa and Aldo Pacchiano. Fairness with continuous optimal transport. *arXiv preprint arXiv:2101.02084*, 2021.

Erhan Çinlar and E. ðCınlar. *Probability and stochastics*, volume 261. Springer, 2011.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Nicolae Dinculeanu. *Vector integration and stochastic integration in Banach spaces*, volume 48. John Wiley & Sons, 2000.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, volume 20, 2007.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, volume 3734, pp. 63–77, 2005.

Steffen Grünewälder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *International conference on machine learning*, 2012.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems*, volume 30, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. URL http://arxiv.org/abs/1312.6114.

Ronny Kohavi and Barry Becker. Uci adult data set. *UCI Meachine Learning Repository*, 1996.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, volume 97, pp. 4382–4391. PMLR, 2019.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, pp. 21247–21259, 2020.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2007.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1219–1226, 2006.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Ilya Shpitser and Judea Pearl. Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence,*, pp. 514–521, 2009.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Le Song, Jonathan Huang, Alexander J. Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*, volume 382, pp. 961–968, 2009.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:233:1–233:29, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, pp. 16196–16208, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268, 2020. URL https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf.

Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. Algorithmic decision making with conditional fairness. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, pp. 2125–2135, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.