Journal of Machine Learning Research 25 (2024) 1-65 Submitted 9/21; Revised 3/24; Published 8/24

Desiderata for Representation Learning: A Causal Perspective

Yixin Wang YIXINW@UMICH.EDU
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
EECS and Statistics, University of California, Berkeley, CA 94720, USA
INRIA, 75013 Paris, France

Editor: Silvia Chiappa

Abstract

Representation learning constructs low-dimensional representations to summarize essential features of high-dimensional data. This learning problem is often approached by describing various desiderata associated with learned representations; e.g., that they be non-spurious, efficient, or disentangled. It can be challenging, however, to turn these intuitive desiderata into formal criteria that can be measured and enhanced based on observed data. In this paper, we take a causal perspective on representation learning, formalizing non-spuriousness and efficiency (in supervised representation learning) and disentanglement (in unsupervised representation learning) using counterfactual quantities and observable consequences of causal assertions. This yields computable metrics that can be used to assess the degree to which representations satisfy the desiderata of interest and learn non-spurious and disentangled representations from single observational datasets.

Keywords: Causal inference, representation learning, non-spuriousness, disentanglement, probabilities of causation

1. Introduction

Representation learning refers to the problem of constructing low-dimensional representations that summarize essential features of high-dimensional data. For example, one may be interested in learning a low-dimensional representation of MNIST images, where each image is a 784-dimensional vector of pixel values. Alternatively, one may be interested in a corpus of text reviews of products; each review is a $5,000$-dimensional word count vector. Given an m-dimensional data point, $\mathbf{X} = (X_1, \ldots, X_m) \in \mathbb{R}^m$, the goal is to find a d-dimensional representation, $\mathbf{Z} = (Z_1, \ldots, Z_d) \triangleq (f_1(\mathbf{X}), \ldots, f_d(\mathbf{X}))$, that captures d important features of the data, where $f_j : \mathbb{R}^m \to \mathbb{R}, j = 1, \ldots, d$ are d deterministic functions and $d \ll m$.

A heuristic approach to the problem has been to fit a neural network that maps from the high-dimensional data to a set of labels, and then take the top layer of the neural network as the representation of the data point. When labels are not available, a related heuristic is to fit a latent variable model (e.g., a variational autoencoder (Kingma & Welling, 2014)) and output a low-dimensional representation based on the inferred latent variables. In both cases, the hope is that these low-dimensional representations will be useful in the performance of downstream tasks and also provide insight into the statistical relationships underlying the data.

These heuristic approaches do not, however, always succeed in producing representations with desirable properties. For example, as we will discuss in detail, common failure modes involve capturing spurious features that do not transfer well or finding dimensions that are entangled and are hard to interpret. For example, in fitting a neural network to images of animals, with the goal of producing a labeling of the species found in the images, a network may capture spurious background features (e.g., grass) that are highly correlated with the animal features (e.g., the face of a dog). Such spurious features can often predict the label well. But they are generally not useful for prediction in a different dataset or for performing other downstream tasks. The learned representation may also be entangled—a single dimension of the representation may encode information about multiple features (e.g., animal fur and background lighting). Such entangled representations are hard to interpret. The representation itself does not provide guidance on how to separate learned dimensions into more informative dimensions that describe animal fur and background lighting.

While non-spuriousness or disentanglement are natural desiderata of representations, they are often intuitively defined, challenging to evaluate, and hard to optimize over algorithmically. Lacking formal metrics for these desiderata, and not having access to manually labeled features (e.g., grass, animal fur, or background lighting) that provide empirical guidance, prevents us from developing representation learning algorithms that satisfy natural desiderata.

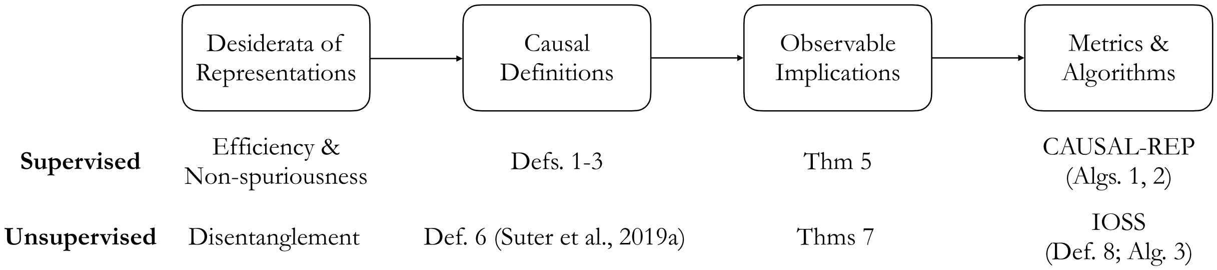
In this work, we take a causal inference perspective on representation learning. This perspective allows us to formalize representation learning desiderata using causal notions, specifically causal relationships among the label Y and the d features captured by $Z_1, \ldots, Z_d$. This yields calculable metrics obtained from the observable implications of the underlying causal relationships. These metrics then enable representation learning algorithms that target these desiderata. (Figure 1 illustrates this workflow.)

We focus on two sets of desiderata in representation learning: (1) efficiency and non-spuriousness in supervised representation learning—i.e., the representations shall efficiently capture non-spurious features of the data—and (2) disentanglement in unsupervised representation learning—i.e., the representations shall encode different features along separate dimensions.

In the supervised setting, the key idea is to view representations as capturing features that are potential *causes* of the label. From this perspective, a non-spurious representation should capture features that are *sufficient causes* of the label. This property ensures that the same representation is still likely to be informative of the label when employed on a new dataset. In the running example of images, a representation that captures the dog-face feature is non-spurious because the presence of a "dog-face" is sufficient to determine the dog label (i.e., whether a dog is present in the image). In contrast, a representation based on the "grass" feature is spurious because it cannot causally determine the dog label, even though it is highly correlated. We formalize this connection between non-spurious representations and sufficient causes by defining non-spuriousness using the notion of *probability of sufficiency* (PS), due to Pearl (2011).

The same causal perspective also allows us to formalize the efficiency of representations, which are those that do not pick up redundant features. Again viewing the representation as a potential cause of the label, its efficiency corresponds to the necessity of the cause. In the image example, a representation of both "dog-face" and "four-leg" is inefficient because it is not necessary to know both "dog-face" and "four-leg" to determine the dog label. As dogs always have four legs (in our simplified world), "four-leg" is a redundant feature given "dog-face." We therefore formalize efficiency of a representation using the notion of *probability of necessity* (PN). As we will discuss, PN and PS are aspects of a general notion of *probabilities of causation*, also due to Pearl (2011).

While these causal definitions formalize efficiency and non-spuriousness, they do not immediately imply calculable metrics for these desiderata, because not all causal quantities are estimable from



Supervised	Efficiency & Non-spuriousness	Defs. 1-3	Thm 5	CAUSAL-REP (Algs. 1, 2)
Unsupervised	Disentanglement	Def. 6 (Suter et al., 2019a)	Thms 7	IOSS (Def. 8; Alg. 3)

Figure 1: A causal perspective of representation learning: From desiderata to algorithms

observational data. To obtain calculable metrics, we have to study observable implications of these causal definitions, a problem known as *causal identification* (Pearl, 2011). We discuss the unique challenges of causal identification with high-dimensional data—specifically, high-dimensional data are often rank-degenerate (Stewart, 1984; Golub et al., 1976)—and we develop identification strategies to address these challenges. These strategies lead to calculable metrics of efficiency and non-spuriousness in the high-dimensional setting, along with the conditions under which they are valid.

Based on these definitions, we develop CAUSAL-REP, an algorithmic framework that formulates representation learning as a task of finding necessary and sufficient causes. We consider both a supervised variant of CAUSAL-REP for supervised learning and an unsupervised counterpart for instance discrimination, the latter of which is closely related to contrastive learning. In a range of empirical studies, we find that CAUSAL-REP is more successful at finding non-spurious representations in both images and text compared to standard benchmarks.

In the unsupervised setting we will take the causal perspective further, focusing on the desideratum of *disentangled representations*. Intuitively, disentanglement requires that different dimensions of the learned representation correspond to independent degrees of freedom of objects; that is, we can view disentangled representations as making it possible to generate new examples of objects by separately manipulating the values of the features encoded by such representations.

To develop metrics for disentanglement, we begin with a causal definition of disentanglement due to Suter et al. (2019): different dimensions of the learned representation should correspond to different features that do not causally affect each other. (These features may still be correlated.) While a useful starting place, this definition unfortunately leaves us short of the goal of measuring and optimizing for disentanglement. Naively, it requires one to assess whether causal connections exist among different features, which is a generally impossible task without substantial knowledge about the underlying causal structure (Pearl, 2011; Imbens & Rubin, 2015). In particular, we may observe correlation among variables, with or without causal connections. Due to this difficulty, existing disentanglement metrics often rely on ground-truth features or auxiliary labels (Suter et al., 2019; Thomas et al., 2018), which we often do not have access to in practice.

To obtain an operational measure of disentanglement, we again turn to studying the observable implications of its causal definition. The key observation is that the observable implications can exist in the *support* of the representation, i.e., the set of values the representation can take. Specifically, we find that disentanglement—the absence of causal connections among the features captured by different dimensions of the representation—implies that their support must be independent in observational data (under a standard positivity condition), namely the set of possible values each feature can take does not depend on the values of other features. Visually, features with independent support must have a scatter plot that occupies a hyperrectangular region; see Figures 8a to 8c for examples.

Building on this connection between disentanglement and independent support, we develop the *independence-of-support score* (IOSS) as an unsupervised disentanglement metric; it evaluates whether different dimensions of the representation have independent support. It also enables us to enforce disentanglement in representation learning via an IOSS penalty.

Through the study of non-spuriousness/efficiency and disentanglement, we illustrate how causality can provide a fruitful perspective for representation learning. Many intuitively defined desiderata in representation learning can be formalized using causal notions. These causal notions can further lead to metrics that quantify these desiderata and representation learning algorithms that enforce these desiderata.

Related work. There has been a flurry of recent work at the intersection of representation learning and causal inference. We provide a brief review, highlighting the contrast between this literature and our work.

Learning non-spurious representations via invariance. Many works have considered causal formulations of representation learning given datasets from multiple environments (Khasanova & Frossard, 2017; Zhao et al., 2019; Moyer et al., 2018; Lu et al., 2021; Mitrovic et al., 2020; Moraffah et al., 2019; Arjovsky et al., 2019; Cheng & Lu, 2017; Veitch et al., 2021; Creager et al., 2021; Puli et al., 2021). The basic idea is to enforce the invariance of the mapping between the learned representation and the outcome label, thereby encouraging non-spurious representations and enabling out-of-distribution prediction. Similar to these works, our CAUSAL-REP algorithm targets non-spurious representations. However, it differs in its focus on the setting in which only a single observational dataset is available. That is, we do not assume access to datasets from multiple environments, nor do we leverage the invariance principle.

Reverse causal inference. Our work is also related to a body of work on reverse causal inference, a task that aims to find "causes of effects" (Schölkopf et al., 2013; Janzing & Schölkopf, 2015; Weichwald et al., 2014; Kilbertus et al., 2018; Schölkopf et al., 2012; Paul, 2017; Wang & Culotta, 2020, 2021; Kommiya Mothilal et al., 2021; Galhotra et al., 2021; Watson et al., 2021; Chalupka et al., 2015, 2017; Gelman & Imbens, 2013). Existing approaches formulate the search for causes as causal hypotheses generation and testing. In contrast, our CAUSAL-REP algorithm formulates this search as maximizing probabilities of causation (POC) (Pearl, 2011, 2019b; Tian & Pearl, 2000; Mueller et al., 2021), specifically in the context of representation learning. We develop identification conditions for these POC for high-dimensional data such as images and text (Nabi et al., 2020; Puli et al., 2020; Wang & Blei, 2019a, 2020; Ranganath & Perotte, 2018; Wang & Blei, 2021; Pryzant et al., 2020; Grimmer & Fong, 2021; Fong & Grimmer, 2016; Wang & Culotta, 2020, 2021). Our results build on existing identification results around multiple causes with shared unobserved confounding (Wang & Blei, 2019a, 2020; Puli et al., 2020) and its positivity issues (D'Amour et al., 2020a; D'Amour, 2019b,a; Imai & Jiang, 2019), but are tailored to a representation learning setting where no unobserved confounding is present.

Disentangled representation learning. Our independence-of-support score (IOSS) relates to the broad literature on disentangled representation learning. Early approaches to disentanglement often enforce statistical independence among different dimensions of the representation (Bengio et al., 2013; Achille & Soatto, 2018; Higgins et al., 2017; Burgess et al., 2018; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2018). However, Locatello et al. (2019a) show that this inductive bias of statistical independence is insufficient for disentanglement due to its non-identifiability. Many recent approaches to disentanglement thus incorporate auxiliary information to enable identifiability in disentanglement (Locatello et al., 2019b; Khemakhem et al., 2020; Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2019; Locatello et al., 2020a; Träuble et al., 2020; Yang et al., 2020; Shen et al., 2020). In contrast to these works, IOSS relies on the causal perspectives of Suter et al. (2019) and establishes identifiability in disentanglement without the need for supervision. Focusing on compactly supported representations, we show that representations with independent support are identifiable under suitable conditions.

Causal structure learning. Our causal approach to disentanglement also relates to causal structure learning; both involve assessing the existence of causal connections between variables. Traditional approaches to causal structure learning relies on independence tests or score-based methods, often assuming all variables are observed; see Heinze-Deml et al. (2018) for a review. IOSS differs from these works in that we allow for unobserved common causes among the observed variables; we also rely on a different observable implication of the lack of causal connections.

Representation learning for causal inference. Representation learning and dimensionality reduction have significantly improved the estimation efficiency of causal inference with high-dimensional covariates and treatments (Johansson et al., 2019; Nabi et al., 2020; Johansson et al., 2020; Shi et al., 2020; Wu & Fukumizu, 2021; Veitch et al., 2019, 2020). The focus in these work is on how representation learning can help causal inference with high-dimensional covariates. In contrast, we focus on how causal inference can help produce useful representations.

This paper. The rest of the paper is organized as follows. Section 2 develops a causal approach to efficiency and non-spuriousness in supervised representation learning. Section 2.1 formalizes efficiency and non-spuriousness using counterfactual notions, namely the probability of necessity and sufficiency in causal inference. Section 2.2 discusses the identification of these counterfactuals, which leads to metrics for efficiency and non-spuriousness. Section 2.4 formulates representation learning as a task of finding necessary and sufficient causes of the label and develops CAUSAL-REP, an algorithm that targets efficient and non-spurious features. Section 2.5 provides an empirical study of CAUSAL-REP, showing that it captures non-spurious features and improves downstream out-of-distribution prediction. Section 3 switches gears to develop a causal approach to disentanglement in unsupervised representation learning. Section 3.1 reviews a causal definition of disentanglement. Section 3.2 studies its observable implications, showing that causal disentanglement implies independent support of the representation. This leads to IOSS, an unsupervised disentanglement metric. Section 3.3 discusses the identifiability of representations with independent support, which enables disentangled representation learning with an IOSS penalty. Section 3.4 demonstrates the effectiveness of IOSS in empirical studies. Finally, Section 4 concludes the paper with a discussion of causal perspectives on representation learning.

2. Supervised Representation Learning: Efficiency and Non-spuriousness

We begin by formulating a set of desiderata for supervised representation learning. As a running example, we consider a dataset of images and their labels. We focus on settings where the labels are obtained by annotators viewing each image and then labeling; they describe the perceived features of the image, as opposed to the intended features of the image. The goal is to construct a low-dimensional image representation that efficiently captures its essential features.

Formally, we consider a dataset that contains n m-dimensional data points and their labels, $\{\mathbf{x}_i, y_i\}_{i=1}^n \in \mathcal{X}^m \times \mathcal{Y}$, assumed to be sampled i.i.d., with $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$. The variable x_{il} might be the value of the lth pixel of an image, with $\mathcal{X} = [0, 255]$; it might also be the number of times the lth word in the vocabulary occur in the document, with $\mathcal{X} = \mathbb{Z}^+$. The goal of representation learning is to find a deterministic function $f : \mathcal{X}^m \rightarrow \mathbb{R}^d$ mapping the m-dimensional data into a d-dimensional space ($d \ll m$). Thus each data point $\mathbf{x}_i$ has a d-dimensional representation $\mathbf{z}_i \triangleq f(\mathbf{x}_i)$, or equivalently $\mathbf{z}_i = (z_{i1}, \ldots, z_{id}) \triangleq (f_1(\mathbf{x}_i), \ldots, f_d(\mathbf{x}_i))$, with $f_j : \mathcal{X}^m \rightarrow \mathbb{R}$.

Ideally, such a representation should be efficient and non-spurious. It should efficiently capture features of the image that are essential for determining the label. Below we use counterfactual notions to formalize the desiderata of efficiency and non-spuriousness.

2.1 Defining Efficiency and Non-spuriousness using Counterfactuals: Criteria

To formalize the desiderata of efficiency and non-spuriousness, we posit a structural causal model (SCM) that describes the data generating process. This SCM will enable us to define these properties through counterfactual quantities.

2.1.1 A structural causal model of supervised representation learning

We describe a structural causal model (SCM) for supervised representation learning and the counterfactual quantities therein.

The structural causal model (SCM) of the labeled dataset. We posit Figure 2 as the SCM of a data-generating process. (For notational simplicity, we suppress the data point index i and focus on categorical Y.) The figure embodies the assumption that the high-dimensional object $\mathbf{X} = (X_1, \ldots, X_m)$ causally affects its label Y. Moreover, there is no confounder between the two; the label Y can be fully determined by the object $\mathbf{X}$. These assumptions hold because we assume the labels are obtained by annotators viewing the images; the labeling is solely based on the images.

Figure 2 also posits that different dimensions of the high-dimensional object, $X_1, \ldots, X_m$, are correlated due to some unobserved common cause $\mathbf{C}$, which can potentially be multi-dimensional. For example, $\mathbf{C}$ can represent the design of an image; nearby pixels tend to be highly correlated if an image has a smooth design. This assumption holds because pixel values of images are not independent; they are determined by common latent variables (Kingma & Welling, 2014; Goodfellow et al., 2014).

The representation $f(\mathbf{X})$ and its interventions. The representation $\mathbf{Z} = f(\mathbf{X})$ of a high-dimensional data point is often a low-dimensional vector that captures important features of the high-dimensional data point $\mathbf{X}$. For example, $\mathbf{Z} = (Z_1, Z_2) = (\mathbb{I}\{X_{125} \times X_{38} < 0.01\}, X_{164} + 0.3 \cdot X_{76}^2)$ is a two-dimensional representation of $\mathbf{X}$, whose first dimension represents whether pixels 125 and 38 are close to black. More complex representations may capture more meaningful features; e.g., whether a dog face is present in the image in the running example.

The representation $\mathbf{Z} = f(\mathbf{X})$ is not included in the SCM as a separate causal variable. The reason is that the representation $\mathbf{Z} = f(\mathbf{X})$ is a *deterministic* function of the image pixels $\mathbf{X}$ and the causal variables of a SCM must be related by functional relationships perturbed by *random* disturbances (Pearl, 2011, 1995). In particular, $\mathbf{Z}$ is not a descendant of $\mathbf{X}$ because a value change in $\mathbf{Z}$ would imply a value change in $\mathbf{X}$ (unless f is a constant function). Nor is $\mathbf{Z}$ an ancestor of $\mathbf{X}$ because a change in the value of $\mathbf{X}$ can also imply a change in $\mathbf{Z}$. Despite $\mathbf{Z}$ being not explicitly included in the SCM, Figure 2 does imply that $\mathbf{Z} = f(\mathbf{X})$ is an ancestor of the label Y. This is because $\mathbf{X}$ is an ancestor of Y.

A reader may wonder: How is an intervention on features defined? There exist many ways to intervene on the same feature. For example, to turn off the feature "dog-face," one can either erase the dog from the image or turn the dog face away. Despite this ambiguity, interventions on the representation $\mathbf{Z} = f(\mathbf{X})$ are well defined; they are *functional interventions* (Puli et al., 2020; Correa & Bareinboim, 2020; Eberhardt & Scheines, 2007), also referred to as *stochastic policies* (Pearl, 2011, Ch. 4.2). At a high level, each intervention on features follows a probabilistic intervention on pixels, where the distribution involves all such possible interventions on pixels that can lead to the desired feature change. (We will define them rigorously in Section 2.2.)

Roughly, suppose a one-dimensional representation $Z = f(\mathbf{X})$ captures the univariate binary feature of whether the grass is present in the image. Then the intervention $\mathrm{do}(Z = 1)$—equivalently $\mathrm{do}(f(\mathbf{X}) = 1)$—means "turning on" the grass feature in the image. That is, holding the design of the image $\mathbf{C}$ fixed, we change the pixels of the image $\mathbf{X}$ such that the grass is present: $Z = 1$. Accordingly, the intervention $\mathrm{do}(Z = 0)$ means "turning off" the grass feature. Holding the design $\mathbf{C}$ fixed, we change the pixels $\mathbf{X}$ such that the grass is absent, $Z = 0$.

Counterfactuals in the SCM. The SCM (Figure 2) results in a family of counterfactuals, "what would the value of a variable be if we intervene on some variables of the causal model." Here we focus on the counterfactuals obtained when we intervene on the image features captured by the representation $\mathbf{Z}$. These counterfactuals will allow us to define the efficiency and non-spuriousness of the representations.

We denote $Y(\mathbf{Z} = \mathbf{z})$ as the counterfactual label of an image if we force its representation $\mathbf{Z}$ to take value $\mathbf{z}$.[1] For example, if a one-dimensional representation Z captures the feature "whether grass is present in the image," then $Y(Z = 0)$ is the counterfactual label had the image had no grass. Accordingly, $Y(Z = 1)$ is the counterfactual label had the image had grass being present.

The key idea here is to evaluate the quality of the representation $\mathbf{Z}$ by reasoning about the counterfactual labels: What would the label Y be had $\mathbf{Z}$ taken different values? We will show how the properties of non-spuriousness and efficiency can be formalized via these counterfactuals. This connection will allow us to develop metrics and algorithms for these properties.

2.1.2 Counterfactual definitions of non-spuriousness and efficiency

We next discuss how these counterfactual notions can allow us to formalize the non-spuriousness and efficiency of a representation[2] $f(\mathbf{x})$, for a single observed data point $(\mathbf{X} = \mathbf{x}, Y = y)$.

Non-spuriousness and its counterfactual definition. A non-spurious representation captures features that can (causally) determine the label. In the causal model, we say that a representation captures non-spurious features if, given an image without the feature, including this feature would change its label.

Returning to the running example of recognizing dogs from images, suppose the label Y indicates whether the image contains a dog. Then a representation Z capturing the presence of dog-face is a non-spurious feature. Given a non-dog image without a dog face $(Y = 0, Z = 0)$, adding a dog face to the image turns on the dog label, i.e., $Y(Z = 1) = 1$. In other words, dog-face is a sufficient cause of a dog label. In contrast, a representation Z that captures the presence of grass is a spurious

[1] The counterfactual label $Y(\mathbf{Z} = \mathbf{z})$ is also commonly written as $Y_{\mathbf{z}}$ (Pearl, 2011). Here we employ the parenthesis notation $Y(\mathbf{Z} = \mathbf{z})$ to avoid subscripts.

[2] We use the word *representation* and *feature* interchangeably. The features are captured by representations.

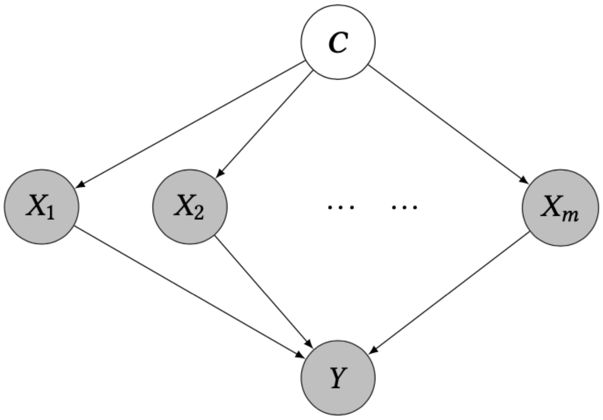


Figure 2: A SCM for supervised representation learning. $\mathbf{X} = (X_1, \ldots, X_m)$ represents the high-dimensional object (e.g., pixels of an image), Y represents the outcome label, and $\mathbf{C}$ denotes the unobserved common cause of $X_1, \ldots, X_m$.

feature. Though the presence of grass and the dog label may be highly correlated, adding grass to a non-dog image does not turn on the dog label.

Viewing the representation $\mathbf{Z}$ as a potential cause of the label, a non-spurious representation shall be a *sufficient cause* of the label. Turning on the feature (captured by) $\mathbf{Z}$ should be sufficient to turn on the label. We thus measure the non-spuriousness of a representation by the probability of sufficiency (PS) of the feature $\mathbf{Z}$ causing the label Y.

Definition 1 (Non-spuriousness of representations)**.** *Suppose we observe a data point with representation $\mathbf{Z} = \mathbf{z}$ and label $Y = y$. Then the non-spuriousness of the representation $\mathbf{Z}$ for label Y is the probability of sufficiency (PS) of $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ for $\mathbb{I}\{Y = y\}$:*

$$PS_{\mathbf{Z}=\mathbf{z},Y=y} = P(Y(\mathbf{Z} = \mathbf{z}) = y \mid \mathbf{Z} \neq \mathbf{z}, Y \neq y). \tag{1}$$

When both the representation Z and the label Y are univariate binary with $z = 1, y = 1$, then Equation (1) *coincides with classical definition of PS (Definition 9.2.2 of Pearl (2011)).*

$PS_{\mathbf{Z}=\mathbf{z},Y=y}$ is the probability of the representation $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ being a *sufficient cause* of the label $\mathbb{I}\{Y = y\}$. In Pearl's language, it describes the capacity of the representation to "produce" the label. PS measures the probability of a positive counterfactual label if we *make* the feature $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ be present (and all else equal), conditional on the feature being absent and the label being negative $\{\mathbb{I}\{\mathbf{Z} = \mathbf{z}\} = 0, \mathbb{I}\{Y = y\} = 0\}$. A non-spurious representation should have a high PS for the label of interest.

Efficiency and its counterfactual definition. An efficient representation captures only essential features of the data; it does not capture any redundant features. In the causal model, we say that the representation is efficient if, given an image with the feature it captures, removing this feature would change its label.

Again returning to the running example of recognizing dogs from images, a representation that captures "dog-face + four-leg" is inefficient. The reason is that, for an image with both a dog face and four legs, removing one of the dog legs and hence turning off the "dog-face + four-leg" feature does not necessarily turn off the "dog label." In contrast, the feature "dog-face" is efficient as removing a dog face from an image will turn off the "dog label."

An efficient representation must capture features that are *necessary causes* of the label. We can therefore define the efficiency of a representation by the probability of necessity (PN) of the feature causing the label (Pearl, 2011).

Definition 2 (Efficiency of representations)**.** *Suppose we observe a data point with representation* $\mathbf{Z} = \mathbf{z}$ *and label* $Y = y$*. Then the efficiency of the representation* $\mathbf{Z}$ *for the label* Y *is the probability of necessity (PN) of* $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ *for* $\mathbb{I}\{Y = y\}$*:*[3]

$$PN_{\mathbf{Z}=\mathbf{z},Y=y} = P(Y(\mathbf{Z} \neq \mathbf{z}) \neq y \,|\, \mathbf{Z} = \mathbf{z}, Y = y). \tag{2}$$

When both the representation Z *and the label* Y *are univariate binary with* $z = 1, y = 1$*, then Equation* (2) *coincides with classical definition of PN (Definition 9.2.1 of Pearl (2011)).*

$PN_{\mathbf{Z}=\mathbf{z},Y=y}$ is the probability of the representation $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ being a *necessary cause* of the label $\mathbb{I}\{Y = y\}$. It measures the probability of a negative counterfactual label if we *remove* the feature (i.e., setting $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\} = 0$ and all else equal), conditional on the feature being present and the label being positive, $\{\mathbb{I}\{\mathbf{Z} = \mathbf{z}\} = 1, \mathbb{I}\{Y = y\} = 1\}$. An efficient representation should have a high PN for the label of interest.

2.1.3 Simultaneous assessment of non-spuriousness and efficiency

Representation learning often targets representations that are both non-spurious and efficient. How can one assess non-spuriousness and efficiency simultaneously?

Assessing non-spuriousness and efficiency simultaneously. We invoke the notion of probability of necessity and sufficiency (PNS). Intuitively, non-spuriousness measures how effectively we can turn on a label by turning on the feature captured by the representation; efficiency measures how effectively we can turn off a label by turning off this feature. Thus a representation is both non-spurious and efficient if the label responds to the feature in both ways. If the feature is turned on, then the label will be turned on; if the feature is turned off, then the label will be turned off too. PNS calculates exactly this probability.

Definition 3 (Efficiency & non-spuriousness of representations)**.** *Suppose we observe a data point with representation* $\mathbf{Z} = \mathbf{z}$ *and label* $Y = y$*. Then the efficiency and non-spuriousness of the representation* $\mathbf{Z}$ *for label* Y *is the probability of necessity and sufficiency (PNS) of* $\mathbb{I}\{\mathbf{Z} = \mathbf{z}\}$ *for* $\mathbb{I}\{Y = y\}$*:*

$$PNS_{\mathbf{Z}=\mathbf{z},Y=y} = P(Y(\mathbf{Z} \neq \mathbf{z}) \neq y, Y(\mathbf{Z} = \mathbf{z}) = y)). \tag{3}$$

When both the representation Z *and the label* Y *are univariate binary with* $z = 1, y = 1$*, then Equation* (3) *coincides with classical definition of PNS (Definition 9.2.3 of Pearl (2011)).*

Requiring both necessity and sufficiency of the cause is a stronger requirement than requiring only necessity (or only sufficiency). Accordingly, PNS is a weighted combination of PN and PS,

$$PNS_{\mathbf{Z}=\mathbf{z},Y=y} = P(\mathbf{Z} = \mathbf{z}, Y = y) \cdot PN_{\mathbf{Z}=\mathbf{z},Y=y} + P(\mathbf{Z} \neq \mathbf{z}, Y \neq y) \cdot PS_{\mathbf{Z}=\mathbf{z},Y=y},$$

[3] The distribution of the counterfactual $Y(\mathbf{Z} \neq \mathbf{z})$ is defined as that of a soft intervention (Correa & Bareinboim, 2020; Eberhardt & Scheines, 2007); a.k.a. a stochastic policy (Pearl, 2011, Ch. 4.2):

$$P(Y(\mathbf{Z} \neq \mathbf{z})) = \int P(Y(\mathbf{Z}))P(\mathbf{Z} \,|\, \mathbf{Z} \neq \mathbf{z})\, \mathrm{d}\mathbf{Z}.$$

as per Lemma 9.2.6 of Pearl (2011). We note that our definitions of PN, PS, PNS (Equations (1) to (3)) generalize those in Pearl (2011, Ch. 9) from univariate binary causes to general (continuous, discrete, or multi-dimensional) causes. We next consider two further extensions of the PNS notion.

Extension: Efficiency and non-spuriousness over a dataset. The discussion of efficiency and non-spuriousness (Definitions 1 to 3) has focused on individual data points up to this point, reflecting the fact that probabilities of causation (POC) notions are most commonly discussed with respect to a single occurred event—given that the event $\mathbf{Z} = \mathbf{z}$ and $Y = y$ has occurred, what is the probability that $\mathbf{Z} = \mathbf{z}$ is a sufficient and/or necessary cause of $Y = y$?

In practice, however, we are often interested in whether a representation, $f : \mathcal{X}^m \to \mathbb{R}^d$, can produce an efficient and non-spurious summary for datasets of n i.i.d. data points, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. We thus extend Definition 3 to this setting:

$$PNS_n(\mathbf{Z}, Y) \triangleq \prod_{i=1}^n PNS_{\mathbf{Z}=\mathbf{z}_i, Y=y_i} = \prod_{i=1}^n P(Y(\mathbf{Z} \neq \mathbf{z}_i) \neq y_i, Y(\mathbf{Z} = \mathbf{z}_i) = y_i)), \tag{4}$$

where $\mathbf{z}_i = f(\mathbf{x}_i)$ is the representation for data point i. Loosely, this means that a representation is efficient and non-spurious for a dataset if its efficiency and non-spuriousness holds jointly across all the data points. (One can similarly extend Definitions 1 and 2 for PS and PN.)

Extension: Conditional efficiency and non-spuriousness. For multi-dimensional representations, one is often interested in the efficiency and non-spuriousness of each of its dimensions. We expect each dimension of the representation to be efficient and non-spurious conditional on all other dimensions.

We thus extend Definition 3 to formalize a notion of *conditional efficiency and non-spuriousness*. Consider a d-dimensional representation $\mathbf{Z} = (Z_1, \ldots, Z_d) = (f_1(\mathbf{X}), \ldots, f_d(\mathbf{X}))$. The conditional efficiency and non-spuriousness of the jth dimension Z_j for data point $(\mathbf{x}_i, y_i)$ is

$$PNS_{Z_j=z_{ij}, Y=y_i \,|\, \mathbf{Z}_{-j}=\mathbf{z}_{i,-j}} = P(Y(Z_j \neq z_{ij}, \mathbf{Z}_{-j} = \mathbf{z}_{i,-j}) \neq y_i, Y(Z_j = z_{ij}, \mathbf{Z}_{-j} = \mathbf{z}_{i,-j}) = y_i), \tag{5}$$

where $z_{ij} = f_j(\mathbf{x}_i)$ is the jth dimension of the representation, and $z_{i,-j} = (z_{ij'})_{j' \in \{1,\ldots,d\} \setminus j}$. Accordingly, the conditional efficiency and non-spuriousness of Z_j across all n data points is

$$PNS_n(Z_j, Y \,|\, \mathbf{Z}_{-j}) \triangleq \prod_{i=1}^n PNS_{\mathbf{Z}=\mathbf{z}_i, Y=y_i \,|\, \mathbf{Z}_{-j}=\mathbf{z}_{i,-j}}. \tag{6}$$

Conditional efficiency and non-spuriousness describes how the label responds to the jth feature captured by the representation, holding all other features fixed. Its definition resembles the definition of (unconditional) efficiency and non-spuriousness. The only difference is that the conditional notion contrasts the counterfactual label of $\{Z_j \neq z_{ij}, \mathbf{Z}_{-j} = \mathbf{z}_{i,-j}\}$ and $\{Z_j = z_{ij}, \mathbf{Z}_{-j} = \mathbf{z}_{i,-j}\}$, while the unconditional notion contrasts that of $\{Z_j \neq z_{ij}, \mathbf{Z}_{-j} \neq \mathbf{z}_{i,-j}\}$ and $\{Z_j = z_{ij}, \mathbf{Z}_{-j} = \mathbf{z}_{i,-j}\}$.

The conditional efficiency and non-spuriousness across all dimensions of a representation is generally a stronger requirement than the (unconditional) efficiency and non-spuriousness. For example, suppose a dataset of $(\mathbf{X}, Y)$ with $\mathbf{X} = (X_1, X_2, X_3)$ is generated by $Y = X_1 + \epsilon_y$. Now consider a two-dimensional representation, $\mathbf{Z} = (Z_1, Z_2)$, where $Z_1 = X_1$ and $Z_2 = Z_1 + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$. When σ^2 is small, Z_1 and Z_2 are highly correlated. In this case, the (unconditional) efficiency and non-spuriousness of $\mathbf{Z}$ is high because (Z_1, Z_2) is indeed non-spurious and (nearly) efficient; Z_2 only introduces a negligible amount of useless information. However, the conditional efficiency and non-spuriousness of Z_2 given Z_1 is low because Z_2 is completely useless given Z_1.

We will leverage this conditional efficiency and non-spuriousness notion for representation learning in Section 2.4. First, however, we study how to measure (conditional) efficiency and non-spuriousness from observational data.

2.2 Identifying Efficiency and Non-spuriousness in Observational Datasets: Identification

Definitions 1 to 3 formalize the efficiency and non-spuriousness of representations using POC. These POC are counterfactual quantities; calculating them requires access to a causal structural equation (i.e., the true data generating process), which is rarely available in practice.

In this section, we study the observable implications of Definitions 1 to 3. They lead to strategies to evaluate the efficiency and non-spuriousness of representations with observational data; these strategies are known as *causal identification* strategies (Pearl, 2011). For simplicity of exposition, we focus on identifying the PNS in Definition 3. (Identification formulas for PN and PS in Definitions 1 and 2 can be similarly derived using Theorem 9.2.15 of Pearl (2011).)

To identify the counterfactual quantity PNS from observational data, we perform two steps of reduction, climbing down the ladder of causation (Pearl, 2019a): (1) connect PNS to interventional distributions, and (2) identify these interventional distributions from observational data. We describe these steps in detail in the next sections. The end product is Theorem 5, which provides an algorithm for calculating a lower bound on the PNS.

2.2.1 From counterfactuals to interventional distributions

To connect the counterfactual quantity PNS with intervention distributions, we generalize the classical identification results for PNS that have been developed for univariate binary causes and outcomes. It turns out that the PNS cannot be point-identified by interventional distributions; for each set of interventional distributions, there may exist many values of PNS that are consistent with the interventional distributions (Pearl, 2011). However, PNS can be bounded by the difference between these two interventional distributions.

Lemma 1 (A lower bound on PNS)**.** *Assuming the causal graph in Figure 2, the PNS is lower bounded by the difference between two intervention distributions:*

$$\begin{aligned} PNS_{\mathbf{Z}=\mathbf{z},Y=y} &= P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})\neq y) \\ &\geq P(Y=y\,|\,\mathrm{do}(\mathbf{Z}=\mathbf{z})) - P(Y=y\,|\,\mathrm{do}(\mathbf{Z}\neq\mathbf{z})). \end{aligned} \tag{7}$$

The inequality becomes an equality when the outcome Y is monotone in the representation $\mathbf{Z}$ (in the binary sense); i.e., $P(Y(\mathbf{Z}=\mathbf{z})\neq y, Y(\mathbf{Z}\neq\mathbf{z})=y)=0$.

Theorem 1 generalizes Theorem 9.2.10 of Pearl (2011) to non-binary $\mathbf{Z}$. (The proof is in Appendix A.) It connects the counterfactual quantity PNS to the intervention distribution $P(Y\,|\,\mathrm{do}(\mathbf{Z}=\mathbf{z}))$. An upper bound on PNS can be similarly obtained by generalizing the upper bound in Theorem 9.2.10 of Pearl (2011). We focus on PNS here because a larger lower bound on PNS implies a large PNS. In contrast, a larger upper bound on PNS does not necessarily imply a large PNS.

Reducing the counterfactual quantity PNS to the lower bound in Theorem 1 has climbed down one level of the ladder of causation (Pearl, 2019a); we have converted a counterfactual quantity (level three) to one with interventional distributions (level two). Next we will descend one level further, discussing how to identify the interventional distributions $P(Y\,|\,\mathrm{do}(\mathbf{Z}=\mathbf{z}))$ from the observational data distribution $P(\mathbf{X}, Y)$ (level one); $\mathbf{Z} = f(\mathbf{X})$ is the representation of $\mathbf{X}$ via some known function f. These identification results would also enable the identification of $P(Y \mid \mathrm{do}(\mathbf{Z} \neq \mathbf{z}))$:

$$P(Y \mid \mathrm{do}(\mathbf{Z} \neq \mathbf{z})) = \int P(Y \mid \mathrm{do}(\mathbf{Z}))P(\mathbf{Z} \mid \mathbf{Z} \neq \mathbf{z})\,\mathrm{d}\mathbf{Z} = \int P(Y \mid \mathrm{do}(f(\mathbf{X})))P(f(\mathbf{X}) \mid f(\mathbf{X}) \neq \mathbf{z})\,\mathrm{d}f(\mathbf{X}).$$

The functional intervention $\mathrm{do}(f(\mathbf{X}))$ and its counterfactuals $Y(f(\mathbf{X}) = z)$. We conclude this section by laying out the formal definitions of counterfactuals under functional interventions $Y(f(\mathbf{X}) = z)$.

We first write down the structural causal model M that governs the variable $\mathbf{C}, \mathbf{X} = (X_1, \ldots, X_m), Y$ in Figure 2:

$$\begin{aligned} c &= u_c, && (8)\\ x_j &= h_{x_j}(c, u_{x_j}), j = 1, \ldots, m, && (9)\\ y &= h_y(x_1, \ldots, x_m, u_y), && (10) \end{aligned}$$

where $\mathbf{U} = (u_c, u_y, \{u_{x_j}\}_{j=1}^m)$ are the set of background exogenous variables. We further consider a probability distribution $P(\mathbf{U})$ defined over the domain of $\mathbf{U}$.

The counterfactual $Y(f(\mathbf{X}) = z)$ is thus the potential response of Y to action $\mathrm{do}(f(\mathbf{X}) = z)$, following the definition of counterfactuals in Pearl (2011) (Definition 7.1.1–5). This functional intervention on $f(\mathbf{X})$ is a soft intervention on $\mathbf{X}$ (Correa & Bareinboim, 2020; Eberhardt & Scheines, 2007; Puli et al., 2020) conditional on all its parents $\mathbf{C}$, with a stochastic policy $p(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z})$ (Pearl, 2011, Ch. 4.2). In other words, a functional intervention $\mathrm{do}(f(\mathbf{X}) = \mathbf{z})$ considers all interventions on $\mathbf{X}$ that are consistent with the functional constraint $f(\mathbf{X}) = \mathbf{z}$ and its parental structure $\mathbf{C}$.

Following Section 3 of Correa & Bareinboim (2020), we define the counterfactual $Y(f(\mathbf{X}) = z)$ as the solution for Y of the set of equations,

$$\begin{aligned} c &= u_c, && (11)\\ x_j &= \tilde{h}_{x_j}(c, u_{x_j}), j = 1, \ldots, m, && (12)\\ y &= h_y(x_1, \ldots, x_m, u_y). && (13) \end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_m)$ follows a pre-specified distribution under $\tilde{h}_{x_j}(\cdot)$ in Equation (12),

$$\mathbf{X} \sim P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = z). \tag{14}$$

The distribution $P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = z)$ is derived using Bayes rule from the conditional $P(\mathbf{X} \mid \mathbf{C})$ in the original structural causal model (Equations (8) to (10)):

$$P(\mathbf{X} = \mathbf{x}^* \mid \mathbf{C}) = \sum_{\mathbf{U}: \{h_{x_j}(c, u_{x_j}) = x_j^*\}_{j=1}^m} P(\mathbf{U}) \qquad \forall \mathbf{x}^*. \tag{15}$$

These equations (Equations (11) to (14)) constitute the structural causal model $M_{f(\mathbf{X})}$.

Comparing functional interventions with *do* interventions. The functional intervention action $\mathrm{do}(f(\mathbf{X}) = z)$ is different from the usual do intervention $\mathrm{do}(\mathbf{X} = \mathbf{x}^*)$ with $\mathbf{x}^* = (x_1^*, \ldots, x_m^*)$, where one would replace Equation (9) with $x_j = x_j^*$. Rather, it is more akin to a stochastic intervention when we replace the function $h_{x_j}(\cdot)$ in Equation (9) with $\tilde{h}_{x_j}(\cdot)$ that induces a pre-specified distribution

$P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = z)$. We note that the constrained distribution $P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = z)$ is introduced only after the original causal model M ((Equations (8) to (10))) is defined. That said, it can be derived, following Equation (15), from the distribution $P(\mathbf{U})$ of the exogenous variables in the original model.

Moreover, the probabilities PN, PS, and PNS with functional interventions include randomness both from inference about the individual-level random noise variables in the SCM and the selection of the functional intervention. For example, the classical PNS would only involve the exogenous variable in the y-equation (Equation (10)),

$$P(Y(\mathbf{X} = \mathbf{x}) = y, Y(\mathbf{X} \neq \mathbf{x}) = y) = \sum_{u_y : Y(\mathbf{X}=\mathbf{x})=y, Y(\mathbf{X}\neq\mathbf{x})=y} P(u_y), \tag{16}$$

but the PNS with functional interventions would involve the exogenous variables in both the x-equation (Equation (12)) and the y-equation (Equation (13)),

$$\begin{aligned} &P(Y(f(\mathbf{X}) = z) = y, Y(f(\mathbf{X}) \neq z) \neq y) \\ &\qquad = \sum_{u_y, \{u_{x_j}\}_{j=1}^m : Y(f(\mathbf{X})=z)=y, Y(f(\mathbf{X})\neq z)=y} P(u_y, \{u_{x_j}\}_{j=1}^m). \end{aligned} \tag{17}$$

Functional intervention counterfactuals and out-of-distribution generalization. As a consequence of this additional dependence on x's exogenous variable, $Y(f(\mathbf{X}) = z)$ does not necessarily enjoy the same invariance properties that $Y(\mathbf{X} = x)$ enjoys. Specifically, while the classical counterfactual $P(Y(\mathbf{X} = x))$ is invariant to any changes to $P(\mathbf{X})$, the functional intervention counterfactual $P(Y(f(\mathbf{X}) = z))$ is not; the definition of the latter relies on $P(\mathbf{X} \mid \mathbf{C})$ as in Equation (14). Therefore, the counterfactual nature of $P(Y(f(\mathbf{X}) = z))$ does not necessarily improve out-of-generalization performance under covariate shift. Whether it does would rely on the nature of the distribution shift and the underlying causal structure of the data; for example, one would require $\int P(\mathbf{X} \mid f(\mathbf{X}) = z, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}\mathbf{C}$ to stay invariant to achieve such improvements.

For the same reason, the notions of necessity (efficiency) and sufficiency (non-spuriousness) of representations is also dependent on the population. Specifically, in the running example, the probabilities of causation for the feature of "dog face" depends on the distribution of images without dog face. Suppose that dog images without dog face predominantly have the dog face turned away in the dataset. Human labelers would still label the image as the dog. Then dog face would not be a necessary feature of the dog label. In contrast, if the images without dog face are all labeled non-dog, then dog face would be a necessary feature of the dog label.

Finally, we note that this non-invariance to changes in $P(\mathbf{X})$ is not specific to functional interventions or counterfactuals; it also appears in many other causal quantities that only involve classical *do* interventions. For example, the intervention distribution $P(Y \mid \mathrm{do}(X_1 = x_1))$ in Figure 2 has a similar dependence on $P(\mathbf{X})$:

$$P(Y \mid \mathrm{do}(X_1 = x_1)) = \int P(Y \mid X_1, X_{2:m}) P(X_{2:m}) \,\mathrm{d}X_{2:m}, \tag{18}$$

following back-door adjustment. Similar to $P(Y(f(\mathbf{X}) = z))$, this intervention distribution $P(Y \mid \mathrm{do}(X_1 = x_1))$ is thus not necessarily invariant to covariate shift and may not always improve out-of-generalization performance.

In more detail, this dependence of *do* intervention distributions (and, as a consequence, population quantities like average treatment effects) on the population $P(\mathbf{X})$ is of the same nature as that of functional intervention counterfactuals. Suppose we are interested in the interventional outcome distribution under some medical treatments. Further, both the patient's age and treatment status affect the patient's outcome. Then, the intervention distribution of the patient's outcome under treatment would depend on the age distribution of the population.

2.2.2 From interventional distributions to observational data distributions

To calculate the lower bound of PNS in Equation (7), we need to identify the intervention distribution $P(Y \mid \mathrm{do}(\mathbf{Z} = \mathbf{z}))$—equivalently $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$—from observational datasets $(\mathbf{X}, Y)$. Below we first state the formal definition of functional interventions $\mathrm{do}(f(\mathbf{X}) = \mathbf{z})$. We then discuss the challenges in identifying $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$ for high-dimensional $\mathbf{X}$. To tackle this challenge, we further provide an identification strategy and discuss its theoretical and practical requirements.

Identification of functional intervention distributions. We begin with defining the functional intervention distribution $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$, following the definition of the counterfactual $Y(f(\mathbf{X}) = z)$.

Definition 4 (Functional intervention distribution (Puli et al., 2020))**.** *The intervention distribution under a functional intervention $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$ is defined as*

$$P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z})) \triangleq \int P(Y \mid \mathrm{do}(\mathbf{X}), \mathbf{C}) P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z}) P(\mathbf{C}) \, \mathrm{d}\mathbf{X} \, \mathrm{d}\mathbf{C}, \tag{19}$$

where $\mathbf{C}$ denotes all parents of $\mathbf{X}$.

Following this definition, one can write the intervention distribution of interest, $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$, as follows:

$$P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z})) = \int P(Y \mid \mathbf{X}) \cdot \left[\int P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z}) P(\mathbf{C}) \, \mathrm{d}\mathbf{C} \right] \mathrm{d}\mathbf{X}. \tag{20}$$

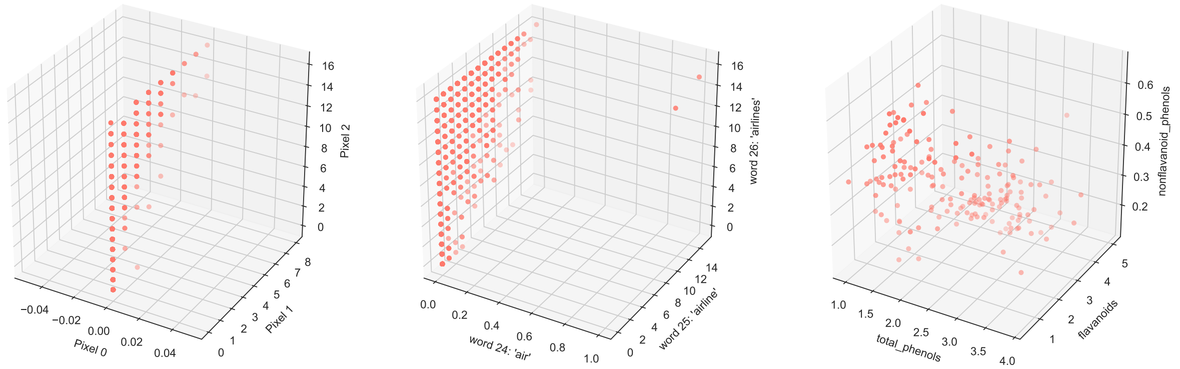
This equality is due to the SCM in Figure 2: there is no unobserved confounding between $\mathbf{X}$ and Y, which implies $P(Y \mid \mathrm{do}(\mathbf{X}), \mathbf{C}) = P(Y \mid \mathbf{X})$.

Functional intervention distributions recover the standard backdoor adjustment as special cases if we take the function f to be an identity function $f(\mathbf{X}) = \mathbf{X}$ or one that returns a subset $f(\mathbf{X}) = \mathbf{X}_S, S \subset \{1, \ldots, m\}$; see Appendix B for detailed derivations.

Equation (20) provides a way to identify $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$, provided that one can calculate $P(Y \mid \mathbf{X})$ and $\int P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z}) P(\mathbf{C}) \, \mathrm{d}\mathbf{C}$. At first sight, both quantities seem easy to calculate: one can estimate $P(Y \mid \mathbf{X})$ from observational sampling of $P(Y, \mathbf{X})$. To calculate $\int P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z}) P(\mathbf{C}) \, \mathrm{d}\mathbf{C}$, one may leverage probabilistic factor models (e.g., probabilistic principal component analysis (PPCA), Gaussian mixture model (GMM), or variational autoencoder (VAE)) because the latent $\mathbf{C}$ renders $X_1, \ldots, X_m$ conditionally independent in Figure 2. One can often read off $P(\mathbf{X} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z})$ and $P(\mathbf{C})$ from the probabilistic factor model fit, if the factor model is identifiable.

However, $P(Y \mid \mathbf{X})$ turns out to be challenging to calculate in practice, especially when $\mathbf{X}$ represents high-dimensional objects such as images or text, as we now discuss.

The challenge of identifying $P(Y \mid \mathbf{X})$ with high-dimensional $\mathbf{X}$. The key challenge in identifying $P(Y \mid \mathbf{X})$ lies in the high-dimensionality of $\mathbf{X}$. High-dimensional objects (e.g., images or



(a) High-dim. image data: MNIST (Deng, 2012)

(b) High-dim. text data: Airline tweets

(c) Low-dim. data: Wine features

Figure 3: High-dimensional data[5] such as images (Figure 3a) or text (Figure 3b) are often only supported on a low-dimensional manifold. Low-dimensional data, e.g., the wine dataset (Figure 3c), is often supported in the whole space.

text) are often only supported on a low-dimensional manifold. For example, it may be the case that $X_j = g_0(\{X_1, \ldots, X_m\}\backslash X_j)$ for some $j \in \{1, \ldots, m\}$ and some function g; see Figure 3 for examples. We refer to this problem as the *rank-degeneracy problem* (Stewart, 1984; Golub et al., 1976). It is akin to the classical challenge of having highly correlated variables in a linear regression; it is also an example of the underspecification problem in deep learning (D'Amour et al., 2020b).

The rank degeneracy of high-dimensional $\mathbf{X} = (X_1, \ldots, X_m)$ challenges the identification of $P(Y \,|\, \mathbf{X})$, even if both $\mathbf{X}$ and Y are observed. Intuitively, this non-identifiability problem occurs because the data can only inform $P(Y \,|\, \mathbf{X})$ on the low-dimensional manifold in $\mathcal{X}^m$ where $P(\mathbf{X})$ is supported. The behavior of $P(Y \,|\, \mathbf{X})$ outside this manifold is unconstrained, hence non-identifiable.[4]

In more detail, suppose $p(\mathbf{x}, y)$, $p(\mathbf{x})$, $p(y \,|\, \mathbf{x})$ denote the relevant densities (and assume they exist). Any $h_0(\cdot, \cdot)$ function that satisfies

$$h_0(\mathbf{x}, y) \cdot p(\mathbf{x}) = p(\mathbf{x}, y) \qquad \forall \mathbf{x} \in \mathcal{X}^m \tag{21}$$

is a valid density for $p(y \,|\, \mathbf{x})$. Under rank degeneracy, it turns out that there exist many different functions $h_0(\mathbf{x}, y)$ that satisfy Equation (21). Hence $p(y \,|\, \mathbf{x})$ is non-identifiable. The reason is that rank degeneracy implies $p(\mathbf{x}) = 0$ for $\mathbf{x} \in \widetilde{\mathcal{X}} \subseteq \mathcal{X}^m$, where $\widetilde{\mathcal{X}}$ is a set with positive measure. If some function h_0' satisfies Equation (21), then the function h_0'' would also satisfy Equation (21) if they differ only on the set $(\mathbf{x}, y) \in \widetilde{\mathcal{X}} \times \mathcal{Y}$. Note that h_0' and h_0'' are not almost surely equal; they differ on a set $\widetilde{\mathcal{X}}$ with positive measure. Thus, h_0' and h_0'' are two different densities both valid for $p(y \,|\, \mathbf{x})$, which implies the non-identifiability of $p(y \,|\, \mathbf{x})$.

As a more concrete example, consider a high-dimensional vector of image pixels $\mathbf{X}$ that lives on a low-dimensional manifold; i.e., such that $X_j - g_0(\{X_1, \ldots, X_m\}\backslash X_j)$ is identically zero in the observational data (Kingma & Welling, 2014; Goodfellow et al., 2014). This rank degeneracy implies

[4] This non-identifiability of $P(Y \,|\, \mathbf{X})$ is also why, for high-dimensional $\mathbf{X}$, directly fitting a neural network for $P(Y \,|\, \mathbf{X})$ can pick up spurious correlations and fail in out-of-distribution prediction. Similar failure can also occur with linear regression when $\mathbf{X}$ is approximately low rank; e.g., when different dimensions of $\mathbf{X}$ are highly correlated.

[5] https://www.kaggle.com/crowdflower/twitter-airline-sentiment, https://archive.ics.uci.edu/ml/datasets/wine

that for any $p(y \mid \mathbf{x}) = h_0(\mathbf{x}, y)$ compatible with the observational data distribution, the conditional $p(y \mid \mathbf{x}) = h_0(\mathbf{x}, y) + \alpha \cdot (x_j - g_0(\{x_1, \ldots, x_m\} \backslash x_j)), \forall \alpha \in \mathbb{R}$, is also compatible with the observational data.

This fundamental non-identifiability of $P(Y \mid \mathbf{X})$ with high-dimensional $\mathbf{X}$ suggests that we cannot hope to identify functional interventions $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$ where the function f non-trivially depends on all $X_1, \ldots, X_m$, or any of its subsets that is also rank degenerate.

Causal identification of $P(Y \mid \mathrm{do}(f(\mathbf{X}))$ for a restricted set of f. Given the fundamental non-identifiability of $P(Y \mid \mathbf{X})$ with high-dimensional $\mathbf{X} = (X_1, \ldots, X_m)$, we restrict our attention to representations that only nontrivially depends on a "full-rank" subset; i.e., $\mathbf{Z} = f(\mathbf{X}) = \tilde{f}((X_j)_{j \in S})$, for some function $\tilde{f} : \mathcal{X}^{|S|} \to \mathbb{R}^d$, and a set $S \subseteq \{1, \ldots, m\}$, where $p((x_j)_{j \in S}) > 0$ for all values $(x_j)_{j \in S} \in \mathcal{X}^{|S|}$. We term this requirement "observability."

Focusing on such representations $f(\mathbf{X}) = \tilde{f}((X_j)_{j \in S})$, we calculate its intervention distributions by returning to the definition of functional interventions (Definition 4),

$$P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z})) = \int P(Y \mid (X_j)_{j \in S}, \mathbf{C}) P((X_j)_{j \in S} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z}) P(\mathbf{C}) \,\mathrm{d}(X_j)_{j \in S} \,\mathrm{d}\mathbf{C}. \quad (22)$$

Equation (22) implies that $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$ is identifiable as long as one can identify the unobserved common cause $\mathbf{C}$ from data. This condition is often called the "pinpointability" (or "effective observability") of $\mathbf{C}$ (Wang & Blei, 2020, 2019a). The following proposition summarizes the identification of $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$ under these conditions.

Proposition 1 (Identification of $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$. *Assume the causal graph in Figure 2. Suppose the representation only effectively depends on a subset $(X_j)_{j \in S}$ of $(X_1, \ldots, X_m)$; i.e., $f(\mathbf{X}) = \tilde{f}((X_j)_{j \in S})$ for some function $\tilde{f} : \mathcal{X}^{|S|} \to \mathbb{R}^d$ and some set $S \subseteq \{1, \ldots, m\}$. Then the intervention distribution $P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z}))$ is identifiable by*

$$P(Y \mid \mathrm{do}(f(\mathbf{X}) = \mathbf{z})) = \int P(Y \mid f(\mathbf{X}) = \mathbf{z}, h(\mathbf{X})) \cdot P(h(\mathbf{X})) \,\mathrm{d}h(\mathbf{X}), \quad (23)$$

if the following conditions are satisfied:

1. *(pinpointability) the unobserved common cause $\mathbf{C}$ is pinpointable; i.e., $P(\mathbf{C} \mid \mathbf{X}) = \delta_{h(\mathbf{X})}$ for a deterministic function h known up to bijective transformations,*
2. *(positivity) $(X_j)_{j \in S}$ satisfies the positivity condition given $\mathbf{C}$; i.e., $P((X_j)_{j \in S} \in \widetilde{\mathcal{X}} \mid \mathbf{C}) > 0$ for any set $\widetilde{\mathcal{X}} \subset \mathcal{X}^{|S|}$ such that $P((X_j)_{j \in S} \in \widetilde{\mathcal{X}}) > 0$,*
3. *(observability) $P((X_j)_{j \in S} \in \widetilde{\mathcal{X}}) > 0$ for all subsets $\widetilde{\mathcal{X}} \subset \mathcal{X}^{|S|}$ with a positive measure.*

Proposition 1 describes a set of representations $\mathbf{Z} = f(\mathbf{X})$ for which can we evaluate their intervention distribution, and hence the lower bound of their PNS for Y. (The proof is in Appendix C.)

Proposition 1 is based on three conditions. The pinpointability ensures that the terms in Equation (22) involving $\mathbf{C}$ (e.g., $P(\mathbf{C}), P((X_j)_{j \in S} \mid \mathbf{C}, f(\mathbf{X}) = \mathbf{z})$) are estimable from observational data. We note that pinpointability serves a different purpose here than in Wang & Blei (2019a, 2020) and Puli et al. (2020). We invoke pinpointability for Proposition 1 because functional interventions require the knowledge of $\mathbf{C}$; there is no unobserved confounding issue. In contrast, Wang & Blei (2019a, 2020) and Puli et al. (2020) invoke pinpointability to handle unobserved confounders.

Pinpointability is often approximately satisfied when the high-dimensional $\mathbf{X}$ is driven by a low-dimensional factor $\mathbf{C}$. In this case, $P(\mathbf{C} \mid \mathbf{X})$ is often close to a point mass (Chen et al., 2020b). For example, when the m-dimensional data $\mathbf{X}$ is generated by a known d-dimensional probabilistic factor model, then the latent factor $\mathbf{C}$ is approximately pinpointable when $m \gg d$. In these probabilistic factor models, $P(\mathbf{C} \mid \mathbf{X})$ becomes increasingly close to a point mass as $m \to \infty$ but d fixed (Chen et al., 2020b; Bai & Li, 2016; Wang & Blei, 2019b). In practice, one can assess pinpointability by fitting probabilistic factor models to $\mathbf{X}$. We will discuss these practical details in the next section.

The second condition of Proposition 1 is the positivity of $(X_j)_{j\in S}$ given $\mathbf{C} = h(\mathbf{X})$, also known as the overlap condition (Imbens & Rubin, 2015). This condition ensures that $P(Y \mid (X_j)_{j\in S}, \mathbf{C})$ in Equation (22) is estimable from observational data. Positivity loosely requires that all values of $(X_j)_{j\in S}$ are possible conditional on $\mathbf{C} = h(\mathbf{X})$. For example, it is violated when $S = \{1, \ldots, n\}$ because it is impossible to observe $\mathbf{X} = \mathbf{x}$ and $\mathbf{C} = \mathbf{c}$ simultaneously when $h(\mathbf{x}) \neq \mathbf{c}$. In practice, positivity is more likely to be satisfied when $(X_j)_{j\in S}$ is a low-dimensional vector; e.g., the image representation only depends on a few image pixels that are far away from each other.

Finally, the observability condition requires that it must be possible to observe all possible values of $(X_j)_{j\in S}$. Together with the positivity condition, this implies that $P((X_j)_{j\in S} \mid \mathbf{C}) > 0$ for all $(X_j)_{j\in S}$ and, therefore, $P(Y \mid (X_j)_{j\in S}, \mathbf{C})$ in Equation (23) is estimable from observational data. This condition is violated if the observed $(X_j)_{j\in S}$ is rank-degenerate; e.g., $(X_j)_{j\in S}$ with $S = \{1, \ldots, m\}$ represents the whole high-dimensional image vector supported only on a low-dimensional manifold. This rank degeneracy would render $P(Y \mid (X_j)_{j\in S}, C)$ non-identifiable, for the same reason as why $P(Y \mid \mathbf{X})$ is non-identifiable with rank-degenerate $\mathbf{X}$. In practice, this observability condition is more likely to hold when $(X_j)_{j\in S}$ is low-dimensional.

Proposition 1 describes a class of representations $\mathbf{Z} = f(\mathbf{X})$ whose intervention distributions are identifiable from observational data. Combining Proposition 1 and Theorem 1 allows us to lower bound the PNS of these representations and evaluate their efficiency and non-spuriousness.

Theorem 5 (Evaluating efficiency and non-spuriousness with observational data)**.** *Under the assumptions of Proposition 1, the efficiency and non-spuriousness of the representation $\mathbf{Z} = f(\mathbf{X}) = \tilde{f}((X_j)_{j\in S})$ in the dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is lower bounded by*

$$\begin{aligned} PNS_n(f(\mathbf{X}), Y) &\geq \underline{PNS_n(f(\mathbf{X}), Y)} \\ &\triangleq \prod_{i=1}^n \int [P(Y = y_i \mid f(\mathbf{X}) = f(\mathbf{x}_i), \mathbf{C}) \\ &\qquad - P(Y = y_i \mid f(\mathbf{X}) \neq f(\mathbf{x}_i), \mathbf{C})] \cdot P(\mathbf{C}) \, \mathrm{d}\mathbf{C}, \end{aligned} \tag{24}$$

where $\mathbf{C} = h(\mathbf{X})$ is the unobserved common cause pinpointed by the observational data $\mathbf{X}$ in Proposition 1. Similarly, the conditional efficiency and non-spuriousness of the jth dimension of $(f_1(\mathbf{X}), \ldots, f_d(\mathbf{X}))$ is lower bounded by

$$\begin{aligned} PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X})) &\geq \underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))} \\ &\triangleq \prod_{i=1}^n \int [P(Y = y_i \mid f_j(\mathbf{X}) = f_j(\mathbf{x}_i), f_{-j}(\mathbf{X}) = f_{-j}(\mathbf{x}_i), \mathbf{C}) \\ &\qquad - P(Y = y_i \mid f_j(\mathbf{X}) \neq f_j(\mathbf{x}_i), f_{-j}(\mathbf{X}) = f_{-j}(\mathbf{x}_i), \mathbf{C})] \cdot P(\mathbf{C}) \, \mathrm{d}\mathbf{C}. \end{aligned} \tag{25}$$

Theorem 5 is an immediate consequence of Proposition 1 and theorem 1. It suggests that we evaluate efficiency and non-spuriousness using Equation (24) and Equation (25) in practice. We operationalize this result in the next section. We conclude this section with two clarifications.

Why is any causal adjustment needed? A reader might ask: There is no confounding in Figure 2. Why do we need the causal adjustment that pinpoints the latent common causes and adjusts for it? If Y were a binary label, the one-dimensional representation $\mathbb{E}\left[Y \mid \mathbf{X}\right]$ would be the minimal sufficient statistic for the Bayes optimal classifier, and thus the optimal necessary and sufficient representation.

The reason lies in the challenge of identifying $\mathbb{E}\left[Y \mid \mathbf{X}\right]$ (or more generally $P(Y \mid \mathbf{X})$) in practice. For example, when $\mathbf{X}$ is an image that often lives on a low-dimensional manifold, one cannot estimate $\mathbb{E}\left[Y \mid \mathbf{X}\right]$ from data for all $\mathbf{X}$. It is because we never observe any $\mathbf{X}$ that is off the manifold, and hence $\mathbb{E}\left[Y \mid \mathbf{X}\right]$ is not identifiable. Many functions of $\mathbf{X}$ would fit the observed data equally well, even if they differ in their values for off-manifold $\mathbf{X}$. But only some of them would capture the optimal necessary and sufficient representation.

For this reason, we have to place additional constraints on the allowable representations to circumvent the need to identify $\mathbb{E}\left[Y \mid \mathbf{X}\right]$. Specifically, we considered a restricted set of f in $P(Y \mid \mathrm{do}(f(\mathbf{X})))$ that only non-trivially depends on a "full-rank" subset $S \subset \{1, \ldots, m\}$. This restriction makes $E[Y \mid X_S]$ identifiable, and so does $P(Y \mid \mathrm{do}(f(\mathbf{X}))$ (under conditions in Proposition 1).

Would pinpointing C prevent representation learning? Another common concern surrounds pinpointing and adjusting for $\mathbf{C}$ in CAUSAL-REP. In particular, much of representation learning is motivated by trying to recover latent variables like $\mathbf{C}$, but the identification assumptions of CAUSAL-REP preclude representation learning from picking up information in $\mathbf{C}$.

We clarify this concern by first discussing why pinpointing $\mathbf{C}$ was needed. The intention of invoking C is to handle the non-identifiability of $P(Y \mid \mathbf{X})$ when X is high-dimensional and nearly degenerately lives on a low-dimensional manifold. In other words, the different components of $\mathbf{X}$ are extremely highly correlated and redundant of each other. In such cases, pinpointing C would separate out the shared redundant component in $\mathbf{X}$ from the exogeneous non-redundant component in $\mathbf{X}$, transforming $\mathbf{X}$ into a lower dimensional vector that does not suffer from this near-degeneracy.

Consider a toy example for illustration. Suppose $\mathbf{X} = (X_1, \ldots, X_m)$ is a set of high-dimensional covariates that satisfies

$$X_j = X_1 + \epsilon_j, \qquad j = 2, \ldots, m,$$

where $\epsilon_{2:m} \perp X_1$ and $\mathrm{Var}(\epsilon_j) \ll \mathrm{Var}(X_1)$. This is an example where the m-dimensional vector $\mathbf{X}$ is nearly degenerate, approximately living on a one-dimensional manifold governed by X_1. In this case, setting $\mathbf{C} = X_1$ would render $X_1, \ldots, X_m$ conditionally independent, as the causal graph in Figure 2. Hence, the factor modeling step of CAUSAL-REP would absorb the shared redundant component X_1 into $\mathbf{C}$ and separate out the residual exogenous components of $\mathbf{X}$ conditional on $\mathbf{C}$, namely $\epsilon_{2:m}$. Moreover, these exogenous components $\epsilon_{2:m}$ are no longer nearly degenerate as $X_{2:m}$. The representation learning algorithm then learns a function of these exogenous components $\epsilon_{2:m}$ to capture necessary and sufficient features.

This example shows that pinpointing $\mathbf{C}$ is intended to handle the near-degeneracy of $\mathbf{X}$ by identifying and conditioning on the redundant components in $\mathbf{X}$. Bringing this intuition to image examples, $\mathbf{C}$ will likely capture the redundancy among pixels. If two pixels always take highly correlated values across images, then $\mathbf{C}$ would absorb one of them. For example, $\mathbf{C}$ would likely capture the redundancy among the pixels that are near the image boundary; they often take highly correlated values to render the background color and texture of the image. That said, for center pixels of images that often render different objects, they may not be as highly correlated as boundary pixels. $\mathbf{C}$ may still capture the redundant component among these center pixels but the redundancy is minimal; much variations (and co-variations) among these pixel will remain intact conditional on $\mathbf{C}$, enabling representation learning of necessary and sufficient features.

Following this reasoning, it is indeed true that CAUSAL-REP may not be able to find *all* necessary and sufficient features. If a necessary and sufficient feature is also the source of high correlations among certain pixels, then CAUSAL-REP will absorb it in $\mathbf{C}$ and preclude it from being picked up by $f(\mathbf{X})$. In this sense, CAUSAL-REP can only find *some* features that have high necessity and sufficiency; it will miss out many necessary and sufficient features if they also drives the correlations among pixels, e.g. background features or texture features in images.

This reasoning also suggests when CAUSAL-REP will be sensitive to the pinpointing step. Empirically, we find that the learned necessary and sufficient representation can sometimes be sensitive to configurations of the pinpointing step (e.g., model choice) and the subsequent representation learning (e.g., regularization strengths). It most often happens when the necessary and sufficient representation entails features that are highly correlated with the pinpointed $\mathbf{C}$. Such features, for example, can include image features that are highly correlated with the background textures. Otherwise, CAUSAL-REP can be robust to the choices of models and their hyperparameters, as is illustrated in empirical studies (Figure 6b).

We finally note that pinpointing $\mathbf{C}$ is only one of the ways to address the near-degeneracy issue of $\mathbf{X}$. There exists other approaches to this problem that enables the identification of $P(Y \mid \mathbf{X})$. For example, one could consider finding a bijective map $l(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^{d_0}$ between the space of $\mathbf{X}$ and $\mathbb{R}^{d_0}$, where $d_0 < m$. One can then perform representation learning on $\tilde{\mathbf{X}} = l(\mathbf{X})$: it will no longer suffer from the near-degeneracy and $P(Y \mid l(\mathbf{X}))$ will be identifiable. We did not take this approach because finding such a bijection would require specific characterizations of the $\mathbf{X}$-space in different applications.

Take a further step back, pinpointing $\mathbf{C}$ and/or handling the near-degeneracy of $\mathbf{X}$ may not always be necessary in representation learning. In CAUSAL-REP, pinpointing $\mathbf{C}$ is only performed to handle the near-degeneracy of $\mathbf{X}$ in practice, which prevents the identification of $P(Y \mid \mathbf{X})$. If $\mathbf{X}$ does not suffer from this near-degeneracy in the data, then one does not need to invoke the full CAUSAL-REP algorithm (Algorithm 1). Rather, one can directly maximize Eq. (9) to find necessary and sufficient features without the need to pinpoint $\mathbf{C}$.

2.3 Measuring efficiency and non-spuriousness in practice: Estimation

We operationalize Theorem 5 to measure efficiency and non-spuriousness for a representation $f(\mathbf{X})$. The full algorithm is in Algorithm 1. The algorithm involves three steps: (1) Pinpoint the unobserved common cause $\mathbf{C}$; (2) Estimate the conditional $P(Y \mid f(\mathbf{X}), \mathbf{C})$; (3) Calculate the lower bounds in Theorem 5 (Equations (24) and (25)) to measure efficiency and non-spuriousness. We discuss each step below.

Pinpointing the unobserved common cause C. Theorem 5 requires that the unobserved common cause $\mathbf{C}$ must be pinpointable by the observational data: $P(\mathbf{C} \mid \mathbf{X}) = \delta_{h(\mathbf{X})}$ for some deterministic function h. Moreover, this pinpointing function h needs to be known up to bijective transformations. How can we assess pinpointability and extract h in practice?

One can assess pinpointability by fitting a probabilistic factor model, viewed as a causal generative model; e.g., probabilistic principal component analysis (PPCA) (Tipping & Bishop, 1999), variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014), mixture models (McLachlan &

Basford, 1988), or mixed membership models (Pritchard et al., 2000; Blei et al., 2003; Airoldi et al., 2008; Erosheva & Fienberg, 2005).

Specifically, suppose the dataset of n i.i.d. data points $(\mathbf{x}_i)_{i=1}^n$ is assumed to be generated by a PPCA model, then one can assess pinpointability by first fitting a PPCA model,

$$\mathbf{c}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_K), \qquad i = 1, \ldots, n, \tag{26}$$

$$x_{il} \,|\, \mathbf{c}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{c}_i^\top \theta_l, \sigma^2), \qquad l = 1, \ldots, m, \tag{27}$$

where $\mathbf{c}_i$ is the latent variable for each data point $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is the set of parameters. One then infers the parameters $\boldsymbol{\theta}$ and the posterior $p(\mathbf{c}_i \,|\, \mathbf{x}_i)$ for each i, using standard posterior inference algorithms such as variational inference (Blei et al., 2017) or Markov chain Monte Carlo methods (Gilks et al., 1995). Pinpointability then approximately holds if $p(\mathbf{c}_i \,|\, \mathbf{x}_i)$ is close to a point mass for all i's.

The choice of the latent dimensionality K is a challenging task. We choose K in practice by searching for a factor model that can both fit data well (e.g., via a posterior predictive check (Gelman et al., 1996)) and approximately satisfy pinpointability. A larger d tends to have a better fit to the data, while a smaller d is more likely to satisfy pinpointability. We proceed if pinpointability holds, along with the other conditions—observability and positivity—from Theorem 5.

Given a fitted factor model, one can extract the pinpointing function h by computing the posterior mean, $h(\mathbf{x}) \approx \mathbb{E}\left[\mathbf{c}_i \,|\, \mathbf{x}_i = \mathbf{x}\right]$. The same procedure applies if the data is generated by other probabilistic factor models.

A reader might ask: Why probabilistic factor models? The reason is that the causal graph in Figure 2 encodes a conditional independence structure that coincides with the defining feature of probabilistic factor models. In more detail, Figure 2 assumes that each dimension of the data $\mathbf{X} = (X_1, \ldots, X_m)$ is rendered conditionally independent given an unobserved common cause $\mathbf{C}$. This conditional independence is precisely the defining structure of probabilistic factor models,

$$\mathbf{c}_i \sim p(\cdot\,;\, \lambda_{\mathbf{C}}), \qquad i = 1, \ldots, n, \tag{28}$$

$$x_{il} \,|\, \mathbf{c}_i \sim p(\cdot \,|\, \mathbf{c}_i\,;\, \theta_l), \qquad l = 1, \ldots, m, \tag{29}$$

with i.i.d. data $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ and latent $\mathbf{c}_i$. We therefore use probabilistic factor models to assess the pinpointability of $\mathbf{C}$ in Figure 2.

Calculate the conditional $P(Y \,|\, f(\mathbf{X}), \mathbf{C})$. After pinpointing $\mathbf{C}$, we still need the conditional $P(Y \,|\, f(\mathbf{X}), \mathbf{C})$ to calculate the lower bound in Theorem 5.

To estimate $P(Y \,|\, f(\mathbf{X}), \mathbf{C})$, we can fit a model to the observational data $(\mathbf{x}_i, y_i)_{i=1}^n$, with $f(\cdot)$ being the representation function of interest and $\mathbf{c}_i = h(\mathbf{x}_i)$ estimated from the first step. As an example, we may posit a linear model,

$$P(Y \,|\, f(\mathbf{X}), \mathbf{C}) = \mathcal{N}((\beta_0 + \boldsymbol{\beta}^\top f(\mathbf{X}) + \boldsymbol{\gamma}^\top \mathbf{C}), \sigma^2), \tag{30}$$

and estimate the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)$ and $\boldsymbol{\gamma}$ via maximum likelihood. One may also posit other more flexible models like a partially linear model (with a nonlinear f_C) and target a categorical outcome:

$$P(Y \,|\, f(\mathbf{X}), \mathbf{C}) = \text{Categorial}(\text{softmax}(\beta_0 + \boldsymbol{\beta}^\top f(\mathbf{X}) + \gamma^\top f_C(\mathbf{C})). \tag{31}$$

Algorithm 1: Calculating the (lower bound of) efficiency and non-spuriousness of a representation

```
input  : The observational data and its label $\{\mathbf{x}_i, y_i\}_{i=1}^n$; representation function $f(\cdot)$
         ; the probabilistic factor model that generates the data $P(\mathbf{X}, \mathbf{C})$.
output : The (lower bound of) efficiency and non-spuriousness of representation $f(\mathbf{X})$ for label $Y$
Fit a probabilistic factor model (Equations (28) and (29)) and infer $p(\mathbf{c}_i \mid \mathbf{x}_i)$;
if Pinpointability holds, i.e., $p(\mathbf{c}_i \mid \mathbf{x}_i)$ is close to a point mass for all $i$ then
    if Observability and positivity (Theorem 5) hold then
        foreach datapoint $i$ do
            Pinpoint the unobserved common cause: $\hat{\mathbf{c}}_i = h(\mathbf{x}_i) \approx \mathbb{E}[\mathbf{c}_i \mid \mathbf{x}_i]$;
        end
        Calculate the conditional $P(Y \mid f(\mathbf{X}), \mathbf{C})$ by fitting a model to $\{\mathbf{x}_i, \hat{\mathbf{c}}_i, y_i\}_{i=1}^n$ (e.g., Equations (30)
          and (31));
    end
    Calculate the (lower bound of) efficiency and non-spuriousness $\underline{PNS_n(f(\mathbf{X}), Y)}$ and
      $\underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))}, j = 1, \dots, d$ (Equations (24) and (25));
end
```

Calculate the lower bounds in Theorem 5. We finally calculate the lower bounds in Equations (24) and (25) using the pinpointed $\mathbf{C}$ and the estimated conditional $P(Y \mid f(\mathbf{X}), \mathbf{C})$ from the previous two steps.

As an example, we calculate the lower bound of the conditional efficiency and non-spuriousness $\underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))}$ with the linear model (see Appendix E for the detailed derivation):

$$
\begin{aligned}
&\log \underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))} \\
&\approx \Bigg(\frac{1}{2\sigma^2} \Bigg[\sum_{i=1}^{n} (\beta_j \cdot (f_j(\mathbf{x}_i) - \mathbb{E}[f_j(\mathbf{x}_i)]))^2 \\
&\qquad + 2 \cdot \sum_{i=1}^{n} \beta_j \cdot (f_j(\mathbf{x}_i) - \mathbb{E}[f_j(\mathbf{x}_i)]) \cdot \gamma^\top (\mathbf{c}_i - \mathbb{E}[\mathbf{c}_i]) \Bigg] \Bigg) + \text{constant}.
\end{aligned} \tag{32}
$$

Similarly, we can obtain the lower bound of the (unconditional) efficiency and non-spuriousness,

$$
\begin{aligned}
&\log \underline{PNS_n(f(\mathbf{X}), Y)} \\
&\approx \Bigg(\frac{1}{2\sigma^2} \sum_{i=1}^{n} \Bigg[(\sum_{j=1}^{d} \beta_j \cdot (f_j(\mathbf{x}_i) - \mathbb{E}[f_j(\mathbf{x}_i)]))^2 \\
&\qquad + 2 \cdot \sum_{j=1}^{d} \beta_j \cdot (f_j(\mathbf{x}_i) - \mathbb{E}[f_j(\mathbf{x}_i)]) \cdot \gamma^\top (\mathbf{c}_i - \mathbb{E}[\mathbf{c}_i]) \Bigg] \Bigg) + \text{constant}.
\end{aligned} \tag{33}
$$

Equations (32) and (33) illustrate how conditional efficiency and non-spuriousness differ from the unconditional version. The conditional notion considers the $\beta_j \cdot f_j(\mathbf{X})$ one at a time. The unconditional notion lumps together all d dimensions of the representation and considers $\sum_{j=1}^{d} \beta_j \cdot f_j(\mathbf{X})$ as a whole. In this sense, the conditional PNS is finer-grained notion than the (unconditional) PNS.

A Linear example. We illustrate Algorithm 1 on a toy rank-degenerate image dataset. (We analyze more complex and higher-dimensional data in Section 2.5.) Imagine we collect a dataset of

$n = 1000$ images. Each image is characterized by its values on $m = 5$ chosen representative pixels, $\mathbf{X} = (X_1, \ldots, X_5)$, accompanied by a label about the image brightness Y. Both the pixel values $\mathbf{X}$ and the labels Y are real-valued.

We simulate such a dataset of image pixels and labels. As the pixel values are often highly correlated, we generate $\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})$ from a multivariate Gaussian distribution with strong (positive or negative) correlations—all pairwise correlations are > 0.8. As the brightness label only depends on a small number of pixels, we simulate y_i from a linear model that only uses two of the five pixels

$$y_i = \beta_1^* x_{i1} + \beta_2^* x_{i2} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{34}$$

where $\beta_1^* = 0.5, \beta_2^* = 1.0$.

The goal is to evaluate the efficiency and non-spuriousness of a representation $f(\mathbf{X}) = (X_2, 0.5 \cdot X_1 + X_4)$. We apply Algorithm 1: pinpoint the unobserved common cause $\mathbf{C}$, estimate $P(Y \mid f(\mathbf{X}), \mathbf{C})$, and finally calculate the lower bound of efficiency and non-spuriousness $\underline{PNS_n(f(\mathbf{X}), Y)}$ and $\underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))}, j = 1, \ldots, d$.

We first pinpoint $\mathbf{C}$. Suppose it is known that the data $\mathbf{X}$ is generated by a PPCA. We fit a one-dimensional PPCA (Equation (26)) to $\{\mathbf{x}_i\}_{i=1}^n$; the latent variable c_i is thus a scalar. The fit leads to the posterior of the unobserved common cause, $p(c_i \mid \mathbf{x}_i)$. We assess the pinpointability of $\mathbf{C}$ by calculating $\mathrm{Var}(c_i \mid \mathbf{x}_i)$ for all i. We find the variance is smaller than 0.01 for all i, implying that $p(c_i \mid \mathbf{x}_i)$ is fairly close to a point mass and pinpointability is approximately satisfied. (The threshold of 0.01 is a subjective choice.) We then calculate $\hat{c}_i = \mathbb{E}[c_i \mid \mathbf{x}_i]$ for all i.

We next estimate $P(Y \mid f(\mathbf{X}), \mathbf{C})$ by fitting a linear model to the dataset $\{(y_i, f(\mathbf{x}_i), \hat{c}_i)\}_{i=1}^n$ with $f(\mathbf{x}_i) = (x_{i2}, 0.5 \cdot x_{i1} + x_{i4})$,

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot (0.5 \cdot x_{i1} + x_{i4}) + \gamma \cdot \hat{c}_i, \sigma^2). \tag{35}$$

Fitting this linear model returns the regression coefficients $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}\}$.

Finally, plugging in these regression coefficients to Equations (32) and (33) gives the efficiency and non-spuriousness, with $f_1(\mathbf{x}_i) = x_{i2}$ and $f_2(\mathbf{x}_i) = 0.5 \cdot x_{i1} + x_{i4}$.

2.4 CAUSAL-REP: Learning Efficient and Non-spurious Representations

We have developed a strategy to evaluate the efficiency and non-spuriousness of a given representation in Section 2.2. Here we utilize this strategy to improve representation learning in both supervised and unsupervised settings.

2.4.1 Representation learning as finding necessary and sufficient causes

As representation learning requires efficient and non-spurious representations, we formulate representation learning as a task of *finding necessary and sufficient causes* of the label.

Representation learning as finding necessary and sufficient causes. To operationalize this formulation, we search for a representation that maximizes the *conditional* efficiency and non-spuriousness for a given dataset, following Definition 3 and Equation (6). Thus we view an ideal representation as one in which each dimension of the representation captures features that are essential and non-spurious given all other dimensions.

We perform representation learning by maximizing the sum of log PNS across all dimensions of the representation,

$$\max_{f} \sum_{j=1}^{d} \log PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X})), \tag{36}$$

where $PNS_n(\cdot, Y \mid \cdot)$ measures the conditional efficiency and non-spuriousness of $f_j(\mathbf{X})$ as in Equation (6).

Classes of representation functions. To perform this maximization over representations in practice, we consider parameterized classes of representation functions $f(\cdot)$. One option is to consider a class of neural network functions with a fixed architecture, subject to the constraint that each output dimension has zero mean and unit standard deviation, e.g.,

$$\begin{aligned}\{f : \text{multilayer perceptrons } \mathbb{R}^m \to \mathbb{R}^d &\text{ with two hidden layers of size 512} \\ &\text{s.t. } \{f(\mathbf{x}_i)\}_{i=1}^{n} \text{ has sample mean 0 and standard error 1}\}.\end{aligned} \tag{37}$$

Another option is to consider representations that are convex combinations of the m-dimensional data,

$$\{f : f(\mathbf{X}) = \mathbf{X}_{1\times m}\mathbf{W}_{m\times d}, \sum_{l=1}^{m} W_{lj} = 1, W_{lj} \geq 0, \forall l \in \{1, \ldots, m\}, j \in \{1, \ldots, d\}\}. \tag{38}$$

Each dimension of such representations is a convex combination of $X_1, \ldots, X_m$, that is, $f_j(\mathbf{X}) = \sum_{l=1}^{m} W_{lj} X_l$, $j = 1, \ldots, d$.

A third option is to consider representations that select relevant subsets of the m-dimensional data,

$$\{f : f(\mathbf{X}) = \mathbf{X}_{1\times m}\mathbf{W}_{m\times d}, W_{lj} \in \{0, 1\}, \sum_{l=1}^{m} W_{lj} \leq 1, \forall l \in \{1, \ldots, m\}, j \in \{1, \ldots, d\}\}. \tag{39}$$

Each dimension of such representations selects one (or none) of $X_1, \ldots, X_m$. The d-dimensional representation $f(\mathbf{X})$ thus selects at most d features from $X_1, \ldots, X_m$.

CAUSAL-REP: Maximizing PNS in practice. Solving Equation (36) involves calculating counterfactual quantities such as conditional PNS. As is discussed in Section 2.2, these quantities are not directly calculable without access to the causal structural equations.

To employ Equation (36) in practice, we employ the lower bound of PNS derived in Equation (7). Specifically, we restrict our attention to representations whose lower bound of PNS is *identifiable* via Theorem 5, and find the representation that maximizes this lower bound:

$$\textbf{(CAUSAL-REP objective)} \qquad \max_{f} \sum_{j=1}^{d} \log \underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))} + \lambda \cdot R(f\,;\, \{\mathbf{x}_i, y_i, \mathbf{c}_i\}_{i=1}^{n}), \tag{40}$$

where $\underline{PNS_n(f_j(\mathbf{X}), Y \mid f_{-j}(\mathbf{X}))}$ is the PNS lower bound (Equation (25)) in Theorem 5. The parameter $\lambda \geq 0$ indicates regularization strength. The term $R(f\,;\, \{\mathbf{x}_i, y_i, c_i\}_{i=1}^{n})$ is a regularization penalty,

$$R(f\,;\,\{\mathbf{x}_i, y_i, \mathbf{c}_i\}_{i=1}^n) \triangleq \frac{1}{d}\sum_{j=1}^{d} \log(1 - \mathrm{Rsq}(\{f_j(\mathbf{x}_i)\,;\,\mathbf{c}_i\}_{i=1}^n)) - \alpha \cdot ||W(f)||_2^2\,, \tag{41}$$

where $\mathrm{Rsq}(\{f_j(\mathbf{x}_i)\,;\,\mathbf{c}_i\}_{i=1}^n)$ is the R-squared of regressing the jth-dimension of the representation $f_j(\mathbf{x}_i)$ against the unobserved common cause $\mathbf{c}_i$. The quantity $||W(f)||_2$ represents the L_2-norm of the f function's parameters, e.g., all the weight parameters of a neural network f; α is the relative weight of the two regularization penalties.

The first part of the objective (Equation (40)) is the lower bound of conditional PNS developed in Theorem 5. Following the practical discussions in Section 2.2, its calculation requires pinpointing the unobserved common cause $\mathbf{C} = h(\mathbf{X})$ and fitting a model to $P(Y \mid f(\mathbf{X}), \mathbf{C})$ (e.g., Equation (30)).

The second part of the objective is the regularization penalty. It encourages representations whose lower bound of PNS is identifiable from observational data. Specifically, these regularization penalties aim to enforce the positivity and observability conditions in Proposition 1. (The pinpointability condition is enforced in a separate step.)

In more detail, the first penalty $\frac{1}{d}\sum_{j=1}^{d}\log(1-\mathrm{Rsq}(\{f_j(\mathbf{x}_i)\,;\,\mathbf{c}_i\}_{i=1}^n))$ in Equation (41) encourages the positivity of $f(\mathbf{X})$ given $\mathbf{C}$ in Proposition 1. The sample R-squared $\mathrm{Rsq}(\{f_j(\mathbf{x}_i)\,;\,\mathbf{c}_i\}_{i=1}^n)$ evaluates the variation of $f_j(\mathbf{X})$ explainable by $\mathbf{C}$. When the R-squared is equal to one, then the positivity condition is violated. Accordingly, a close-to-one value of R-squared implies that $f_j(\mathbf{X})$ nearly violates the positivity condition. The penalty $\frac{1}{d}\sum_{j=1}^{d}\log(1-\mathrm{Rsq}(\{f_j(\mathbf{x}_i)\,;\,\mathbf{c}_i\}_{i=1}^n))$ takes a large negative value if any dimension of the representation $f(\mathbf{X})$ nearly violates the positivity condition.

The second penalty $-\alpha \cdot ||W(f)||_2^2$ encourages representations that satisfy the observability condition. Specifically, it penalizes the coefficients in front of $(X_1, \ldots, X_m)$ in the representation function f. Imposing a large regularization parameter $\alpha > 0$ leads to representations $f(\mathbf{X})$ that only effectively depend on a small subset of $(X_1, \ldots, X_m)$; i.e., $f(\mathbf{X}) = \tilde{f}((X_j)_{j\in S})$ for some function $\tilde{f}$, with $S \subseteq \{1, \ldots, m\}$ being a set with only a few elements. Such representations are more likely to satisfy the observability condition—namely, $P((X_j)_{j\in S}) > 0$ for all values of $(X_j)_{j\in S}$—because lower-dimensional $(X_j)_{j\in S}$ are more likely to have full rank (Udell & Townsend, 2019).

Out-of-distribution prediction with CAUSAL-REP. Given a representation returned by CAUSAL-REP, how do we make predictions, especially on out-of-distribution data?

To make predictions, we train a prediction function that maps the CAUSAL-REP representation to the labels. We note that the predictions are only made using the CAUSAL-REP representation; it does not involve the unobserved common cause $\mathbf{C}$. This prediction function is different from the prediction model fitted for $P(Y \mid f(\mathbf{X}), \mathbf{C})$ (cf. Equations (30) and (31)). The rationale is that CAUSAL-REP encourages non-spurious representations; it implies that the relationship between the representation $f(\mathbf{X})$ and the label Y should generalize to out-of-distribution data. However, CAUSAL-REP places no constraints on the relationship between $\mathbf{C}$ and Y.

Specifically, suppose CAUSAL-REP returns $\hat{f}(\cdot)$ as the representation function; i.e., it maximizes Equation (40). Then we posit a model for $P(Y \mid \hat{f}(\mathbf{X}))$; e.g., a linear model:

$$P(Y \mid \hat{f}(\mathbf{X})) = \mathcal{N}(\beta_0^{\mathrm{pred}} + (\boldsymbol{\beta}^{\mathrm{pred}})^\top \hat{f}(\mathbf{X}), (\sigma^{\mathrm{pred}})^2), \tag{42}$$

or a flexible exponential family model:

$$P(Y \mid \hat{f}(\mathbf{X})) = \mathrm{EF}(g^{\mathrm{pred}}(\hat{f}(\mathbf{X}))), \tag{43}$$

where EF indicates an exponential family distribution and g^{pred} indicates a link function. We fit this model to the training data $\{(\hat{f}(\mathbf{x}_i), y_i)_{i=1}^n\}$ using maximum likelihood estimation, and then use this fitted model to predict on out-of-distribution data; e.g., we calculate $\mathbb{E}\left[(Y \mid \hat{f}(\mathbf{X}_{\mathrm{test}}))\right]$.

CAUSAL-REP with perfectly correlated spurious and non-spurious features. We have described CAUSAL-REP, a representation learning algorithm that targets non-spurious and efficient representations. One might ask: What if spurious and non-spurious features are perfectly correlated in the training dataset? That is, what if the spurious feature is on if and only if the non-spurious feature is also on? Intuitively, we can not hope to tease apart spurious and non-spurious representations in this case. How would CAUSAL-REP fare?

In such a perfect correlation case, the CAUSAL-REP algorithm would capture neither the spurious feature nor the non-spurious one. The reason is that both features are excluded from the feasible set of representations considered by CAUSAL-REP. Their PNS lower bounds are both not identifiable, but CAUSAL-REP only considers representations whose PNS lower bound is identifiable.

In more detail, suppose $f_s((X_j)_{j\in T_s})$ captures the spurious feature and $f_n((X_j)_{j\in T_n})$ captures the non-spurious features, where $T_s, T_n \subset \{1, \ldots, m\}$. They depend on disjoint subsets of $\mathbf{X}$, $T_s \cap T_n = \emptyset$, but they are perfectly correlated: $f_s((X_j)_{j\in T_s}) = f_n((X_j)_{j\in T_n})$. Then both features must be measurable with respect to the unobserved common cause $\mathbf{C}$: $f_s((X_j)_{j\in T_s}) = f_n((X_j)_{j\in T_n}) = q(\mathbf{C})$ for some function q.[6] This measurability makes their PNS lower bounds non-identifiable; it violates the positivity assumption required by Theorem 5.

2.4.2 CAUSAL-REP and the linear example continued

We described each step of CAUSAL-REP in the previous section. Algorithm 2 summarizes these steps. The key step is to maximize the CAUSAL-REP objective (Equation (40)). This step involves a nested loop of optimization: an outer loop with respect to the representation parameters $W(f)$, and an inner loop to estimate $P(Y \mid f(\mathbf{X}), \mathbf{C})$. This nesting is required because calculating the CAUSAL-REP objective involves estimating $P(Y \mid f(\mathbf{X}), \mathbf{C})$ by maximum likelihood.

To maximize the CAUSAL-REP objective, one can use standard gradient-based algorithms to handle this nested loop. When the inner loop—the maximum likelihood estimation of $P(Y \mid f(\mathbf{X}), \mathbf{C})$—has a closed-form solution (e.g., under the linear model in Equation (30)), we can plug in the solution to the CAUSAL-REP objective and directly apply a gradient-based method for optimization. When the inner loop does not admit a closed form, we can alternate between the gradient updates of the two optimizations, e.g., alternating between (1) one gradient update step to maximize the CAUSAL-REP objective, and (2) multiple gradient update steps until convergence to estimate $P(Y \mid f(\mathbf{X}), \mathbf{C})$ with maximum likelihood.

The linear example continued. We illustrate CAUSAL-REP (Algorithm 2) with the linear example in Section 2.3: pinpoint the unobserved common cause $\mathbf{C}$ from the training

[6] The reason is that $\mathbf{C}$ must render all X_j's conditionally independent due to the causal graph (Figure 2). It must also render the spurious and non-spurious features conditionally independent, because they depend on disjoint subsets of $\mathbf{X}$. As the two features are perfectly correlated, the only way to render them conditionally independent is to make them measurable with respect to $\mathbf{C}$.

Algorithm 2: CAUSAL-REP (Supervised)

```
input  : The training data and its label $\{\mathbf{x}_i^{\text{train}}, y_i^{\text{train}}\}_{i=1}^{n_{\text{train}}}$; the (out-of-distribution) test data
         $\{\mathbf{x}_i^{\text{test}}\}_{i=1}^{n_{\text{test}}}$; the probabilistic factor model that generates the training data $P(\mathbf{X}, \mathbf{C})$
output : CAUSAL-REP representation function $\hat{f}(\cdot)$; predictions on the test data $\hat{y}_i, i = 1, \ldots, n_{\text{test}}$
# Representation learning with CAUSAL-REP
Fit a probabilistic factor model (Equations (28) and (29)) to the training data and infer
  $p(\mathbf{c}_i \,|\, \mathbf{x}_i^{\text{train}}), i = 1, \ldots, n_{\text{train}}$;
if Pinpointability holds, i.e. $p(\mathbf{c}_i \,|\, \mathbf{x}_i^{\text{train}})$ is close to a point mass for all $i$ then
    foreach training datapoint $i$ do
        Pinpoint the unobserved common cause $\mathbf{C}$: $\mathbf{c}_i = h(\mathbf{x}_i^{\text{train}}) \triangleq \mathbb{E}\left[\mathbf{c}_i \,|\, \mathbf{x}_i^{\text{train}}\right]$ for $i = 1, \ldots, n_{\text{train}}$;
    end
    Maximize Equation (40) to obtain the CAUSAL-REP representation $\hat{f}$;
end
# Out-of-distribution prediction with the CAUSAL-REP representation
Estimate $P(Y \,|\, \hat{f}(\mathbf{X}))$ by fitting a model (e.g. Equations (42) and (43)) to $\{\hat{f}(\mathbf{x}_i^{\text{train}}), y_i^{\text{train}}\}_{i=1}^{n_{\text{train}}}$
Predict on test data using the fitted model: $\hat{y}_i = \mathbb{E}\left[Y \,|\, \hat{f}(\mathbf{x}_i^{\text{test}})\right], i = 1, \ldots, n_{\text{test}}$
```

data $\{\mathbf{x}_i^{\text{train}}, y_i^{\text{train}}\}_{i=1}^{n_{\text{train}}}$, obtain the CAUSAL-REP objective $\hat{f}$, and predict on some out-of-distribution test data $\{\mathbf{x}_i^{\text{test}}\}_{i=1}^{n_{\text{test}}}$. We focus on learning two-dimensional representations, $f(\mathbf{X}) = (f_1(\mathbf{X}), f_2(\mathbf{X}))$.

We first perform the same pinpointability step to obtain $\hat{c}_i$ as in Section 2.3.

We next maximize the CAUSAL-REP objective (Equation (40)). We calculate the CAUSAL-REP objective by plugging in the calculation of $\underline{PNS_n(f_j(\mathbf{X}), Y \,|\, f_{-j}(\mathbf{X}))}$ in Equation (32), where we adopt a linear model for $P(Y \,|\, f(\mathbf{X}), \mathbf{C})$. In more detail, to optimize the CAUSAL-REP objective, we need to write (the parameters of) the fitted model of $P(Y \,|\, f(\mathbf{X}), \mathbf{C})$ as a function of the representation function f. We use the closed-form estimates of these linear model parameters when fitted to the training data $\{\mathbf{x}_i^{\text{train}}, y_i^{\text{train}}\}_{i=1}^{n^{\text{train}}}$:

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}) &= \arg\min \sum_{i=1}^{n} (y_i^{\text{train}} - \beta_0 - \beta_1 \cdot f_1(\mathbf{x}_i^{\text{train}}) - \beta_2 \cdot f_2(\mathbf{x}_i^{\text{train}}) - \gamma \cdot \hat{c}_i)^2 \\ &= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}^\top \tilde{Y}),\end{aligned}$$

where $\tilde{\mathbf{X}}$ is an $n^{\text{train}} \times 4$ matrix with $\tilde{\mathbf{X}}_{i1} = 1$, $\tilde{\mathbf{X}}_{i2} = f_1(\mathbf{x}_i^{\text{train}})$, $\tilde{\mathbf{X}}_{i2} = f_2(\mathbf{x}_i^{\text{train}})$, $\tilde{\mathbf{X}}_{i2} = \hat{c}_i$, and $\tilde{Y}$ is an $n \times 1$ vector with $\tilde{Y}_{i1} = y_i^{\text{train}}$, $i = 1, \ldots, n$. (When closed-form solutions are not available, we need to perform gradient-based optimization to obtain the fitted model parameters.)

Plugging in these terms, we maximize the CAUSAL-REP objective via gradient descent with respect to the parameters of the representation function $W(f)$. The optimal $\hat{W}(f)$ give the CAUSAL-REP representation function $\hat{f}$.

Finally, we perform out-of-distribution prediction with the CAUSAL-REP representation $\hat{f}$. We train a prediction function by fitting a linear model to $P(Y \,|\, \hat{f}(\mathbf{X}))$,

$$y_i^{\text{train}} = \beta_0^{\text{pred}} + \beta_1^{\text{pred}} \cdot \hat{f}_1(\mathbf{x}_i^{\text{train}}) + \beta_2^{\text{pred}} \cdot \hat{f}_2(\mathbf{x}_i^{\text{train}}) + \epsilon, \epsilon \sim \mathcal{N}(0, (\sigma^{\text{pred}})^2). \tag{44}$$

We obtain the estimated regression coefficients $\{\hat{\beta}_0^{\text{pred}}, \hat{\beta}_1^{\text{pred}}, \hat{\beta}_2^{\text{pred}}, \sigma^{\text{pred}}\}$ by maximum likelihood. They allow us to make predictions on the test data $\{\mathbf{x}_i^{\text{test}}\}_{i=1}^{n_{\text{test}}}$,

$$\hat{y}_i = \mathbb{E}\left[Y \mid \hat{f}(\mathbf{x}_i^{\text{test}})\right] = \hat{\beta}_0^{\text{pred}} + \hat{\beta}_1^{\text{pred}} \cdot \hat{f}_1(\mathbf{x}_i^{\text{test}}) + \hat{\beta}_2^{\text{pred}} \cdot \hat{f}_2(\mathbf{x}_i^{\text{test}}). \tag{45}$$

2.4.3 Extending CAUSAL-REP to unsupervised settings

We extend CAUSAL-REP to unsupervised settings where labels are not available. We focus on the task of instance discrimination in unsupervised representation learning (Hadsell et al., 2006), where the goal is to find representations that can distinguish the different subjects.

The unsupervised CAUSAL-REP. We focus on the unsupervised setting where some form of data augmentation is available. That is, the dataset contains one raw observation for each subject; this raw observation is then augmented multiple times, which leads to multiple observations per subject. Given this unsupervised dataset, we attach the subject ID to each observation as a supervised label. We can thereby formulate the instance discrimination problem as a supervised task, one that finds representations that are informative of the subject IDs and extend CAUSAL-REP to an unsupervised setting. The resulting algorithm turns out to be closely related to contrastive learning (Chen et al., 2020a).

In more detail, we begin with a dataset with n i.i.d. samples, $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Assume that each data point is associated with a different subject i. Next, we augment each data point with $U-1$ augmentations, producing an augmented dataset, $\{\{\mathbf{x}_i^u\}_{u=1}^U\}_{i=1}^n$, $U \geq 2$. We then create an n-dimensional label, $\mathbf{y}_i^u = (y_{i1}^u, \ldots, y_{in}^u)$, for each data point in the augmented dataset, where

$$y_{is}^u = \mathbb{I}\{\mathbf{x}_i^u \text{ belongs to subject } s\}, s = 1, \ldots, n. \tag{46}$$

This labeling turns the unsupervised problem into a supervised one with respect to the augmented labeled dataset $\{\{\mathbf{x}_i^u, \mathbf{y}_i^u\}_{u=1}^U\}_{i=1}^n$. We then employ the CAUSAL-REP objective to find representations that are necessary and sufficient causes of the n outcomes $(Y_1, \ldots, Y_n)$:

$$\max_f \sum_{s=1}^n \sum_{j=1}^m \log \underline{PNS_{n\cdot U}(f_j(\mathbf{X}), Y_s \mid f_{-j}(\mathbf{X}))} + \lambda \cdot R(f\,;\, \{\{\mathbf{x}_i^u, y_{is}^u, \mathbf{c}_i^u\}_{u=1}^U\}_{i=1}^n), \tag{47}$$

where $\underline{PNS_{n\cdot U}(f_j(\mathbf{X}), Y_s \mid f_{-j}(\mathbf{X}))}$ is the conditional efficiency and spuriousness (Equation (6)) for the sth outcome; it is calculated on the augmented dataset $\{\{\mathbf{x}_i^u, \mathbf{y}_i^u\}_{u=1}^U\}_{i=1}^n$. The penalty term $R(\cdot)$ is the same penalty as in the CAUSAL-REP objective (Equations (40) and (41)). Solving Equation (47) produces non-spurious and efficient representations of $\mathbf{X}$ for instance discrimination. (We summarize the algorithm in Algorithm 4.)

The optimization objective in Equation (47) considers the n outcomes $\{Y_s\}$ one at a time and then averages over them. It focuses on the contrast of one-vs-all, finding representations that distinguish each subject from the rest. This objective does not consider the n outcomes as a single n-dimensional outcome $(Y_1, \ldots, Y_n)$ and calculates its PNS.

Unsupervised CAUSAL-REP and contrastive learning. The unsupervised CAUSAL-REP objective (Equation (47)) is closely related to contrastive learning (Chen et al., 2020a), whose learning objective is to maximize

$$\sum_{u_1=1}^U \sum_{u_2=1}^U -\log \frac{\exp(d(f(\mathbf{x}_i^{u_1}), f(\mathbf{x}_i^{u_2})))}{\sum_{i'\neq i} \sum_{u_*=1}^U \exp(d(f(\mathbf{x}_i^{u_1}), f(\mathbf{x}_{i'}^{u_*})))}, \tag{48}$$

where $d(\cdot, \cdot)$ is some distance function. This contrastive learning objective targets the same instance discrimination task as the unsupervised CAUSAL-REP does; it aims for representations that can determine whether two data points come from the same subject. Contrastive learning thus is also able to predict whether each data point belongs to subject s, for $s = 1, \ldots, n$, as is the unsupervised CAUSAL-REP.

That said, the two algorithms only coincide when different dimensions of $\mathbf{X}$ are independent—they do not share an unobserved common cause as in Figure 2. The two algorithms behave very differently when such a common cause exists, especially when the common cause induces some spurious feature (e.g., image color). Concretely, there may be a subset of $(X_1, \ldots, X_m)$ that is highly correlated with the label $\mathbf{Y}$ but cannot causally determine $\mathbf{Y}$. In such cases, contrastive learning may pick up these spurious features if no augmentations are performed for these features (e.g., there is no augmentation that randomly changes the color of images). In contrast, the unsupervised CAUSAL-REP would not capture these spurious features.

This difference illustrates how CAUSAL-REP learns non-spurious representations in a different way than data-augmentation-based methods (including contrastive learning). CAUSAL-REP capitalizes on a causal perspective and its treatment of high-dimensional $\mathbf{X}$. Data-augmentation-based methods instead wipe out the correlation between spurious features and the label by performing data augmentation. We will illustrate this difference further in Section 2.5.5.

2.5 Empirical Studies of CAUSAL-REP

We study CAUSAL-REP in both image and text datasets. We study the following questions:

1. How well do probabilities of causation measure efficiency and non-spuriousness of features? (Section 2.5.1)

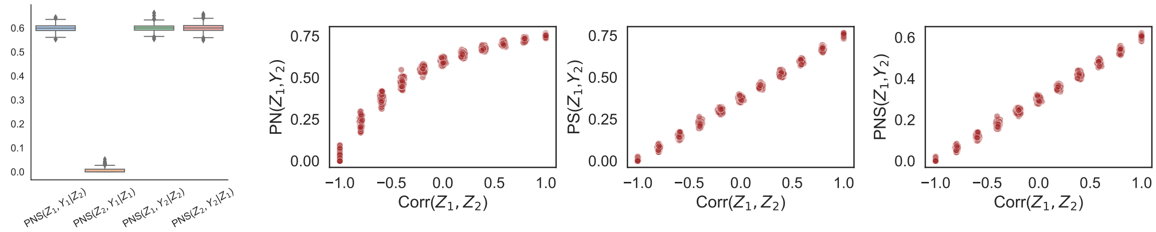
2. Does supervised CAUSAL-REP produce non-spurious representations for synthetic (Section 2.5.2), image (Section 2.5.3), and text (Section 2.5.4) data?

3. How well does unsupervised CAUSAL-REP perform on instance discrimination? (Section 2.5.5)

We find that probabilities of causation (POC) are effective in distinguishing efficient/inefficient and non-spurious/spurious representations. Moreover, CAUSAL-REP finds non-spurious features in both supervised and unsupervised settings; it also outperforms existing unsupervised representation learning algorithms in downstream prediction.

2.5.1 How well do probabilities of causation measure efficiency and non-spuriousness of features?

We first study the correspondence between probabilities of causation (PS, PN, PNS) and the efficiency/non-spuriousness of representations. We generate features with known efficiency and non-spuriousness properties; we find that the (lower bound of) probabilities of causation in Theorem 1 are consistent with these properties.



(a) Conditional PNS

(b) Probabilities of causation (POC) of Z_1 for Y_2

Figure 4: Probabilities of causation can distinguish spurious/non-spurious and efficient/inefficient features. (a) Conditional PNS signals that Z_2 is spurious for Y_1. (b) Probabilities of causation (PNS, PN, PS) of Z_1 for Y_2 increase as Z_1 and Z_2 become increasingly highly correlated.

The simulated data. We simulate two binary features Z_1, Z_2 and two binary outcomes Y_1, Y_2:

$$\begin{aligned}
&Z_1 \sim \text{Bernoulli}(0.4),\\
&Z_2 = Z_1 \oplus \text{Bernoulli}(p), p \in \{0, 0.1, \ldots, 0.9, 1\},\\
&Y_1 = Z_1 \oplus \text{Bernoulli}(0.2),\\
&Y_2 = (Z_1 \& Z_2) \oplus \text{Bernoulli}(0.2),
\end{aligned}$$

where $\oplus$ indicates the XOR operator. We vary the parameter p; a small p implies a high correlation between Z_1 and Z_2. The generative process of Y_1, Y_2 implies that (1) Z_1 is necessary (a.k.a. efficient) and sufficient (a.k.a. non-spurious) for Y_1, but is necessary and insufficient for Y_2; (2) Z_2 is neither necessary nor sufficient for Y_1, but is necessary and insufficient for Y_2; (3) $Z_1 \& Z_2$ is necessary and sufficient for Y_2. Moreover, when the correlation between Z_1 and Z_2 increases, then using both features as a representation become increasingly inefficient; the conditional necessity decreases. We calculate the lower bound of PS, PN, PNS of Z_1, Z_2 for both outcomes Y_1, Y_2; the exact value is not identifiable from data.

Results. Figures 4, 12 and 13 present the probabilities of causation of the features. Figure 4a shows that the (lower bound of) conditional PNS can correctly signal the non-spuriousness of features: Z_2 is spurious for Y_1, but Z_1 is non-spurious for Y_1, and both Z_1, Z_2 are non-spurious for Y_2. Moreover, we study how probabilities of causation (POC) fare when the correlation between Z_1 and Z_2 increases. Figure 4 shows that the (lower bounds of) POC of Z_1 for Y_2 increase as the correlation increases, which is consistent with the intuition that Z_2 becomes less necessary for Y_2 given Z_1 given increasingly higher correlations. Similarly, Figures 12 and 13 show that (the lower bounds of) the unconditional POC of Z_2 for Y_1 also increase. It is also consistent with the intuition: Z_2 is an increasingly better surrogate of Z_1 for Y_1 under higher correlations between Z_1 and Z_2.

2.5.2 Does supervised CAUSAL-REP pick up spurious features in synthetic data?

We next study the performance of the supervised version of CAUSAL-REP in a toy synthetic dataset. We use PPCA as the pinpointing factor model and linear functions as the class of representation functions.

The simulated data. We generate a dataset of core and spurious features $(X_1, \ldots, X_5)$, which are highly correlated in the training set but not so in the test set,

$$\begin{aligned}\bar{X}_1^{\text{train}}, \ldots, \bar{X}_5^{\text{train}} &\sim \mathcal{N}(0, 0.05 \cdot I_5 + 0.95 \cdot J_5),\\ \bar{X}_1^{\text{test}}, \ldots, \bar{X}_5^{\text{test}} &\sim \mathcal{N}(0, 0.05 \cdot I_5 + 0.05 \cdot J_5),\end{aligned}$$

where I_5 is a 5×5 identity matrix and J_5 is a 5×5 all-ones matrix. For both training and test sets, we add noise to features to lower the signal-to-noise ratio of the problem: $X_j \sim \mathcal{N}(\bar{X}_j, 0.4^2), j = 1, 2; X_j \sim \mathcal{N}(\bar{X}_j, 0.3^2), j = 3, 4, 5$. Finally, we generate an outcome Y that only depends on the core features $Y \sim \mathcal{N}(\beta_1 X_1 + \beta_2 X_2, 1)$, where the coefficients are drawn from a uniform $\beta_j \sim \text{Unif}[0, 10], j = 1, 2,$.

Evaluation. We study whether these algorithms pick up spurious features by evaluating their predictive performance on a non-spuriousness test set. This test set is constructed such that the predictive performance indicates whether the representation picks up spurious features: if an algorithm picks up spurious features, then its predictions on this test set will suffer. Specifically, we make the spurious features much less predictive in non-spuriousness test set than in training set: the relationship between the spurious feature and the label in the test set is set to be the exact opposite to that in the training set. In this way, if the algorithm picks up spurious features, then it will perform poorly on the test set. In this sense, the test set serves as a "test strip" for spurious features; the more the algorithm relies on the spurious feature, the lower the predictive performance is on the test set. We name such test sets as "non-spuriousness test sets."

Results. We compare the performance of CAUSAL-REP, linear regression, and an oracle algorithm; the oracle is equipped with the knowledge of which features are spurious and only performs linear regression against the non-spurious features. Figure 5a presents the result: CAUSAL-REP outperforms linear regression in predictive R-squared on non-spuriousness test set; its predictive performance is not far from the oracle algorithm.

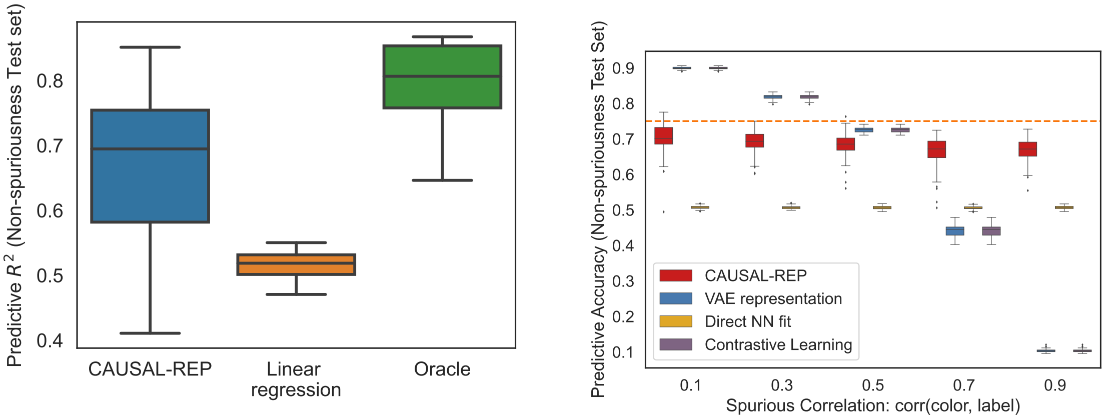
2.5.3 Does supervised CAUSAL-REP produce non-spurious representations for image data? A study on Colored MNIST and CelebA

We next study the supervised version of CAUSAL-REP in image datasets: Colored MNIST (Arjovsky et al., 2019) and CelebA (Liu et al., 2015). We consider an implementation of the supervised CAUSAL-REP algorithm which adopts a VAE with 64 latent dimensions as the probabilistic factor model for pinpointing. For representation functions, we consider a two-layer neural network with 20-dimensional outputs. We again evaluate the non-spuriousness of the representations by the predictive accuracy on non-spuriousness test sets; if the learned representation captures spurious features, then it will suffer in its predictive performance on non-spuriousness test sets.

Competing methods. We compare CAUSAL-REP with a baseline representation learning algorithm that fits a neural network to the labels and extracts its penultimate layer as the representation (Bengio et al., 2013). As we use VAE for pinpointing in CAUSAL-REP in this study, we also compare with directly adopting the VAE representation for prediction.

The Colored MNIST study. We focus on the colored MNIST data with the digits '3' and '8' and colors 'red' and 'green'; see Appendix G.2 for the detailed experimental setup.

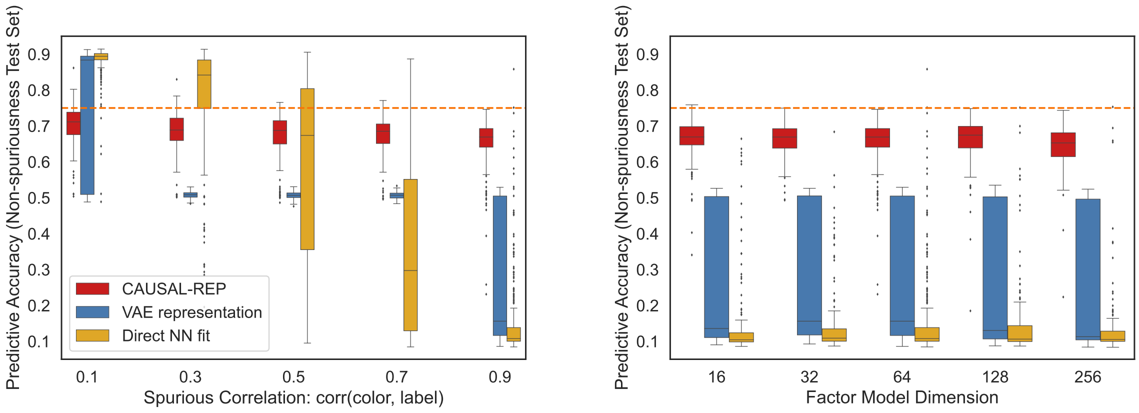
Figure 6a presents the non-spuriousness test sets' predictive accuracy of CAUSAL-REP across different levels of spurious correlations. CAUSAL-REP outperforms other baseline methods in non-spuriousness test sets' predictive accuracy when the spurious correlation is larger than 0.5. Its



(a) Supervised CAUSAL-REP: Toy synthetic data

(b) Unsupervised CAUSAL-REP: Colored and Shifted MNIST

Figure 5: CAUSAL-REP learns non-spurious representations in both supervised and unsupervised settings. (a) In toy synthetic data, CAUSAL-REP outperforms linear regression in predictive performance on non-spuriousness test sets. (b) CAUSAL-REP outperforms baseline representation learning algorithms (e.g. directly fitting neural networks, or performing contrastive learning, or adopting VAE representations) in the colored and shifted MNIST dataset. The dashed yellow line indicates the theoretical maximum predictive accuracy on non-spuriousness test sets. (Higher is better.)



(a)

(b)

Figure 6: (a) The supervised CAUSAL-REP learns non-spurious representations in colored MNIST and outperforms baseline representation learning algorithms (e.g. directly fitting neural networks, or adopting VAE representations) in in predictive performance on non-spuriousness test sets. (b) The performance of CAUSAL-REP is robust to the choice of the latent dimensionality of probabilistic factor models. The dashed yellow line indicates the theoretical maximum of predictive accuracy on non-spuriousness test sets. (Higher is better.)

performance remains close to the theoretical maximum (the yellow dashed line) across different levels of the spurious correlation, suggesting that CAUSAL-REP does not pick up spurious features. In contrast, the non-spuriousness test-set predictive performance of representation learning by

target	spurious	spurious corr. (train)	spurious corr. (test)	CAUSAL-REP	Direct NN fit	VAE rep.
Arched Brows	Eye Bags	0.784	-0.797	**0.539(0.036)**	0.514(0.029)	0.499(0.012)
Arched Brows	Earrings	0.799	-0.791	**0.521(0.025)**	0.504(0.022)	0.494(0.008)
Attractive	Necklace	0.793	-0.787	**0.537(0.022)**	0.505(0.030)	0.485(0.018)
Black Hair	Mouth Open	0.791	-0.795	**0.594(0.033)**	0.566(0.060)	0.505(0.010)
Goatee	Male	0.889	0.053	**0.728(0.073)**	0.566(0.102)	**0.867(0.165)**
Mustache	Black Hair	0.764	-0.778	0.525(0.023)	0.512(0.037)	**0.540(0.005)**
Mustache	Male	0.892	0.088	**0.787(0.066)**	0.555(0.097)	**0.855(0.212)**

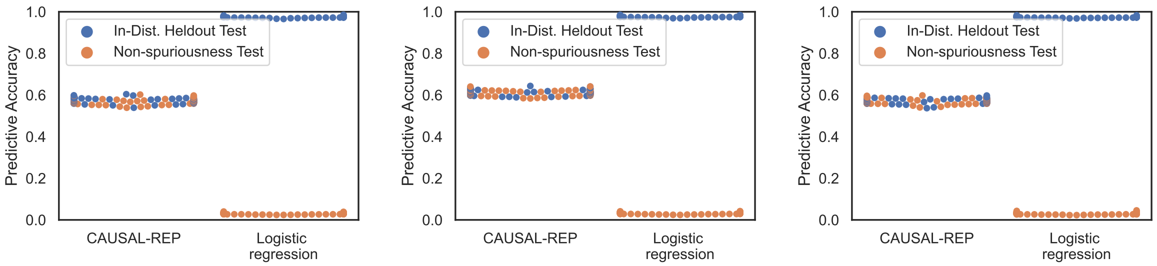
Table 1: CAUSAL-REP learns non-spurious representations in the CelebA dataset and outperforms baseline representation learning algorithms in non-spuriousness test sets. The last three columns report the predictive accuracy on non-spuriousness test sets by different algorithms.

directly fitting neural networks quickly degrades as the spurious correlation increases; so does non-spuriousness test sets' prediction performance with VAE representations.

Figure 6b evaluates the robustness of CAUSAL-REP against the latent dimensionality choice of the pinpointing factor model, fixing the spurious correlation at 0.9. We find that the performance of CAUSAL-REP does not change much as we vary the latent dimensionality of pinpointing VAE. The relative performance of CAUSAL-REP and the competing methods also stay stable.

The CelebA study. We next study CAUSAL-REP on the CelebA dataset (Liu et al., 2015). We create training and test sets, each containing 5,000 data points, by subsampling the CelebA datasets. We focus on face attributes with a relatively balanced distribution in the raw CelebA dataset. We designate pairs of target attributes and spurious attributes, and subsample such that the two are highly correlated in the training set and not as correlated in the test set. We then perform representation learning and prediction for the target labels, using CAUSAL-REP and other competing methods.

Table 1 presents the results for different pairs of target and spurious face attributes. Though the spurious label and the target label are highly correlated, CAUSAL-REP can pick up non-spurious features that inform the target label and outperform the baseline algorithm that directly fits a neural network. In most settings, CAUSAL-REP also outperforms in non-spuriousness test set prediction with VAE representations, indicating that CAUSAL-REP does not rely as much on spurious features. The exception is that VAE representations can outperform CAUSAL-REP representations when the spurious correlations climb up to 0.9, though the performance of CAUSAL-REP is still competitive. The spurious features in these settings violate the positivity condition: Mustache and goatee exclusively appear on males in the subsampled dataset.



(a) Amazon reviews **(b)** Tripadvisor reviews **(c)** Yelp reviews

Figure 7: CAUSAL-REP learns non-spurious representations across reviews text copura; its predictive performance is stable across in-distribution heldout test sets and non-spuriousness test sets.

2.5.4 Does supervised CAUSAL-REP produce non-spurious representations for text data? A study on reviews corpora and sentiment analysis

We next study CAUSAL-REP on text datasets: the Amazon (Wang et al., 2011, 2010), Tripadvisor,[7] and Yelp[8] reviews corpora, and the IMDB-L, IMDB-S, and Kindle dataset as is processed in Wang & Culotta (2020, 2021). In these studies, we convert these corpora into bags of words. We use PPCA as a pinpointing factor model for CAUSAL-REP and consider representations where each dimension is a convex combination of words.

The reviews corpora study. We begin with the raw reviews datasets from Amazon, Tripadvisor, and Yelp. We create a binary label for each review by converting 4 and 5 stars to a positive label and 1 and 2 stars to a negative label. We then inject irrelevant words to the training dataset as spurious features by randomly adding in "as," "also," "am," and "an" to reviews with positive labels; the resulting spurious correlation is around 0.9. We create two test datasets: one is in-distribution heldout test set with the spurious words present as in the training set; the other is non-spuriousness test set without the randomly added spurious words.

Figure 7 presents the predictive accuracy of CAUSAL-REP in both test sets. We find that the predictive performance of CAUSAL-REP is stable across in-distribution heldout test sets and non-spuriousness test sets, suggesting that it learns non-spurious representations. In contrast, logistic regression predicts well in the heldout in-distribution test set but not in the non-spuriousness test set, suggesting that it picks up spurious features that only exist in the training set.

Table 4 presents the most informative words of the (positive or negative) ratings, suggested by the CAUSAL-REP representation and the logistic regression coefficients. Across three reviews corpora, logistic regression returns the spurious words "as," "also," "am," and "an" as the top words. In contrast, CAUSAL-REP extracts words that are more relevant for the ratings.

The sentiment analysis study. We next employ CAUSAL-REP for sentiment classification on the IMDB-L, IMDB-S, and Kindle datasets (Wang & Culotta, 2020, 2021). We evaluate CAUSAL-REP on their raw in-distribution heldout test sets and non-spuriouness test sets, without employing additional data augmentations as in Wang & Culotta (2020, 2021).

[7] http://times.cs.uiuc.edu/ wang296/Data/

[8] https://www.yelp.com/dataset/documentation/main

	In-Dist. Heldout test set		Non-spuriousness test set	
	Logistic Regression	CAUSAL-REP	Logistic Regression	CAUSAL-REP
IMDB-L	**0.669**	0.645	0.591	**0.642**
IMDB-S	**0.836**	0.682	0.570	**0.621**
Kindle	**0.850**	0.618	0.468	**0.572**

Table 2: CAUSAL-REP outperforms naive representation learning algorithms in predicting on non-spuriousness test sets.

Table 2 presents the predictive performance of CAUSAL-REP compared with logistic regression. While logistic regression outperforms in its predictive performance on in-distribution heldout test sets, CAUSAL-REP outperforms on non-spuriouness test sets. Moreover, the predictive performance of CAUSAL-REP is similar across in-distribution heldout and non-spuriouness test sets, suggesting that CAUSAL-REP produces non-spurious representations.

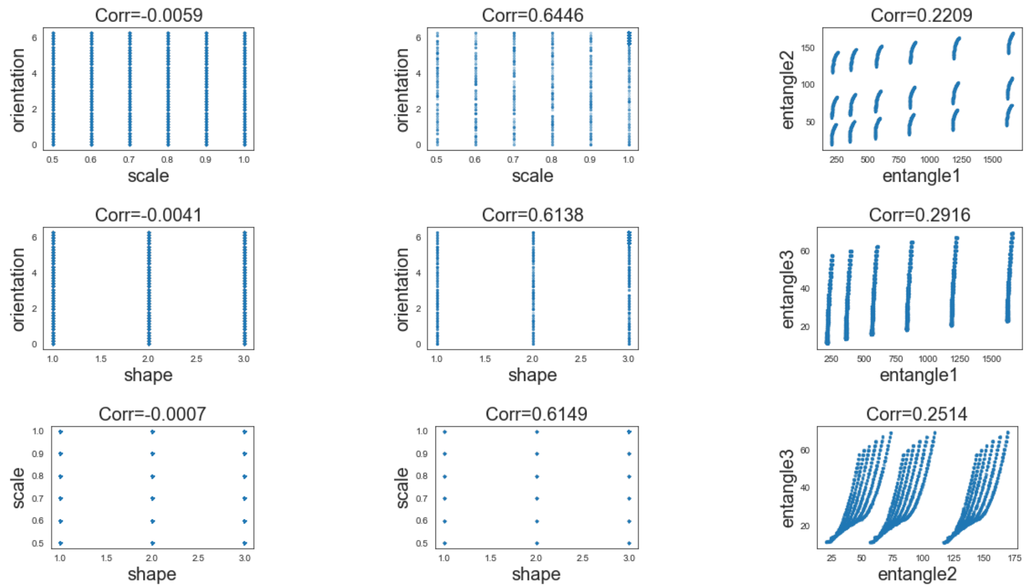
2.5.5 How well does unsupervised CAUSAL-REP perform on instance discrimination? A study on colored and shifted MNIST

Finally, we study CAUSAL-REP in the unsupervised setting. We focus on image datasets in the unsupervised setting because non-spurious features that distinguish different images are more readily defined in the image domain than in the text domain. We evaluate the non-spuriousness of the unsupervised CAUSAL-REP again by its predictive performance on non-spuriousness test sets. Given the unsupervised CAUSAL-REP representation, we fit a prediction model to the target label and test its predictive performance on non-spuriousness test sets.

Competing methods. We compare CAUSAL-REP with baseline representation learning algorithms that fit a neural network to the subject ID label in Section 2.5.5. We also compare with contrastive learning (Chen et al., 2020a) and the VAE representation.

The colored and shifted MNIST study. We construct the colored and shifted MNIST dataset by coloring and shifting digits in a way that is highly correlated with the digit labels; see Appendix G.4 for details. (Representation learning with the unsupervised CAUSAL-REP does not make use of the digit labels; these labels are only used in producing a prediction function from the representations to the labels.)

Figure 5b presents the predictive accuracy of the non-spuriousness test sets for unsupervised CAUSAL-REP on colored and shifted MNIST. CAUSAL-REP outperforms baseline representation learning algorithms when the spurious correlation is high, and their predictive accuracy stays close to the theoretical maximum (the yellow dashed line). Comparing CAUSAL-REP with contrastive learning with shift augmentation, we find that the predictive accuracy of the non-spuriousness test sets for contrastive learning degrades when the spurious correlations increase over 0.7. This is because the shift augmentation can not rule out the spurious color feature; it can only avoid picking up the spurious shift feature. This observation echoes the discussion in Section 2.5.5, illustrating how the unsupervised CAUSAL-REP relies on a different mechanism to produce non-spurious representations than contrastive learning and data augmentation.



(a) Disentangled and uncorrelated (IOSS=0.13)

(b) Disentangled but highly correlated (IOSS=0.14)

(c) Entangled but w/ low correlations (IOSS=0.27)

Figure 8: Disentangled features have independent support even though they may be correlated. Moreover, the independence of support can distinguish disentangled and entangled features. This figure illustrates how entangled and disentangled features differ using pairwise scatter plots. Figure 8a considers the ground truth features (shape, scale, orientation) of the dSprites dataset. These features are disentangled. They also have independent support; e.g., conditional on 'scale', the set of values that 'orientation' can take does not change with 'scale.' Visually, these disentangled features have scatter plots that occupy rectangular (or hyperrectangular) regions. Figure 8b considers the same features but in a subset of the dSprites dataset where the features are correlated. These features, though correlated, are still disentangled; they also have independent support. Figure 8c considers three entangled features, each of which is a nonlinear transformation of the three ground-truth features. These features are not disentangled. Their supports are also not independent. Conditional on 'entangle1,' the possible values 'entangle2' can take depends on the value of 'entangle1.'

3. Unsupervised Representation Learning: Disentanglement

In this section we study the desideratum of disentanglement in unsupervised representation learning. Unsupervised representation learning aims to find a low-dimensional representation for high-dimensional objects without the help of labels. Given an m-dimensional object, $\mathbf{X} = (X_1, \ldots, X_m)$, the goal is again to find a d-dimensional representation, $\mathbf{Z} = (Z_1, \ldots, Z_d) \triangleq (f_1(\mathbf{X}), \ldots, f_d(\mathbf{X}))$.

We focus on a causal definition of disentanglement: different dimensions of the representation shall encode features that do not causally affect each other (Suter et al., 2019). An absence of causal relationships among different dimensions of the representation can be viewed as independent manipulability of each dimension. Other related criteria include the notion of independent mechanisms (Suter et al., 2019; Parascandolo et al., 2018; Schölkopf et al., 2012) and the concept of independently controllable factors (Thomas et al., 2017, 2018). Based on this causal definition, we will develop unsupervised metrics and algorithms for unsupervised disentanglement.

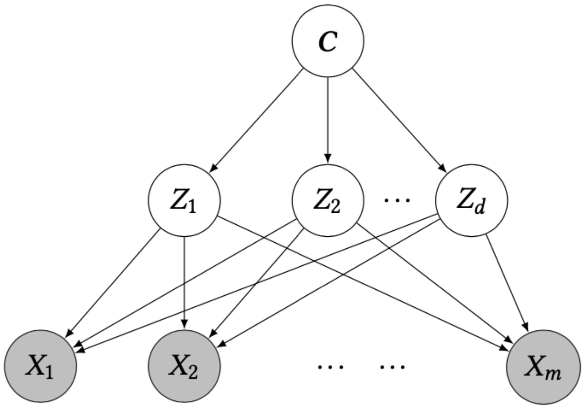


Figure 9: The SCM of unsupervised disentanglement (Suter et al., 2019).

3.1 The Causal Definition of Disentanglement

We begin with reviewing a causal definition of disentanglement. A representation is (causally) *disentangled* if (1) it represents features that can generate the original data and (2) its dimensions correspond to features that do not causally affect each other (Suter et al., 2019). Definition 6 casts these requirements in terms of a structural causal model (SCM) (Pearl, 2011).

Definition 6 ((Causal) Disentanglement (Suter et al., 2019))**.** *A d-dimensional representation,* $\mathbf{Z} = (Z_1, \ldots, Z_d)$*, is disentangled if it represents features that generate the object of interest* $\mathbf{X} = (X_1, \ldots, X_m)$ *according to the following causal SCM:*

$$\begin{aligned} &\mathbf{C} \leftarrow u_{\mathbf{C}}, && \\ &Z_j \leftarrow f_j^z(\mathbf{C}, u_{z,j}), && j = 1, \ldots, d, \qquad (49)\\ &X_l \leftarrow f_l^x(\mathbf{Z}, u_{\mathbf{X}}), && l = 1, \ldots, m, \qquad (50)\end{aligned}$$

where $\mathbf{C} = (C_1, \ldots, C_K)$ *denotes an K-dimensional unobserved confounder that can affect* $\mathbf{Z}$ *and hence induce correlations among its components* $(Z_1, \ldots, Z_d)$*. Figure 2 describes this SCM using a causal DAG.*[9]

As an example to illustrate this definition, we return to the dataset of animal images in Section 1. If we had access to the d ground-truth features that generated the dataset (e.g., fur color, number of legs), we could create a d-dimensional representation that concatenates them. Such a representation would be disentangled, as the ground-truth features do not causally affect each other. That said, they may still be correlated due to some unobserved confounders (a.k.a., common causes). For example, the animal's genetic information may be an unobserved confounder; the genes can causally affect these features, and thus induce correlations among them.

Notably, the causal definition of disentanglement opens up the possibility to consider correlated features that are disentangled. We work with this disentanglement definition in the rest of this section.

[9] This causal SCM allows Z_j to depend on only a subset of the K confounder dimensions because the constraint imposed by a SCM resides in the absence of causal connections.

3.2 Measuring Disentanglement with Observational Data

How do we assess the disentanglement of a representation? In the do notation, a representation $\mathbf{Z}$ is disentangled if, for $j = 1, \ldots, d$,

$$P(Z_j \,|\, \mathrm{do}(\mathbf{Z}_{-j} = \mathbf{z}_{-j})) = P(Z_j), \forall \mathbf{z}_{-j} \in \mathcal{Z}_{-j}, \tag{51}$$

where $P(\cdot \,|\, \mathrm{do}(\cdot))$ denotes the intervention distribution (Pearl, 2011). We cannot, however, estimate $P(Z_j \,|\, \mathrm{do}(\mathbf{Z}_{-j}))$ in practice. The reason is that we only have access to an observational dataset of the representation $\{\mathbf{z}_i\}_{i=1}^n = \{f(\mathbf{x}_i)\}_{i=1}^n$. This dataset is generally insufficient for identifying $P(Z_j \,|\, \mathrm{do}(\mathbf{Z}_{-j}))$ due to the unobserved common cause $\mathbf{C}$ (Pearl, 2011; Imbens & Rubin, 2015). This difficulty of estimating $P(Z_j \,|\, \mathrm{do}(\mathbf{Z}_{-j}))$ is a core challenge in assessing disentanglement.

To tackle this challenge, we consider the possibility of finding observable implications of disentanglement. These observable implications will be necessary conditions associated with Equation (51). Though possibly insuffiecent, they can still help us reject representations that are not disentangled. Specifically, when a representation violates these observable implications, it cannot be disentangled.

3.2.1 Observable implications of disentanglement: The independent support condition

We turn to the specific observable implications that underpin our approach to disentanglement. We will prove that if a representation is disentangled, then its different dimensions must have independent support under a standard positivity condition.

Theorem 7 (Disentanglement ⇒ Independent support). *Assume the unobserved common cause* $\mathbf{C}$ *satisfies a positivity condition: for all* j*, we have* $P(Z_j \,|\, \mathbf{C}) > 0$ *iff* $P(Z_j) > 0$*. Then the support of the interventional distribution coincides with that of the observational distribution:*

$$\mathit{supp}(Z_j \,|\, \mathrm{do}(Z_{j'} = z_{j'})) = \mathit{supp}(Z_j \,|\, Z_{j'} = z_{j'}), \tag{52}$$

where $j, j' \in \{1, \ldots, d\}$*,* $j \neq j'$*, and the density at* $z_{j'}$ *is nonzero,* $p(z_{j'}) > 0$*. As a consequence, different dimensions of a disentangled representation* $\mathbf{Z} = (Z_1, \ldots, Z_d)$ *must have independent support:*

$$\begin{aligned} \mathit{supp}(Z_1, \ldots, Z_d) &= \mathit{supp}(Z_1) \times \cdots \times \mathit{supp}(Z_d), \\ \mathit{supp}(Z_j \,|\, Z_{\mathcal{S}}) &= \mathit{supp}(Z_j) \text{ for all } \mathcal{S} \subseteq \{1, \ldots, d\} \backslash j. \end{aligned} \tag{53}$$

Proof We focus on proving Equation (52). (The remainder of the proof is in Appendix I.)

First notice that the positivity condition on $\mathbf{C}$ guarantees that $\mathrm{supp}(\mathbf{C}) = \mathrm{supp}(\mathbf{C} \,|\, Z_j = z_j)$ for all $p(z_j) > 0$:

$$\begin{aligned} \mathbb{I}\{P(\mathbf{C} \,|\, Z_j) > 0\} &= \mathbb{I}\left\{\frac{P(\mathbf{C})P(Z_j \,|\, \mathbf{C})}{P(Z_j)} > 0\right\} \\ &= \mathbb{I}\{P(\mathbf{C}) > 0\}\,\mathbb{I}\{P(Z_j \,|\, \mathbf{C}) > 0\} \\ &= \mathbb{I}\{P(\mathbf{C}) > 0\}\,\mathbb{I}\{P(Z_j) > 0\} \\ &= \mathbb{I}\{P(\mathbf{C}) > 0\}, \end{aligned} \tag{54}$$

where the third equality is due to the positivity condition. Therefore, Equation (52) holds because

$$\begin{aligned}
&\mathbb{I}\{P(Z_j \,|\, \mathrm{do}(Z_{j'})) > 0\}\\
&= \mathbb{I}\left\{\int P(Z_j \,|\, Z_{j'}, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}\mathbf{C} > 0\right\}\\
&= \int \mathbb{I}\left\{P(Z_j \,|\, Z_{j'}, \mathbf{C}) > 0\right\} \mathbb{I}\left\{P(\mathbf{C}) > 0\right\} \mathrm{d}\mathbf{C}\\
&= \int \mathbb{I}\left\{P(Z_j \,|\, Z_{j'}, \mathbf{C}) > 0\right\} \mathbb{I}\left\{P(\mathbf{C} \,|\, Z_{j'}) > 0\right\} \mathrm{d}\mathbf{C}\\
&= \mathbb{I}\left\{\int P(Z_j \,|\, Z_{j'}, \mathbf{C}) P(\mathbf{C} \,|\, Z_{j'}) \,\mathrm{d}\mathbf{C} > 0\right\}\\
&= \mathbb{I}\{P(Z_j \,|\, Z_{j'}) > 0\},
\end{aligned}$$

where the third equality is due to Equation (54). ■

The intuition behind Theorem 7 is that two variables can have their support depend on each other in only two ways: (1) the causal connection between them, and (2) their common cause induces a dependence among their supports. Therefore, two causally disconnected variables must have independent support under the positivity condition, which guarantees the common cause does not causally affect their support.

Positivity is also known as the overlap condition and it is a standard assumption in causal inference with observational data (Imbens & Rubin, 2015; Pearl, 1995). It implies that the unobserved common cause $\mathbf{C}$ can induce correlations among Z_j's, though it can not affect their support. Any combination of Z_j's values must be possible. While standard, positivity may not always hold in practice. For example, suppose the latent factors Z_j's capture different human traits like skin color and skin texture. Then their unobserved common cause $\mathbf{C}$ would violate the positivity condition if it captures fine-grained genetic information; detailed genetic information will likely affect the support of latent factors, e.g., skin color. However, $\mathbf{C}$ would satisfy the positivity condition if it only captures coarse-grained information like individual probabilistic preferences due to environmental factors.

Theorem 7 shows that the absence of causal arrows between two variables implies that their support must be independent, even in the presence of unobserved confounders. In other words, the relationship between the support of the variables is robust to unobserved confounding. It implies that any disentangled $\mathbf{Z}$ must satisfy the independent support condition despite of the unobserved confounder $\mathbf{C}$.

Consider compactly supported representations: each dimension of the representations must be supported on a closed and bounded interval. Then "independent support" implies a hyperrectangular joint support. As an example, Figures 8a and 8b illustrate the scatterplots of two disentangled discrete features; they occupy rectangular regions, regardless of their correlations. Figure 8c illustrate the scatter plots of two entangled features, which occupy curvilinear, non-rectangular regions.

We note that the implications from disentanglement to independent support in Theorem 7 only hold in one direction; enforcing independent support does not guarantee the latent factors are disentangled. Further, independent support will only yield meaningful constraints on latent factors if they have bounded support; there would be no observable signal to independent and dependent support if all factors have infinite support. Generally, independent support alone is usually insufficient for identifying disentangled latent factors. One can often transform any latent factors to have independent support.

However, independent support could yield meaningful constraints if we restrict our attention to particular classes of latent factors and transformations allowed. For example, if we have only linear non-identifiability of some bounded latent factors, then independent support can be sufficient for identifying the disentangled latent factors (Ahuja et al., 2023); no causally connected factors can be transformed into independent support using linear transformations. We will discuss these considerations in Section 3.3.

Theorem 7 is most suitable for detecting entangled representations whose different dimensions causally affect each other's support. For example, in a dataset of animals, the feature "the animal being a rabbit" can affect the support of the feature "the background being the sea" because rabbits can not swim. These two features are entangled; they also violate the independent support condition. In the next sections, we will leverage these observations stemming from Theorem 7 to develop a disentanglement metric.

We conclude this section with a discussion on alternative motivations of independent support.

Is causality needed to motivate independent support? While we motivate the independent support condition from the causal perspective, this condition can be motivated from many other non-causal perspectives; e.g., the relaxation of independent factors (Roth et al., 2022), or the generalization of the classical independent component analysis to handle correlated latent factors (Comon, 1994). Moreover, the support of variables can be dependent for other reasons than the absence of causal relationships; e.g., the variables admit deterministic functional relationships among them. This fact is reflected in Theorem 7 where the implication from disentanglement to independent support goes only one way.

That said, the causal motivation of independent support aligns with the downstream applications of disentanglement. For example, in computer vision, a common task involves not only disentangling latent objects in images and videos but also understanding their causal relationships (Locatello et al., 2020b; Schölkopf et al., 2021). In this context, the causal graph of Figure 9—latent factors with unobserved common cause—serves as a starting point for these downstream causal tasks (Ahuja et al., 2023). It also aligns well with the probabilistic factor models (e.g., variational autoencoders (Kingma & Welling, 2014)) that are most commonly used for disentanglement.

Finally, Theorem 7 can be extended beyond the unobserved common cause structure in Figure 9. It can, for example, extend to causal models that involves selection bias among $Z_1, \ldots, Z_d$. In that case, one could replace the positivity condition with one that requires the selection operator does not change the latent factors' support.

3.2.2 Assessing disentanglement with the IOSS

To assess disentanglement, we build on Theorem 7 to develop the independence-of-support score (IOSS). It quantifies the extent to which a representation violates the independent support condition.

Definition 8 (Independence-of-support score (IOSS))**.** *Suppose a representation* $\mathbf{Z}$ *has bounded support and* $\sup\ Z_j\ -\ \inf\ Z_j\ >\ 0, j = 1, \ldots, d$*. Then the IOSS of* $\mathbf{Z}$ *is the Hausdorff distance* *between the joint support of $(Z_1, \ldots, Z_d)$ and the product of each individual's support:*

$$\begin{aligned}
&IOSS(Z_1, \ldots, Z_d) \\
&\triangleq d_H(supp(\bar{Z}_1, \ldots, \bar{Z}_d), supp(\bar{Z}_1) \times \cdots supp(\bar{Z}_d)) \\
&= d(supp(\bar{Z}_1) \times \cdots supp(\bar{Z}_d), supp(\bar{Z}_1, \ldots, \bar{Z}_d)),
\end{aligned}$$

where $\bar{Z}_j = (Z_j - \inf Z_j)/(\sup Z_j - \inf Z_j)$ is the standardized Z_j, and $d_H(\cdot, \cdot)$ is the Hausdorff distance.[10] *The second equality is due to $supp(Z_1, \ldots, Z_d) \subseteq supp(\bar{Z}_1) \times \cdots \times supp(\bar{Z}_d)$.*

IOSS calculates the distance between the current support of the representation $\mathbf{Z}$ and the (fictitious) support of $\mathbf{Z}$ if it were disentangled. The larger the IOSS is, the more entangled the representation is. When $\mathbf{Z}$ is disentangled, $IOSS(\mathbf{Z}) = 0$.

IOSS focuses on representations with bounded support, which allows us to standardize each Z_j to have $\inf \bar{Z}_j = 0$ and $\sup \bar{Z}_j = 1$. This standardization step makes IOSS scale-invariant; it also makes the different Z_j equivalent in their importance in IOSS.

To compute IOSS in practice, we can directly compute the support of discrete-valued representations and hence their IOSS. For continuous-valued representations, we compute the sample IOSS. Given n samples, $\{\mathbf{Z}_i\}_{i=1}^n = \{(Z_{i1}, \ldots, Z_{id})\}_{i=1}^n$, we rescale the representation to the unit hypercube,

$$\bar{Z}_{ij} = \frac{Z_{ij} - \min_{i=1,\ldots,n} Z_{ij}}{\max_{i=1,\ldots,n} Z_{ij} - \min_{i=1,\ldots,n} Z_{ij}},$$

and compute the sample IOSS as follows:

$$\widehat{IOSS}(\{\mathbf{Z}_i\}_{i=1}^n) = \max_{k=1,\ldots,K} \min_{i=1,\ldots,n} (U_k - Z_{ij})^2,$$

where $U_k \overset{iid}{\sim} \text{Unif}[0,1]^d, k = 1, \ldots, K$, are K uniform draws from $\prod_j \text{supp}(\bar{Z}_{ij})$. To improve the robustness of IOSS in practice, one may also replace the $\max, \min$ with the $(100 - \alpha)$th and αth quantile, where α is a small positive number.

The sample IOSS is comparable for representations with a fixed sample size n, dimension d, and uniform draws K. As $n \to \infty$ and $K/n \to \infty$, the sample $IOSS$ approaches IOSS almost surely, $\widehat{IOSS}(\{\mathbf{Z}_i\}_{i=1}^n) \to IOSS(\mathbf{Z})$.

To see IOSS in action, we compute the sample IOSS for the three representations in Figure 8 with $K = 10^d \cdot n$. The sample IOSS of the entangled features in Figure 8c is substantially higher than that of the disentangled features in Figures 8a and 8b. In Section 3.4, we will show that IOSS better distinguishes entangled and disentangled representations than existing disentanglement metrics.

[10] The Hausdorff distance between sets $\mathcal{X}$ and $\mathcal{Y}$ is

$$d_H(\mathcal{X}, \mathcal{Y}) = \max\{d(X, Y), d(Y, X)\},$$

where

$$\begin{aligned}
d(x, Y) &= \inf\{d(x, y) \mid y \in Y\}, \\
d(X, Y) &= \sup\{d(x, Y) \mid x \in X\}.
\end{aligned}$$

3.3 Learning Disentangled Representations with IOSS

To learn disentangled representations, we propose to use the sample IOSS as a regularizer in unsupervised representation learning.

Before presenting the algorithm, let us consider why (and when) could IOSS serve as a meaningful regularizer. To develop the intuition , we consider a toy example of a two-dimensional, compactly supported representation (Z_1, Z_2) with independent support: $Z_1 \in [1,2]$, $Z_2 \in [0,2]$. Next consider an entanglement of this representation (Z_1', Z_2'), which is a bijective transformation of (Z_1, Z_2):

$$Z_1' = Z_1 + Z_2, \qquad Z_2' = Z_1 - Z_2.$$

We will show that if (Z_1', Z_2') does not have independent support, then the support of $Z_1 - Z_2$ depends on the value of $Z_1 + Z_2$. To see why, consider the case when $Z_1 + Z_2 = 4$, then we must have $Z_1 = Z_2 = 2$ due to the support constraints on Z_1, Z_2. Hence $Z_1 - Z_2 = 0$, thus the support of $Z_1 - Z_2$ is $\{0\}$. Following a similar argument, the support of $Z_1 - Z_2$ is $\{1\}$ when $Z_1 + Z_2 = 1$. Therefore, the support of $Z_1 - Z_2$ depends on values of $Z_1 + Z_2$, and hence they have dependent support.

More broadly, the independent support condition can be useful for identifying latent factors under linear identification. Consider the setting in which we already can linearly identify the representation but cannot fully identify each latent factor in a separate coordinate. In such cases, one major challenge is to handle rotation non-identifiability in linear identification. Independent support helps in that it enforces a (hyper-)rectangular support of the latents. A rotated rectangular support is no longer rectangular; hence independent support can help resolve rotation non-identifiability.

We refer the readers to Ahuja et al. (2023, 2024) for formal identifiability results regarding the independent support condition with observational and/or interventional data. In particular, Theorem 6.3 of Ahuja et al. (2023) establishes the identifiability of representations with independent support. It shows that, given linear identification, satisfying independent support for pairwise coordinates is sufficient for identifying latent factors up to permutation, scaling, and shifting. In more detail, it focuses on representations that generate the same σ-algebra; two representations $\mathbf{Z}, \mathbf{Z}'$ satisfy $\sigma(\mathbf{Z}) = \sigma(\mathbf{Z}')$ if there exists a bijective function L such that $\mathbf{Z} = L(\mathbf{Z}')$. Representations with the same σ-algebra are "information-equivalent"; they capture the same information about the raw data $\mathbf{X}$. Among these information-equivalent representations, there is a unique representation (up to coordinate-wise bijective transformations) that has independent support. Together with "disentanglement ⇒ independent support" (Theorem 7), this unique representation must also be disentangled (if it exists).

We now illustrate disentangled representation learning with IOSS. Begin with an unsupervised representation learning objective $L(\{x_i\}_{i=1}^n\,;\,\{\mathbf{z}_i\}_{i=1}^n)$, and use the sample IOSS as a regularizer:

$$L_{IOSS}(\{x_i\}_{i=1}^n\,;\,\{\mathbf{z}_i\}_{i=1}^n) = -\underbrace{L(\{x_i\}_{i=1}^n\,;\,\{\mathbf{z}_i\}_{i=1}^n)}_{\text{representation learning loss}} + \lambda \cdot \widehat{IOSS}(\{\mathbf{z}_i\}_{i=1}^n), \tag{55}$$

As a concrete example, the representation learning objective can be the VAE objective, a popular generative modeling approach to unsupervised representation learning (Kingma & Welling, 2014):

$$L(\{x_i\}_{i=1}^n\,;\,\{\mathbf{z}_i\}_{i=1}^n) = -\mathbb{E}_{q_\theta(\mathbf{z}_i \,|\, \mathbf{x}_i)}\left[\log p_\theta(\mathbf{z}_i, \mathbf{x}_i) - \log q_\theta(\mathbf{z}_i \,|\, \mathbf{x}_i)\right]$$

where we adopt the (approximate) posterior mean of the latent variables as the representation $\mathbf{z}_i = \mathbb{E}_{q(\mathbf{z}_i \,|\, \mathbf{x}_i)}\left[\mathbf{z}_i \,|\, \mathbf{x}_i\right]$, and where $\lambda > 0$ is a regularization parameter.

To perform stochastic optimization on the L_{IOSS} objective, we subsample batches $\{\mathbf{z}_i\}_{i\in\mathcal{B}}$ of the data and update the parameters with the gradients of $L_{IOSS}(\{\mathbf{z}_i\}_{i\in\mathcal{B}})$. Algorithm 3 summarizes this algorithm for learning disentangled representations with IOSS. In Section 3.4, we will show that it can produce representations that are more disentangled.

Algorithm 3: Disentangled representation learning with IOSS

```
input  : Training data $\{\mathbf{x}_i\}_{i=1}^n$; Test data $\{\tilde{\mathbf{x}}_i\}_{i=1}^n$
output : Representations of the test data $\{\tilde{\mathbf{z}}_i\}_{i=1}^{n'}$
Minimize Equation (55) to obtain $q_\theta$;
Compute representations for the test data: $\tilde{\mathbf{z}}_i = \mathbb{E}_{q_\theta(\mathbf{z}_i \,|\, \mathbf{x}_i)}\left[\mathbf{z}_i \,|\, \mathbf{x}_i\right]$.
```

3.4 Empirical Studies of IOSS

We study IOSS in unsupervised image datasets. We find that IOSS is more effective at distinguishing between disentangled and entangled representations than other unsupervised disentanglement metrics. Unsupervised representation learning with the IOSS penalty results in representations with better disentanglement.

3.4.1 Can IOSS distinguish entangled and disentangled representations?

We first study whether IOSS can distinguish entangled and disentangled representations.

Generating entangled/disentangled representations. Focusing on datasets with correlated ground truth features, we subsample each dataset so that the correlation among the ground-truth features is > 0.8. Given the subsampled dataset, we take the ground-truth features as disentangled representations. We then generate entangled representations by applying nonlinear transformations to the ground-truth features. Given ground-truth features $Z_1, \ldots, Z_d$, we construct entangled representations by setting $Z_i' = f(Z_1, \ldots, Z_d\,;\,\theta_i) + Z_i$ with $\theta_i \sim \mathrm{Unif}([-2.5, 2.5])$, where $f(\cdot\,;\,\theta)$ is a third-order polynomial with coefficients θ.

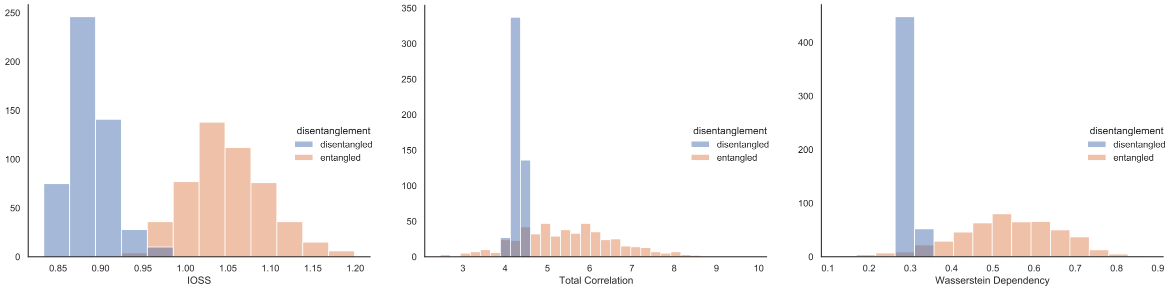
Evaluation metrics. For each pair of generated entangled and disentangled representations, we calculate different disentanglement scores for each. We report the percentage of entangled and disentangled pairs that are correctly labeled by the metric. For example, if IOSS returns a smaller value for the disentangled representation than the entangled one in the pair, then it is deemed correct labeling. Similar procedures apply to other unsupervised disentangled metrics. A metric can effectively distinguish entanglement and disentanglement if this percentage is high.

Competing methods. We compare IOSS with existing unsupervised disentanglement metrics that do not rely on ground truth features: total correlation (Chen et al., 2018) and Wasserstein dependency (Xiao & Wang, 2019). We also compared with an oracle supervised disentanglement metric: the intervention robustness score (Suter et al., 2019); it targets the same causal disentanglement definition as IOSS but relies on ground truth features.

Results. Table 3 presents the results. The IOSS outperforms baseline unsupervised disentanglement metrics in distinguishing disentanglement from entanglement. Figures 10 and 15 to 17 also show that IOSS can better separate disentangled and entangled representations than existing unsupervised disentanglement metrics.

	mpi3d	smallnorb	dsprites	cars3d
IOSS (this paper)	**0.998(0.045)**	**0.968(0.176)**	**0.980(0.140)**	0.892(0.311)
Total Correlation	0.858(0.349)	0.070(0.255)	0.162(0.369)	0.090(0.286)
Wasserstein Dependency	0.956(0.205)	0.478(0.500)	0.310(0.463)	**0.964(0.186)**
Intervention Robustness (oracle)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)

Table 3: IOSS outperforms existing disentanglement metrics in distinguishing entangled and disentangled features. The table presents the proportion of disentangled/entangled pairs that the metric correctly distinguish. Intervention Robustness Score is an oracle metric as it makes use of ground-truth features (Higher is better.)



(a) IOSS **(b)** Total Correlation **(c)** Wasserstein Dependency

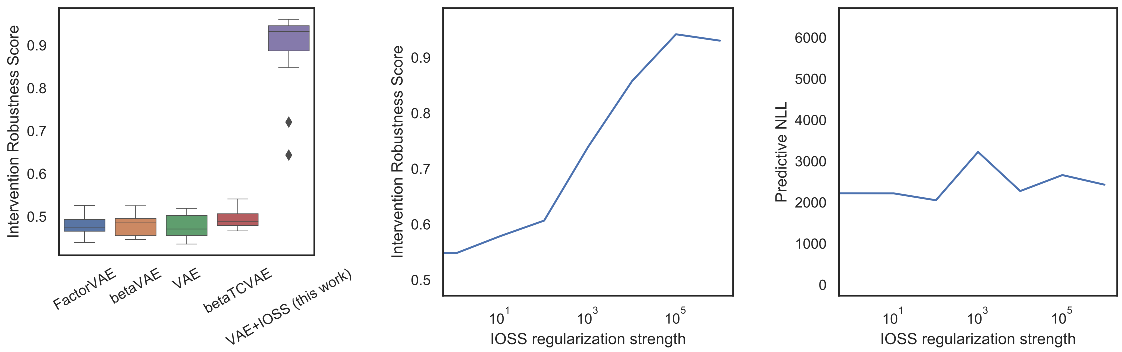
Figure 10: IOSS can better distinguish entangled and disentangled representations than existing unsupervised disentanglement metrics on the mpi3d dataset.

3.4.2 Does the IOSS regularizer encourage disentangled representations?

We next apply the IOSS penalty to learn disentangled representations via VAEs across all four datasets: mpi3d, smallnorb, dsprites, and cars3d. We work with subsampled datasets such that the ground truth features are highly correlated (with correlation ≈ 0.8) following a similar process as in Section 3.4.1. We then fit VAEs with an increasingly strong regularization with the IOSS penalty. We evaluate the disentanglement of the learned representation using the supervised disentanglement score of intervention robustness score (Suter et al., 2019); a higher score implies better disentanglement.

Competing methods. We compare VAE+IOSS with other unsupervised disentanglement algorithms including classical VAE (Kingma & Welling, 2014), FactorVAE (Kim & Mnih, 2018), betaVAE (Higgins et al., 2017), and betaTCVAE (Chen et al., 2018).

Disentanglement of VAE+IOSS-learned representations. Figure 11a shows that VAE+IOSS produce more causally disentangled representations than existing unsupervised disentanglement algorithms when ground truth features are highly correlated. These results corroborate Figures 8b and 8c. When features are highly correlated, they may still be disentangled. Hence enforcing statistical independence like beta-TCVAE and FactorVAE in this setting does not perform disentanglement. This setting violates the core assumption of beta-TCVAE and FactorVAE that the disentangled factors are independent; i.e., their joint distribution is a product of their marginals. Thus beta-TCVAE and FactorVAE do not perform as well in disentanglement with correlated features. In contrast, VAE with IOSS regularization can accommodate correlated factors, as the causal graph Figure 2 allows an unobserved common cause between the factors.



(a) Disentanglement of IOSS learned representations **(b)** Regularization with IOSS penalty **(c)** Informativeness and disentanglement tradeoff

Figure 11: (a) VAE+IOSS outperforms existing unsupervised disentanglement algorithms in producing representations that are more causally disentangled. (b) Regularizing for IOSS encourages learning disentangled representations. The intervention robustness score of the representation increases as we increase the regularization strengths of IOSS. (c) There is no obvious tradeoff between informativeness and disentanglement. The model fit stays stable despite the increase of the regularization strength of IOSS.

Moreover, Figure 11b shows that, across datasets, the VAE+IOSS-learned representation becomes increasingly disentangled as we regularize with the IOSS penalty. It implies that the IOSS penalty indeed encourages learning causally disentangled representations. Figure 18 also illustrates that regularizing for IOSS indeed encourages independent support across different dimensions of the learned representations.

Tradeoff between informativeness and disentanglement. We finally evaluate the tradeoff between informativeness and disentanglement in representation learning. We increase the regularization strengths of IOSS and evaluate the log-likelihood of the fitted VAE. Figure 11c shows that the log-likelihood stays stable despite the increasing regularization for disentanglement. While increasing the regularization strengths of IOSS encourages disentanglement, it does not compromise the informativeness of the learned representations in VAE. This result suggests the orthogonality between the disentanglement desiderata and other informativeness-related desiderata.

4. Discussion

We have shown that desiderata for representation learning, often discussed informally but rarely defined formally, can be usefully formalized from the perspective of causal inference. By studying the observable implications of our causal definitions, we develop metrics that evaluate representational desiderata and algorithms that enforce these desiderata.

We have focused on two examples of representation learning desiderata: (1) efficiency and non-spuriousness in supervised representation learning, and (2) disentanglement in unsupervised representation learning. Our overall workflow can be summarized as follows: “desiderata” → “causal definitions” → “observable implications” → “metrics and algorithms.” We have shown that this workflow can lead to practical evaluation metrics and representation learning algorithms.

Targeting the desiderata of efficiency and non-spuriousness in the supervised setting, we view the representation as a cause of the label and formalize these desiderata as the probability of necessity and sufficiency (PNS) of the cause. Studying the observable implications of PNS enables us to measure efficiency and non-spuriousness with observational data. It also allows us to formulate representation learning as a task of finding necessary and sufficient causes. We operationalize this task by developing the CAUSAL-REP algorithm. En route, we develop identification results for the PNS given high-dimensional (rank-degenerate) data.

To encourage unsupervised disentanglement, we again begin with a causal definition, and we obtain metrics for disentanglement that again rely on studying its observable implications. We capitalize on the observation that causally disentangled variables—which have no causal connections among each other—must have independent support under a standard positivity condition. Based on this insight, we develop the independence-of-support score (IOSS) for assessing causal disentanglement. We further establish the identifiability of representations with independent support, which enables representation learning with an IOSS penalty.

For future work, one promising direction is to identify other desiderata in representation learning (or machine learning tasks in general) that may be formalized using causal notions. One may follow the workflow to obtain evaluation metrics and learning algorithms for the desiderata.

Another direction is to extend the metrics and algorithms developed in this work. One may extend the CAUSAL-REP algorithm to general anti-causal learning, without requiring the assumptions (e.g., pinpointability) required in Proposition 1. Such an extension would enable the treatment of reverse causal inference via probabilities of causation. On the disentanglement side, IOSS can also be extended to learn primitives for compositional generalization. Primitives share similar properties as disentangled features; they cannot causally affect each other. In this sense, IOSS-type ideas can potentially offer new approaches to compositional generalization.

Acknowledgments.

We thank David Blei, Gemma Moran, Shin Oblander, and Eli Weinstein for inspiring discussions and helpful feedback. This work was supported in part by the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941, the Office of Naval Research under grant number N00014-23-1-2590, the National Science Foundation under Grant No. 2231174, No. 2310831, No. 2428059, and a Michigan Institute for Data Science Propelling Original Data Science (PODS) grant.

References

Achille, A. & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(1), 1947–1980.

Ahuja, K., Mahajan, D., Wang, Y., & Bengio, Y. (2023). Interventional causal representation learning. In *International Conference on Machine Learning (ICML)* (pp. 372–407).: PMLR.

Ahuja, K., Mansouri, A., & Wang, Y. (2024). Multi-domain causal representation learning via weak distributional invariances. *Artificial Intelligence and Statistics (AISTATS)*.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9, 1981–2014.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bai, J. & Li, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *Review of Economics and Statistics*, 98(2), 298–309.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Bouchacourt, D., Tomioka, R., & Nowozin, S. (2018). Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI Conference on Artificial Intelligence*, volume 32.

Burgess, C. P., Higgins, I., et al. (2018). Understanding disentangling in beta-VAE. *arXiv preprint arXiv:1804.03599*.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, 44(1), 137–164.

Chalupka, K., Perona, P., & Eberhardt, F. (2015). Visual causal feature learning. *Uncertainty in Artificial Intelligence (UAI)*, (pp. 181–190).

Chen, R. T., Li, X., Grosse, R., & Duvenaud, D. (2018). Isolating sources of disentanglement in VAEs. In *Advances in Neural Information Processing Systems* (pp. 2615–2625).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)* (pp. 1597–1607).

Chen, Y., Li, X., & Zhang, S. (2020b). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532), 1756–1770.

Cheng, P. W. & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing. *The Oxford Handbook of Causal Reasoning*, (pp.$\tilde{6}$5).

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3), 287–314.

Correa, J. & Bareinboim, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (pp. 10093–10100).

Creager, E., Jacobsen, J.-H., & Zemel, R. (2021). Environment inference for invariant learning. In *International Conference on Machine Learning (ICML)* (pp. 2189–2200).: PMLR.

D'Amour, A. (2019a). Comment: Reflections on the deconfounder. *Journal of the American Statistical Association*, 114(528), 1597–1601.

D'Amour, A. (2019b). On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *Artificial Intelligence and Statistics (AISTATS)* (pp. 3478–3486).

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2020a). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221, 644–654.

D'Amour, A., Heller, K., et al. (2020b). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6), 141–142.

Eberhardt, F. & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74(5), 981–995.

Erosheva, E. A. & Fienberg, S. E. (2005). Bayesian mixed membership models for soft clustering and classification. In C. Weihs & W. Gaul (Eds.), *Classification—The Ubiquitous Challenge* (pp. 11–26). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fong, C. & Grimmer, J. (2016). Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1600–1609).

Galhotra, S., Pradhan, R., & Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 577–590).

Gelman, A. & Imbens, G. (2013). *Why ask why? Forward causal inference and reverse causal questions*. Technical report, National Bureau of Economic Research.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, (pp. 733–760).

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. CRC Press.

Golub, G., Klema, V., & Stewart, G. W. (1976). *Rank degeneracy and least squares problems.* Technical report, Stanford University, Department of Computer Science.

Goodfellow, I. J., Pouget-Abadie, J., et al. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661.*

Grimmer, J. & Fong, C. (2021). Causal inference with latent treatments. *American Journal of Political Science.*

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2 (pp. 1735–1742).: IEEE.

Heinze-Deml, C., Maathuis, M. H., & Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, 5, 371–391.

Higgins, I., Matthey, L., et al. (2017). Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations.*

Hosoya, H. (2019). Group-based learning of disentangled representations with generalizability for novel contents. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2506–2513).

Imai, K. & Jiang, Z. (2019). Comment: The challenges of multiple causes. *Journal of the American Statistical Association*, 114(528), 1605–1610.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Janzing, D. & Schölkopf, B. (2015). Semi-supervised interpolation in an anticausal learning scenario. *Journal of Machine Learning Research*, 16(1), 1923–1948.

Johansson, F. D., Shalit, U., Kallus, N., & Sontag, D. (2020). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426.*

Johansson, F. D., Sontag, D., & Ranganath, R. (2019). Support and invertibility in domain-invariant representations. In *Artificial Intelligence and Statistics (AISTATS)* (pp. 527–536).

Khasanova, R. & Frossard, P. (2017). Graph-based isometry invariant representation learning. In *International Conference on Machine Learning (ICML)* (pp. 1847–1856).

Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020). Variational autoencoders and nonlinear ICA: A unifying framework. In *Artificial Intelligence and Statistics (AISTATS)* (pp. 2207–2217).

Kilbertus, N., Parascandolo, G., & Schölkopf, B. (2018). Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524.*

Kim, H. & Mnih, A. (2018). Disentangling by factorising. In *International Conference on Machine Learning (ICML)* (pp. 2649–2658).

Kingma, D. P. & Welling, M. (2014). Auto-encoding variational Bayes. *International Conference on Learning Representations.*

Kommiya Mothilal, R., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 652–663).

Kumar, A., Sattigeri, P., & Balakrishnan, A. (2018). Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Locatello, F., Bauer, S., et al. (2019a). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)* (pp. 4114–4124).

Locatello, F., Poole, B., et al. (2020a). Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning (ICML)* (pp. 6348–6359).

Locatello, F., Tschannen, M., et al. (2019b). Disentangling factors of variations using few labels. In *International Conference on Learning Representations*.

Locatello, F., Weissenborn, D., et al. (2020b). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33, 11525–11538.

Lu, C., Wu, Y., Hernández-Lobato, J. M., & Schölkopf, B. (2021). Invariant causal representation learning.

McLachlan, G. J. & Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, volume 38. M. Dekker New York.

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., & Blundell, C. (2020). Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*.

Moraffah, R., Shu, K., Raglin, A., & Liu, H. (2019). Deep causal representation learning for unsupervised domain adaptation. *arXiv preprint arXiv:1910.12417*.

Moyer, D., Gao, S., Brekelmans, R., Steeg, G. V., & Galstyan, A. (2018). Invariant representations without adversarial training. *arXiv preprint arXiv:1805.09458*.

Mueller, S., Li, A., & Pearl, J. (2021). Causes of effects: Learning individual responses from population data. *arXiv preprint arXiv:2104.13730*.

Nabi, R., McNutt, T., & Shpitser, I. (2020). Semiparametric causal sufficient dimension reduction of high dimensional treatments. *arXiv preprint arXiv:1710.06727*.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2018). Learning independent causal mechanisms. In *International Conference on Machine Learning (ICML)* (pp. 4036–4044).

Paul, M. (2017). Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 163–172).

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688.

Pearl, J. (2011). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2019a). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60.

Pearl, J. (2019b). Sufficient causes: On oxygen, matches, and fires. *Journal of Causal Inference*, 7(2), 1–8.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.

Pryzant, R., Card, D., Jurafsky, D., Veitch, V., & Sridhar, D. (2020). Causal effects of linguistic properties. *arXiv preprint arXiv:2010.12919.*

Puli, A., Perotte, A., & Ranganath, R. (2020). Causal estimation with functional confounders. *Advances in Neural Information Processing Systems*, 33.

Puli, A., Zhang, L. H., Oermann, E. K., & Ranganath, R. (2021). Predictive modeling in the presence of nuisance-induced spurious correlations. *arXiv preprint arXiv:2107.00520.*

Ranganath, R. & Perotte, A. (2018). Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273.*

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning (ICML)*, volume 2.

Roth, K., Ibrahim, M., Akata, Z., Vincent, P., & Bouchacourt, D. (2022). Disentanglement of correlated factors via Hausdorff factorized support. *arXiv preprint arXiv:2210.07347.*

Schölkopf, B., Janzing, D., et al. (2012). On causal and anticausal learning. In *International Conference on Machine Learning (ICML)* (pp. 1255–1262).

Schölkopf, B., Janzing, D., et al. (2013). Semi-supervised learning in causal and anticausal settings. In *Empirical Inference* (pp. 129–141). Springer.

Schölkopf, B., Locatello, F., et al. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634.

Shen, X., Liu, F., et al. (2020). Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637.*

Shi, C., Veitch, V., & Blei, D. (2020). Invariant representation learning for treatment effect estimation. *arXiv preprint arXiv:2011.12379.*

Shu, R., Chen, Y., Kumar, A., Ermon, S., & Poole, B. (2019). Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations.*

Stewart, G. (1984). Rank degeneracy. *SIAM Journal on Scientific and Statistical Computing*, 5(2), 403–413.

Suter, R., Miladinovic, D., Schölkopf, B., & Bauer, S. (2019). Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning (ICML)* (pp. 6056–6065).

Thomas, V., Bengio, E., et al. (2018). Disentangling the independently controllable factors of variation by interacting with the world. *arXiv preprint arXiv:1802.09484*.

Thomas, V., Pondard, J., et al. (2017). Independently controllable factors. *arXiv preprint arXiv:1708.01289*.

Tian, J. & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1), 287–313.

Tipping, M. E. & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611–622.

Träuble, F., Creager, E., et al. (2020). Is independence all you need? On the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*.

Udell, M. & Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1), 144–160.

Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021). Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Veitch, V., Sridhar, D., & Blei, D. (2020). Adapting text embeddings for causal inference. In *Uncertainty in Artificial Intelligence (UAI)* (pp. 919–928).

Veitch, V., Wang, Y., & Blei, D. M. (2019). Using embeddings to correct for unobserved confounding in networks. *arXiv preprint arXiv:1902.04114*.

Wang, H., Lu, Y., & Zhai, C. (2010). Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 783–792).

Wang, H., Lu, Y., & Zhai, C. (2011). Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 618–626).

Wang, Y. & Blei, D. (2021). A proxy variable view of shared confounding. In *International Conference on Machine Learning (ICML)* (pp. 10697–10707).: PMLR.

Wang, Y. & Blei, D. M. (2019a). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528), 1574–1596.

Wang, Y. & Blei, D. M. (2019b). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527), 1147–1161.

Wang, Y. & Blei, D. M. (2020). Towards clarifying the theory of the deconfounder. *arXiv preprint arXiv:2003.04948*.

Wang, Z. & Culotta, A. (2020). Identifying spurious correlations for robust text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 3431–3440).

Wang, Z. & Culotta, A. (2021). Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (pp. 14024–14031).

Watson, D., Gultchin, L., Taly, A., & Floridi, L. (2021). Local explanations via necessity and sufficiency: unifying theory and practice. *arXiv preprint arXiv:2103.14651*.

Weichwald, S., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2014). Causal and anti-causal learning in pattern recognition for neuroimaging. In *2014 International Workshop on Pattern Recognition in Neuroimaging* (pp. 1–4).

Wu, P. & Fukumizu, K. (2021). Identifying treatment effects under unobserved confounding by causal representation learning. *arXiv preprint arXiv:2101.06662*.

Xiao, Y. & Wang, W. Y. (2019). Disentangled representation learning with Wasserstein total correlation. *arXiv preprint arXiv:1912.12818*.

Yang, M., Liu, F., et al. (2020). CausalVAE: Disentangled representation learning via neural structural causal models. *arXiv preprint arXiv:2004.08697*.

Zhao, H., Des Combes, R. T., Zhang, K., & Gordon, G. (2019). On learning invariant representations for domain adaptation. In *International Conference on Machine Learning (ICML)* (pp. 7523–7532).

Supplementary Materials

Appendix A. Proof of Theorem 1

Proof The proof generalizes the proof of Theorem 9.2.10 in Pearl (2011).

We first notice the following:

$$P(Y(\mathbf{Z}=\mathbf{z})=y) = P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})=y) + P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})\neq y) \tag{56}$$

Next denote $\epsilon \triangleq P(Y(\mathbf{Z}=\mathbf{z})\neq y, Y(\mathbf{Z}\neq\mathbf{z})=y) \geq 0$. Then we have

$$\begin{align} P(Y(\mathbf{Z}\neq\mathbf{z})=y) &= P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})=y) + P(Y(\mathbf{Z}=\mathbf{z})\neq y, Y(\mathbf{Z}\neq\mathbf{z})=y) \tag{57}\\ &= P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})=y) + \epsilon. \tag{58} \end{align}$$

Substituting Equation (58) into Equation (56) implies that

$$P(Y(\mathbf{Z}=\mathbf{z})=y) = P(Y(\mathbf{Z}\neq\mathbf{z})=y) + P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})\neq y) - \epsilon, \tag{59}$$

which implies

$$\begin{align} P(Y(\mathbf{Z}=\mathbf{z})=y, Y(\mathbf{Z}\neq\mathbf{z})\neq y) =& P(Y(\mathbf{Z}=\mathbf{z})=y) - P(Y(\mathbf{Z}\neq\mathbf{z})=y) + \epsilon \tag{60}\\ =& P(Y=y \mid \mathrm{do}(\mathbf{Z}=\mathbf{z})) - P(Y=y \mid \mathrm{do}(\mathbf{Z}\neq\mathbf{z})) + \epsilon \tag{61}\\ \geq& P(Y=y \mid \mathrm{do}(\mathbf{Z}=\mathbf{z})) - P(Y=y \mid \mathrm{do}(\mathbf{Z}\neq\mathbf{z})). \tag{62} \end{align}$$

The inequality becomes equality when $\epsilon = 0$ under the monotonicity condition in Theorem 1.

■

Appendix B. The definition of functional interventions recovers backdoor adjustment

The definition of functional interventions (Definition 4) recovers the standard backdoor adjustment as special cases when the function $f(\mathbf{X})$ returns $\mathbf{X}$ or its subset (assuming the causal graph Figure 2).

For example, when $f(\mathbf{X}) = \mathbf{X}$, then

$$\begin{aligned} &P(Y \mid \mathrm{do}(f(\mathbf{X})=\mathbf{z})) = P(Y \mid \mathrm{do}(\mathbf{X}=\mathbf{z})) \\ &= \int P(Y \mid \mathrm{do}(\mathbf{X}), \mathbf{C}) P(\mathbf{X} \mid \mathbf{C}, \mathbf{X}=\mathbf{z}) P(\mathbf{C}) \,\mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{C} \\ &= \int P(Y \mid \mathbf{X}) \cdot \delta_{\mathbf{X}=\mathbf{z}} \cdot P(\mathbf{C}) \,\mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{C} \\ &= P(Y \mid \mathbf{X}=\mathbf{z}), \end{aligned}$$

where the third inequality is due to no unobserved confounding between $\mathbf{X}$ and Y in Figure 2.

As another special case, suppose the function $f(\mathbf{X})$ returns the subset of $\mathbf{X} = (X_1, \dots, X_m)$ except X_1, i.e., $f(\mathbf{X}) = (X_2, \dots, X_m)$. Then Equation (19) recovers backdoor adjustment:

$$\begin{aligned}
&P(Y \,|\, \mathrm{do}(f(\mathbf{X}) = \mathbf{z})) = P(Y \,|\, \mathrm{do}((X_2, \dots, X_m) = \mathbf{z})) \\
&= \int P(Y \,|\, \mathrm{do}(\mathbf{X}), \mathbf{C}) P(\mathbf{X} \,|\, \mathbf{C}, (X_2, \dots, X_m) = \mathbf{z}) P(\mathbf{C}) \,\mathrm{d}\mathbf{X} \,\mathrm{d}\mathbf{C} \\
&= \int P(Y \,|\, \mathbf{X}) \cdot \delta_{(X_2, \dots, X_m) = \mathbf{z}} \cdot P(X_1 \,|\, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}\mathbf{X} \,\mathrm{d}\mathbf{C} \\
&= \int P(Y \,|\, \mathbf{X}_1, (X_2, \dots, X_m) = \mathbf{z}) \cdot P(X_1 \,|\, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}X_1 \,\mathrm{d}\mathbf{C} \\
&= \int P(Y \,|\, \mathbf{X}_1, (X_2, \dots, X_m) = \mathbf{z}, \mathbf{C}) \cdot P(X_1 \,|\, (X_2, \dots, X_m) = \mathbf{z}, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}X_1 \,\mathrm{d}\mathbf{C} \\
&= \int P(Y, \mathbf{X}_1 \,|\, (X_2, \dots, X_m) = \mathbf{z}, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}X_1 \,\mathrm{d}\mathbf{C} \\
&= \int P(Y \,|\, (X_2, \dots, X_m) = \mathbf{z}, \mathbf{C}) P(\mathbf{C}) \,\mathrm{d}\mathbf{C},
\end{aligned}$$

where the third inequality is due to no unobserved confounding between $\mathbf{X}$ and Y in Figure 2, and the fifth inequality is due to the conditional independence of $X_1, \dots, X_m$ given $\mathbf{C}$ in Figure 2.

Appendix C. Proof of Proposition 1

Proof

We calculate the intervention distribution for functional interventions $f(\mathbf{X}) = \tilde{f}((X_j)_{j \in S})$

$$\begin{aligned}
&P(Y \,|\, \mathrm{do}(f(\mathbf{X}))) \\
&= \int P(Y \,|\, (X_j)_{j \in S}, \mathbf{C}) P((X_j)_{j \in S} \,|\, \mathbf{C}, f(\mathbf{X})) P(\mathbf{C}) \,\mathrm{d}(X_j)_{j \in S} \,\mathrm{d}\mathbf{C}, && (63) \\
&= \int P(Y \,|\, (X_j)_{j \in S}, \mathbf{C}, f(\mathbf{X})) P((X_j)_{j \in S} \,|\, \mathbf{C}, f(\mathbf{X})) P(\mathbf{C}) \,\mathrm{d}(X_j)_{j \in S} \,\mathrm{d}\mathbf{C}, && (64) \\
&= \int P(Y \,|\, \mathbf{C}, f(\mathbf{X})) P(\mathbf{C}) \,\mathrm{d}\mathbf{C}, && (65) \\
&= \int P(Y \,|\, f(\mathbf{X}), h(\mathbf{X})) \cdot P(h(\mathbf{X})) \,\mathrm{d}h(\mathbf{X}). && (66)
\end{aligned}$$

The first equation is due to the definition of functional interventions on $f(\mathbf{X})$ (Puli et al., 2020). It is a soft intervention on $\mathbf{X}$ with a stochastic policy conditional on the parents of $\mathbf{X}$. The stochastic policy is $P(\mathbf{X} \,|\, f(\mathbf{X}), PA(\mathbf{X}))$ where $PA(\mathbf{X})$ denotes the parents of $\mathbf{X}$.

The second equation is due to $P(Y \,|\, (X_j)_{j \in S}, \mathbf{C}, f(\mathbf{X})) = P(Y \,|\, (X_j)_{j \in S}, \mathbf{C})$.

The third equation is due to the observability and positivity condition. Observability and positivity guarantee that $P((X_j)_{j \in S}, \mathbf{C}) > 0$ for all $(X_j)_{j \in S}, \mathbf{C}$. Thus we can calculate $P(Y \,|\, (X_j)_{j \in S}, \mathbf{C}, f(\mathbf{X})) = P(Y \,|\, (X_j)_{j \in S}, \mathbf{C})$. Moreover, the two conditions imply that $P(f(\mathbf{X}), \mathbf{C}) > 0$ for all $f(\mathbf{X}), \mathbf{C}$, and hence we can calculate $P((X_j)_{j \in S} \,|\, \mathbf{C}, f(\mathbf{X}))$.

The fourth equation is due to the pinpointability condition.

■

Appendix D. Pinpointing the unobserved common cause C

The CAUSAL-REP algorithm begins with a step of pinpointing the unobserved common cause $\mathbf{C}$. In this step, we infer $\mathbf{C}$ from the observational data $\mathbf{X}$, when $\mathbf{C}$ is low-dimensional and $\mathbf{X}$ is high-dimensional. Operationally, as $\mathbf{C}$ renders $\mathbf{X} = (X_1, \ldots, X_m)$ conditionally independent, we infer $\mathbf{C}$ by fitting a probabilistic factor model to $\mathbf{X}$,

$$p(\mathbf{x}_i, \mathbf{c}_i\,;\,\phi) = p(x_{i1}, \ldots, x_{im}, \mathbf{c}_i\,;\,\phi) = p(\mathbf{c}_i)\prod_{j=1}^{m} p(x_{ij}\,|\,\mathbf{c}_i\,;\,\phi). \tag{67}$$

Specifically, we consider VAE, a flexible probabilistic factor model (Kingma & Welling, 2014),

$$\mathbf{C}_i \sim p(\mathbf{c}_i), \tag{68}$$
$$\mathbf{X}_i\,|\,\mathbf{C}_i \sim p(\mathbf{x}_i\,|\,\mathbf{c}_i\,;\,\theta) = \mathrm{EF}(\mathbf{x}_i\,|\,f_\theta(\mathbf{c}_i)\,;\,\lambda_\theta), \tag{69}$$

where EF is an exponential family distribution, $\theta = (f_\theta, \lambda_\theta)$ are the parameters, and $f_\theta : \mathbf{C} \to \mathbf{X}$ is a flexible neural network.

Next we infer $p(\mathbf{c}_i\,|\,\mathbf{x}_i)$ using variational approximation $p(\mathbf{c}_i\,|\,\mathbf{x}_i) \approx q_{\phi^*}(\mathbf{c}_i\,|\,\mathbf{x}_i)$ where $q_{\phi^*}(\mathbf{c}\,|\,\mathbf{x})$ maximizes the evidence lower bound (ELBO) objective of VAE (Blei et al., 2017),

$$q_{\phi^*} = \arg\max_{q_\phi} \sum_{i=1}^{n} \left[\log p(\mathbf{x}_i, \mathbf{c}_i) - \log q_\phi(\mathbf{c}_i\,|\,\mathbf{x}_i)\right],$$

and $q_\phi(\cdot\,|\,\mathbf{x})$ parametrizes a flexible family of distributions with parameters ϕ. For example, we can have $q_\phi(\cdot\,|\,\mathbf{x}) = \mathcal{N}(Z_{1,\phi}(\mathbf{x}), Z_{2,\phi}(\mathbf{x}) \cdot I)$ or even more flexible non-Gaussian distributions via normalizing flows.

Finally, we set

$$\mathbf{C}_i = h(\mathbf{X}_i) \approx \mathbb{E}_{q_{\phi^*}(\mathbf{c}_i\,|\,\mathbf{x}_i)}\left[\mathbf{C}_i\,|\,\mathbf{X}_i\right], i = 1, \ldots, n. \tag{70}$$

Though we use variational approximation for $p(\mathbf{c}_i\,|\,\mathbf{x}_i)$, Equation (70) can often give a good approximation of $\mathbf{C}_i$ when $\dim(\mathbf{X}) \gg \dim(\mathbf{C})$, or more precisely when the pinpointability condition holds (Chen et al., 2020b; Wang & Blei, 2019b).

Appendix E. Calculating PNS lower bounds with linear models

We calculate the lower bound of the conditional efficiency and non-spuriousness $\underline{PNS_n(f_j(\mathbf{X}), Y\,|\,f_{-j}(\mathbf{X}))}$ with the linear model (Equation (30)):

$$\begin{aligned}
&\underline{PNS_n(f_j(\mathbf{X}), Y\,|\,f_{-j}(\mathbf{X}))}\\
&= \prod_{i=1}^{n}\int \big[P(Y = y_i\,|\,f_j(\mathbf{X}) = f_j(\mathbf{x}_i), f_{-j}(\mathbf{X}) = f_{-j}(\mathbf{x}_i), \mathbf{C})\\
&\qquad - P(Y = y_i\,|\,f_j(\mathbf{X}) \neq f_j(\mathbf{x}_i), f_{-j}(\mathbf{X}) = f_{-j}(\mathbf{x}_i), \mathbf{C})\big] \cdot P(\mathbf{C})\,\mathrm{d}\mathbf{C}\\
&= \prod_{i=1}^{n}\int \Bigg[\mathcal{N}(y_i\,;\,\beta_0 + \beta_j f_j(\mathbf{x}_i) + \sum_{j' \neq j}\beta_{j'} f_{j'}(\mathbf{x}_i) + \boldsymbol{\gamma}^\top\mathbf{C}, \sigma^2)
\end{aligned}$$

$$-\mathcal{N}(y_i\,;\,\beta_0+\beta_j\mathbb{E}\left[f_j(\mathbf{x}_i)\right]+\sum_{j'\neq j}\beta_{j'}f_{j'}(\mathbf{x}_i)+\boldsymbol{\gamma}^\top\mathbf{C},\sigma^2)\Bigg]\cdot P(\mathbf{C})\,\mathrm{d}\mathbf{C}$$

$$=\prod_{i=1}^{n}\int\Bigg[\exp\left(-\frac{(\gamma^\top(\mathbf{c}_i-\mathbf{C})+\epsilon_i)^2}{2\sigma^2}\right)$$

$$-\exp\left(-\frac{(\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])+\gamma^\top(\mathbf{c}_i-\mathbf{C})+\epsilon_i)^2}{2\sigma^2}\right)\Bigg]\cdot P(\mathbf{C})\,\mathrm{d}\mathbf{C}\times(2\pi\sigma^2)^{-\frac{n}{2}} \tag{71}$$

$$\approx\prod_{i=1}^{n}\int\Bigg[\left(1-\frac{(\gamma^\top(\mathbf{c}_i-\mathbf{C})+\epsilon_i)^2}{2\sigma^2}\right)$$

$$-\left(1-\frac{(\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])+\gamma^\top(\mathbf{c}_i-\mathbf{C})+\epsilon_i)^2}{2\sigma^2}\right)\Bigg]\cdot P(\mathbf{C})\,\mathrm{d}\mathbf{C}\times(2\pi\sigma^2)^{-\frac{n}{2}} \tag{72}$$

$$=\prod_{i=1}^{n}\int\left(\frac{(\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2+2\cdot\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])\cdot(\gamma^\top(\mathbf{c}_i-\mathbf{C})+\epsilon_i)}{2\sigma^2}\right)$$

$$\cdot P(\mathbf{C})\,\mathrm{d}\mathbf{C}\times(2\pi\sigma^2)^{-\frac{n}{2}} \tag{73}$$

$$=\prod_{i=1}^{n}\left(\frac{(\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2+2\cdot\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])\cdot(\gamma^\top(\mathbf{c}_i-\mathbb{E}\left[\mathbf{C}\right])+\epsilon_i)}{2\sigma^2}\right)$$

$$\times(2\pi\sigma^2)^{-\frac{n}{2}} \tag{74}$$

$$=\prod_{i=1}^{n}\exp\left(\frac{(\beta_j(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2+2\beta_j(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])(\gamma^\top(\mathbf{c}_i-\mathbb{E}\left[\mathbf{C}\right])+\epsilon_i)}{2\sigma^2}-1\right)$$

$$\times(2\pi\sigma^2)^{-\frac{n}{2}} \tag{75}$$

$$=\exp\left(\frac{\sum_{i=1}^{n}(\beta_j(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2+2\sum_{i=1}^{n}\beta_j(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])(\gamma^\top(\mathbf{c}_i-\mathbb{E}\left[\mathbf{C}\right]))}{2\sigma^2}-n\right)$$

$$\times(2\pi\sigma^2)^{-\frac{n}{2}}, \tag{76}$$

where $\mathcal{N}(\cdot)$ denotes the Gaussian density, $\epsilon_i=y_i-(\beta_0+\boldsymbol{\beta}^\top f(\mathbf{x}_i)+\boldsymbol{\gamma}^\top\mathbf{c}_i)$ is the residual of the regression in Equation (30). Equations (72) and (75) make use of Taylor approximation $\exp(x)\approx 1+x$. Equation (76) makes use of $\sum_{i=1}^{n}\epsilon_i\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])\approx\mathbb{E}\left[\epsilon\cdot(f_j(\mathbf{X})-\mathbb{E}\left[f_j(\mathbf{X})\right])\right]=0$ in the regression model.

Finally, in CAUSAL-REP Algorithms 2 and 4, we often impose an R-squared penalty to encourage the possitivity of $f_j(\mathbf{X})$ given $\mathbf{C}$. In these cases, we often have $\sum_{i=1}^{n}\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right])\cdot(\gamma^\top(\mathbf{c}_i-\mathbb{E}\left[\mathbf{C}\right]))\approx\beta_j\mathrm{Cov}(f_j(\mathbf{X}),\gamma^\top\mathbf{C})\approx 0$. Thus, we can further approximate the PNS by

$$\underline{PNS_n(f_j(\mathbf{X}),Y\,|\,f_{-j}(\mathbf{X}))}$$

$$\approx\exp\left(\frac{\sum_{i=1}^{n}(\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2}{2\sigma^2}-n\right)\times(2\pi\sigma^2)^{-\frac{n}{2}}, \tag{77}$$

and thus,

$$\log\underline{PNS_n(f_j(\mathbf{X}),Y\,|\,f_{-j}(\mathbf{X}))}\approx\left(\frac{\sum_{i=1}^{n}(\beta_j\cdot(f_j(\mathbf{x}_i)-\mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2}{2\sigma^2}\right)+\text{constant}, \tag{78}$$

Similarly, we can obtain the lower bound of the (unconditional) efficiency and non-spuriousness, similar to Equation (76),

$$\log\underline{PNS_n(f(\mathbf{X}),Y)}$$

Algorithm 4: CAUSAL-REP (Unsupervised)

```
input  : The observational training data (without labels) $\{\mathbf{x}_i\}_{i=1}^n$; the probabilistic factor model that
         generates the training data $P(\mathbf{X}, \mathbf{C})$
output : CAUSAL-REP representation function $\hat{f}(\cdot)$
Augment the unsupervised training dataset into a supervised one $\{\{\mathbf{x}_i^u, \mathbf{y}_i^u\}_{u=1}^U\}_{i=1}^n$ following
  Equation (46);
Fit a probabilistic factor model (Equations (28) and (29)) and infer $\{\{p(\mathbf{c}_i^u \,|\, \mathbf{x}_i^u)\}_{u=1}^U\}_{i=1}^n$;
if Pinpointability holds, i.e. $p(\mathbf{c}_i^u \,|\, \mathbf{x}_i^u)$ is close to a point mass for all $i, u$ then
    foreach training datapoint $i$ do
        Pinpoint the unobserved common cause $\mathbf{C}$: $\mathbf{c}_i^u = h(\mathbf{x}_i^u) \triangleq \mathbb{E}\left[\mathbf{c}_i^u \,|\, \mathbf{x}_i^u\right]$ for all $u = 1, \ldots, U$;
    end
    Maximize Equation (47) to obtain the CAUSAL-REP representation $\hat{f}$;
end
```

$$\approx \left(\frac{1}{2\sigma^2} \sum_{i=1}^n \left[(\sum_{j=1}^d \beta_j \cdot (f_j(\mathbf{x}_i) - \mathbb{E}\left[f_j(\mathbf{x}_i)\right]))^2 \right.\right.$$
$$\left.\left. +2 \cdot \sum_{j=1}^d \beta_j \cdot (f_j(\mathbf{x}_i) - \mathbb{E}\left[f_j(\mathbf{x}_i)\right]) \cdot \gamma^\top (\mathbf{c}_i - \mathbb{E}\left[\mathbf{C}\right]) \right] \right) + \text{constant}. \tag{79}$$

Appendix F. Details of the unsupervised CAUSAL-REP algorithm

We summarize the unsupervised CAUSAL-REP in Algorithm 4.

Appendix G. Details of the empirical studies for CAUSAL-REP and additional empirical results

Across all experiments, we used the Adam optimizer and learning rate 0.01. For text experiments, we consider a bag-of-words representation with standard stop words removal with the NLTK package. To aid reproducibility, we have included source code that can reproduce the results at `https://github.com/yixinwang/representation-causal-public`.

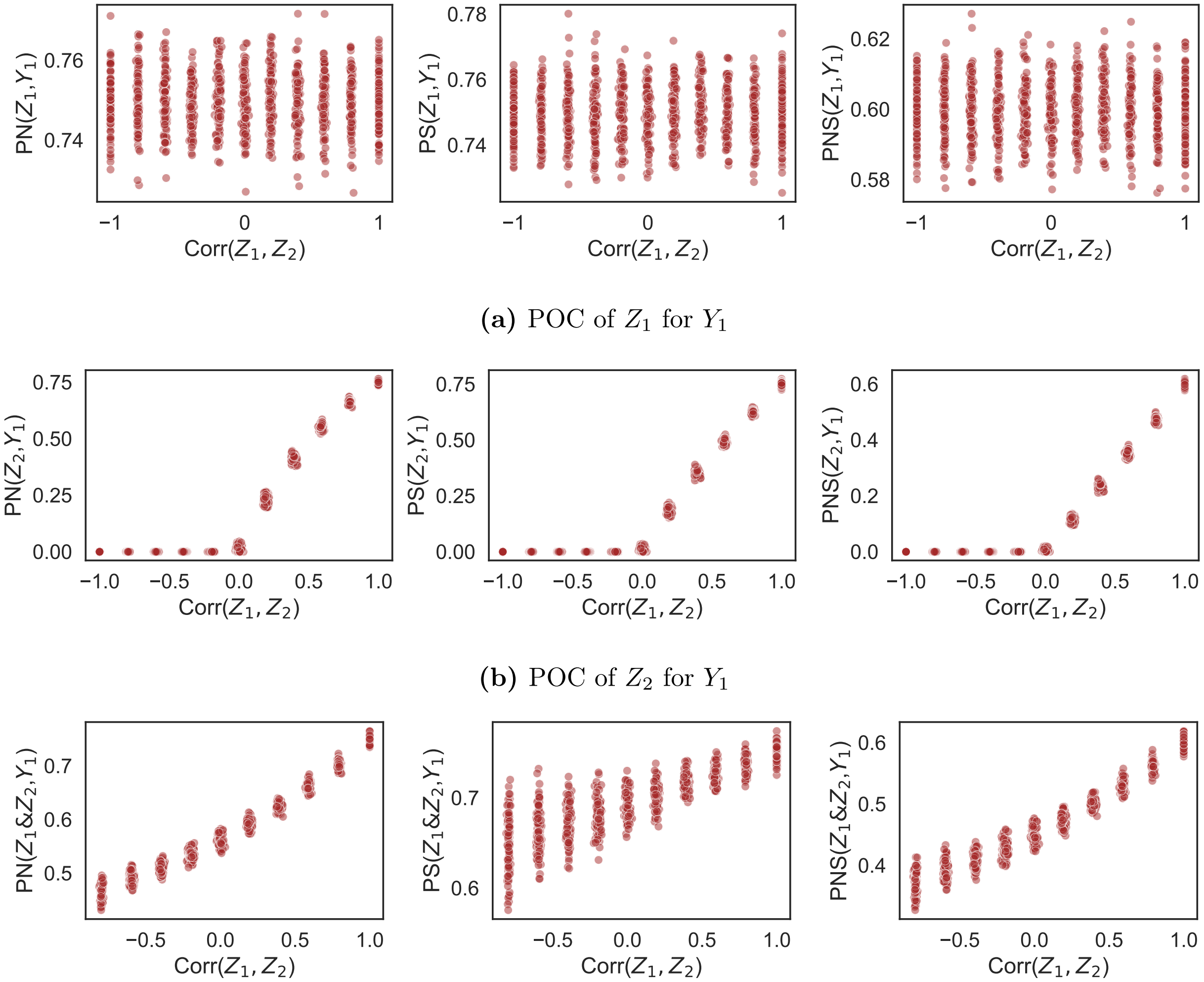
G.1 Details of Section 2.5.1

Figures 12 and 13 present additional results for Section 2.5.1. As Z_1 and Z_2 become increasingly correlated, the (lower bounds of) unconditional POC of Z_2 for Y_1 also increase. It is also consistent with the intuition: Z_2 is an increasingly better surrogate of Z_1 for Y_1 give higher correlations between Z_1 and Z_2.

G.2 Details of the colored MNIST study

We study CAUSAL-REP on the colored MNIST dataset (Arjovsky et al., 2019). The dataset builds on the original MNIST data but color the image in a way that is highly correlated with the digits label. We focus on the digits '3' and '8' and colors 'red' and 'green.'

To create a training set, we color the '3' images in red with probability p and in green with probability $(1 - p)$. Next, we color the '8' images in red with probability $(1 - p)$ and in green with probability p.



(a) POC of Z_1 for Y_1

(b) POC of Z_2 for Y_1

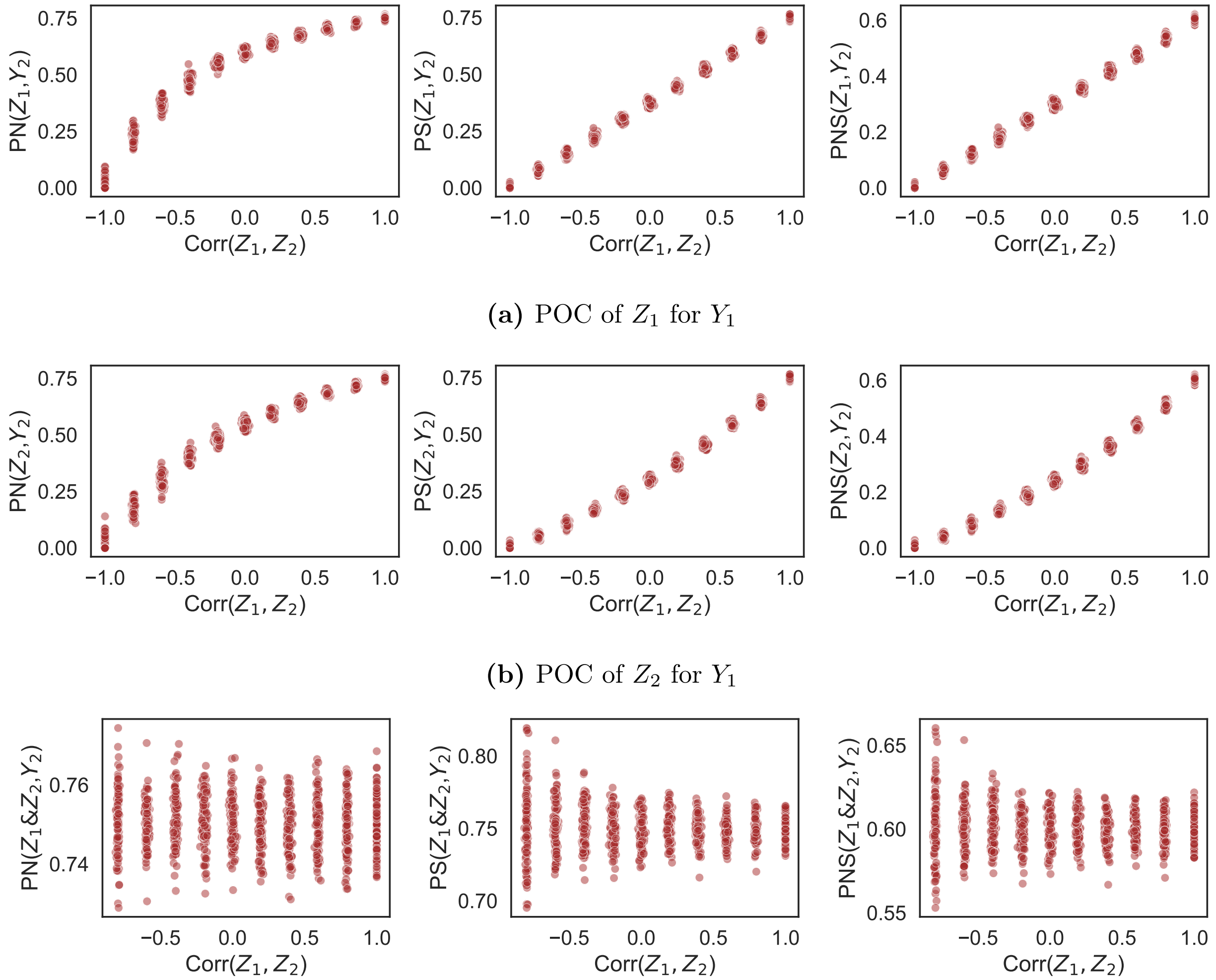
(c) POC of Z_1&Z_2 for Y_1

Figure 12: Lower bounds of probabilities of causation are consistent with intuitive notions of feature necessity and sufficiency. As Z_1 and Z_2 become increasingly highly correlated, (a) the POC of Z_1 for Y_1 stays high, (b) the POC of Z_2 for Y_1 starts to increase when the correlation turns positive, and (c) the POC of Z_1&Z_2 increases.

When $p \in [0, 1]$ is large, then the color of the image is highly correlated with the digit label in such a training set. We further add noise to the ground truth digit label by randomly flipping the labels with a probability of 0.25. The best possible predictive accuracy is thus 0.75 (the yellow dashed line in Figure 6).

The color of the images is a spurious feature in the training set; it has a high correlation with the digit label but does not causally determine the label. In contrast, the features of the digits themselves are non-spurious features; they are highly correlated with the digit label and can causally determine the label.

To create a test set, we color the images such that the color and images are correlated oppositely. We color the '3' images in red with probability $(1 - p)$ and the '8' images in red with probability p.



(a) POC of Z_1 for Y_1

(b) POC of Z_2 for Y_1

(c) POC of $Z_1 \& Z_2$ for Y_1

Figure 13: Lower bounds of probabilities of causation are consistent with intuitive notions of feature necessity and sufficiency. As Z_1 and Z_2 become increasingly highly correlated, (a) the POC of Z_1 for Y_2 increases, (b) the POC of Z_2 for Y_2 increases, and (c) the POC of $Z_1 \& Z_2$ stays high.

As the color-image relationship is very different, a representation learning algorithm will predict poorly in the test set if it only captures color as a feature in the training set.

G.3 Details of the reviews corpora study

Table 4 presents the most informative words of the (positive or negative) ratings, suggested by the CAUSAL-REP representation and the logistic regression coefficients. Across three reviews corpora, logistic regression returns the spurious words "as", "also", "am", "an" as the top words. In contrast, CAUSAL-REP extracts words that are more relevant for the ratings.

Amazon	CAUSAL-REP	Logistic Regression
1	love_this_camera, recommend_this_camera, my_first_digital, great, best_camera, camera_if_you, this_camera_and, camera_have, excellent_camera, camera_bought_this;	am, an, also, as, love_my, the_tracfone,
2	this_camera, camera, camera_is, pictures, picture, the_camera, digital, camera_for, this_camera_is, digital_camera;	it_real, which_is, too, so_much,
3	really_nice, hold_the, excellent_it, this_one_it, easy_it, is_superb, nice_if, returning, too_low, you_need_more;	is_so_much, which_is_pretty,
4	with_this, aa, took, came, yet, pictures_of, camera_in, computer, pictures_in, for_those;	nokia, ear, home, is_must, for_your,
5	camera_was, expect, the_photos, by, camera_are, blurry, sony, have_an, had_some, wife;	faster, must_for, when_use

Tripadvisor	CAUSAL-REP	Logistic Regression
1	front_door, typical, lady, room_is, sized, yogurt, of_italian, in_we, was_at, front_desk_to, was_easy;	am, as, an, also, we_have,
2	lobby, man, directions, door, open, tried, seemed, with_the, by_the, almost;	consideration, and_nice, real,
3	man_at, the_price, is_in, above_and_beyond, we_checked_out, in_central_rome, colliseum, great_for, covered, place_to_sit;	stay_here_again, good
4	the_front, front, front_desk, desk, the_front_desk, at_the_front, desk_staff, front_desk_staff, desk_was, front_of;	
5	the_people_at, the_people, stayed_nights, people_at, was_very_helpful, would_not_recommend, lovely_and, great_location, staff_at, of_my, pricey;	

Yelp	CAUSAL-REP	Logistic Regression
1	but_if, looking, what_you, you_go, but_if_you, that_you, to_do, lot, youll, try, own, do_not;	an, as, am, also, japanese_food,
2	even_if, your_place, this_restaurant, for_you, you_should, thank_you, or_if, thank, lamb, fabulous, is_awesome;	because_its, sure_to_get, ive,
3	want_the, then_you, ahead, hollywood, dont_want, suggest, with, please, check_this_place, all_that;	ive_been, at_night, up_for_it
4	are, you_like, if_you_like, here_if, are_looking, here, are_in, is_the_place, you_need, are_looking_for;	
5	if_you, if, you, want, you_want, you_are, if_you_are, want_to, if_you_want, your, you_want_to, you_dont, dont;	

Table 4: Across the Amazon, Tripadvisor, and Yelp reviews corpa, CAUSAL-REP learns representation that does not rely on the injected words that are spuriously correlated with the sentiment. The table shows the top 12 words identified by the five-dimensional representation from CAUSAL-REP; the five dimensions are not ordered. The representation obtained from logistic regression relies heavily on the spurious words "am", "an", "also", "as."

G.4 Details of the colored and shifted MNIST study

The colored and shifted MNIST dataset is created similarly as in the colored MNIST. We consider four shifts: $(dx, dy) = (0, 1), (1, 0), (0, 1), (1, 1)$—labeled 0,1,2,3—and shift the images such that the shift label is highly correlated with the digit label in the training set.

We evaluate the non-spuriousness of the representations using a downstream prediction task with domain shift. Given a labeled training set where the digit features are no longer correlated with the spurious features, we learn a mapping from the representations to the digit label. A representation can only predict the digit label well if it captured the non-spurious digit features in unsupervised representation learning.

Appendix H. Details of Figure 8

Here we include the full pairwise scatter plot (Figure 14) of the three sets of features in Figure 8.

To generate the factors for Figure 14a, we randomly sample $7,000$ data points from the dsprites dataset.

To generate the factors for Figure 14b, we first sort the data points of the dsprites dataset in descending order by columns ['shape', 'scale', 'orientation', 'positionX', 'positionY']. Next we create the correlated factors dataset by taking the top 5000 data points and then randomly draw 700 data points from the rest.

To generate the factors for Figure 14c, we randomly subsample 7,000 data points and then generate the entangled factors as follows:

$$\begin{aligned}
\text{entangle1} &= 6 \cdot \text{shape}^1 + 8 \cdot (\text{scale}/\text{sd}(\text{scale}))^3 + 1 \cdot (\text{orientation}/\text{sd}(\text{orientation}))^3 + 0.2 \cdot \mathcal{N}(0,1),\\
\text{entangle2} &= 12 \cdot \text{shape}^2 + 1 \cdot (\text{scale}/\text{sd}(\text{scale}))^2 + 8 \cdot (\text{orientation}/\text{sd}(\text{orientation}))^1 + 0.2 \cdot \mathcal{N}(0,1),\\
\text{entangle3} &= 0 \cdot \text{shape}^3 + 4 \cdot (\text{scale}/\text{sd}(\text{scale}))^1 + 4 \cdot (\text{orientation}/\text{sd}(\text{orientation}))^2 + 0.2 \cdot \mathcal{N}(0,1).
\end{aligned}$$

Appendix I. Proof of Theorem 7

Proof We first prove $\text{supp}(Z_1, \ldots, Z_d) = \text{supp}(Z_1) \times \cdots \text{supp}(Z_d)$. Notice that the causal disentanglement of $Z_1, \ldots, Z_d$ (Figure 9) implies that

$$P(Z_1, \ldots, Z_d) = \int P(Z_1, \ldots, Z_d \,|\, C) P(C) \, \mathrm{d}C = \int P(Z_1 \,|\, C) \cdots P(Z_d \,|\, C) P(C) \, \mathrm{d}C. \tag{80}$$

Therefore, we have

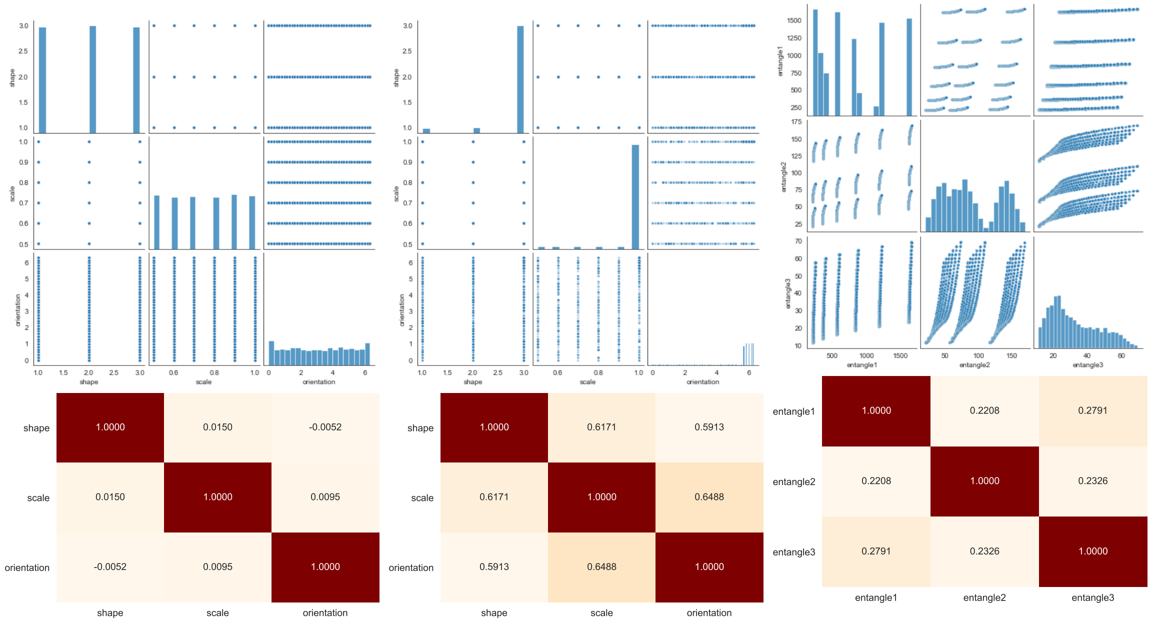
$$P(Z_1, \ldots, Z_d) > 0 \Leftrightarrow P(Z_j \,|\, C \in \mathcal{C}) > 0 \qquad \forall j \in \{1, \ldots, d\} \text{ for some set } P(C \in \mathcal{C}) > 0. \tag{81}$$

Together with the positivity condition, which requires

$$P(Z_j \,|\, C \in \mathcal{C}) > 0 \text{ for any set } P(C \in \mathcal{C}) > 0 \Leftrightarrow P(Z_j) > 0, \tag{82}$$

we have

$$P(Z_1, \ldots, Z_d) > 0 \Leftrightarrow P(Z_j) > 0 \qquad \forall j \in \{1, \ldots, d\}. \tag{83}$$



(a) Disentangled and uncorrelated(IOSS=0.13)

(b) Disentangled but highly correlated(IOSS=0.14)

(c) Entangled but with low correlations (IOSS=0.27)

Figure 14: Disentangled features have independent support even though they may be correlated. Moreover, IOSS can distinguish disentangled and entangled features. This figure illustrates how entangled and disentangled features differ using pairwise scatter plots. Figure 14a considers the ground truth features (shape, scale, orientation) of the dsprites dataset. These features are disentangled. They also have independent support, e.g., conditional on 'scale', the set of values that 'orientation' can take does not change with 'scale.' Visually, these causally disentangled features have scatter plots that occupy rectangular (or hyperrectangular) region. Figure 14b considers the same features but in a subset of the dsprites dataset where the features are correlated. These features, though correlated, are still disentangled; they also have independent support. Figure 14c considers three entangled features, each of which is a nonlinear transformation of the three ground-truth features. These features are not disentangled. Their support are also not independent. Conditional on 'entangle1', the possible values 'entangle2' can take depends on the value of 'entangle1.'

As $\mathrm{supp}(G) = \mathbb{I}\{P(G) > 0\}$, we can rewrite it as

$$\mathrm{supp}(Z_1, \ldots, Z_d) = \mathrm{supp}(Z_1) \times \cdots \mathrm{supp}(Z_d), \tag{84}$$

i.e. $Z_1, \ldots, Z_d$ satisfy the full support condition.

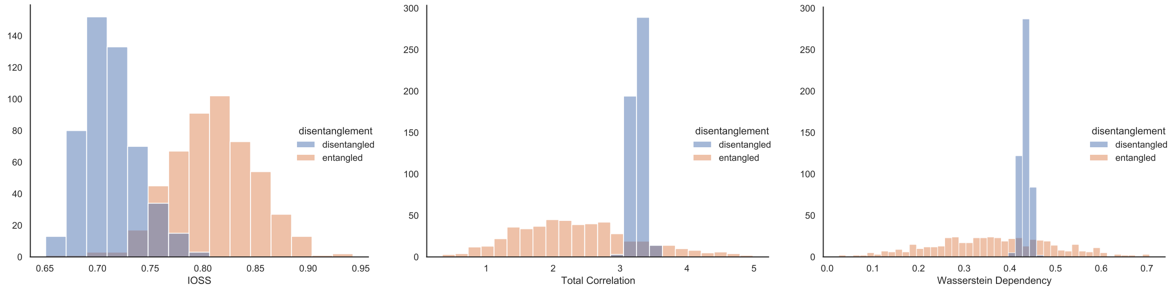
Next we show that $\mathrm{supp}(Z_i \,|\, Z_{\mathcal{S}}) = \mathrm{supp}(Z_i)$.

First notice that

$$\begin{align}\mathbb{I}\{P(Z_i, Z_{\mathcal{S}}) > 0\} &= \mathbb{I}\{P(Z_i \,|\, Z_{\mathcal{S}})P(Z_{\mathcal{S}}) > 0\} \tag{85}\\ &= \mathbb{I}\{P(Z_i \,|\, Z_{\mathcal{S}}) > 0\} \times \mathbb{I}\{P(Z_{\mathcal{S}}) > 0\}. \tag{86}\end{align}$$

Therefore the full support condition $\mathbb{I}\{P(Z_i, Z_{\mathcal{S}}) > 0\} = \mathbb{I}\{P(Z_i) > 0\} \times \mathbb{I}\{P(Z_{\mathcal{S}}) > 0\}$ implies that

$$\mathbb{I}\{P(Z_i) > 0\} = \mathbb{I}\{P(Z_i \,|\, Z_{\mathcal{S}}) > 0\} \text{ for } Z_{\mathcal{S}} \text{ s.t. } \mathbb{I}\{P(Z_{\mathcal{S}}) > 0\}. \tag{87}$$



(a) IOSS **(b)** Total Correlation **(c)** Wasserstein Dependency

Figure 15: IOSS can better distinguish entangled and disentangled representations than existing unsupervised disentanglement metrics on the dsprites dataset.

This implies $\mathrm{supp}(Z_i \,|\, Z_{\mathcal{S}}) = \mathrm{supp}(Z_i)$ for $Z_{\mathcal{S}} \in \mathrm{supp}(Z_{\mathcal{S}})$.

Finally, we consider the support of do interventions.

$$P(Z_{\mathcal{S}} \,|\, \mathrm{do}(Z_i = Z_i)) = \int P(Z_{\mathcal{S}} \,|\, Z_i, C) P(C) \,\mathrm{d}C, \tag{88}$$

$$P(Z_{\mathcal{S}} \,|\, Z_i = Z_i) = \int P(Z_{\mathcal{S}} \,|\, Z_i, C) P(C \,|\, Z_i) \,\mathrm{d}C. \tag{89}$$

Positivity guarantees that the support of $P(C)$ must be the same as $P(C \,|\, Z_i)$. (Because $\mathrm{supp}(C \,|\, Z_i) \subseteq \mathrm{supp}(C)$ and positivity guarantees the other direction.) Therefore, the support of $P(Z_{\mathcal{S}} \,|\, \mathrm{do}(Z_i = Z_i))$ and $P(Z_{\mathcal{S}} \,|\, Z_i = Z_i)$ are the same.

Therefore, if $\mathrm{supp}(Z_{\mathcal{S}} \,|\, Z_i) \neq \mathrm{supp}(Z_{\mathcal{S}})$, then $\mathrm{supp}(P(Z_{\mathcal{S}} \,|\, \mathrm{do}(Z_i = Z_i))) \neq \mathrm{supp}(P(Z_{\mathcal{S}}))$ and hence $P(Z_{\mathcal{S}} \,|\, \mathrm{do}(Z_i = Z_i)) \neq P(Z_{\mathcal{S}})$.

■

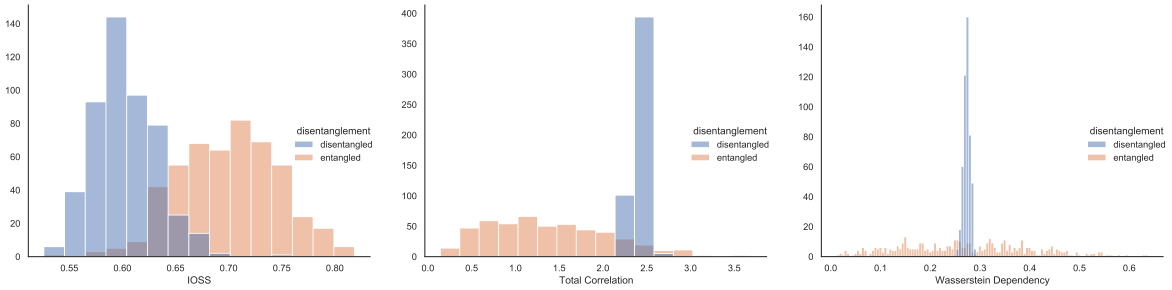
Appendix J. Details of empirical studies of IOSS and additional empirical results

J.1 Details of Section 3.4.1

Figures 10 and 15 to 17 also show that IOSS can better separate disentangled and entangled representations than existing unsupervised disentanglement metrics.

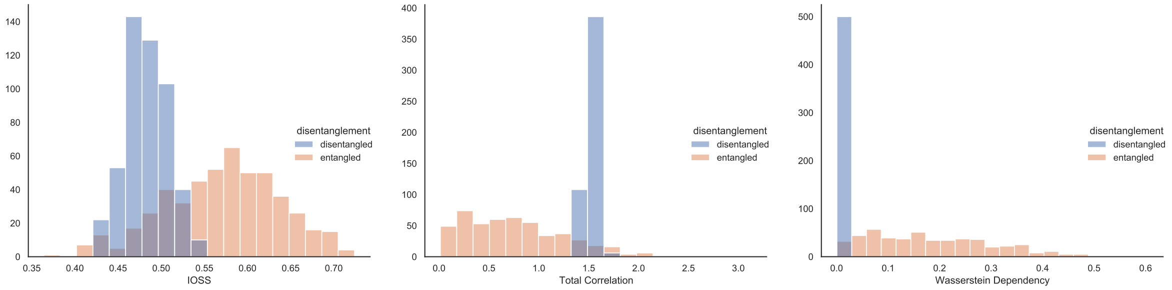
J.2 Details of Section 3.4.2

Figure 18 illustrates that regularizing for IOSS indeed encourages independent support across different dimensions of the learned representations.



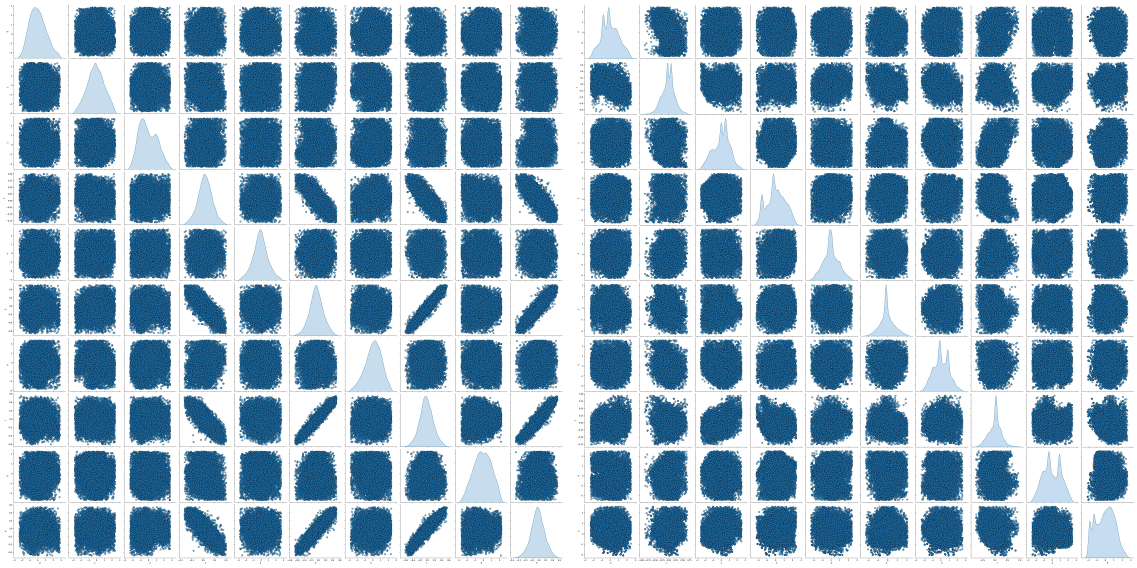
(a) IOSS **(b)** Total Correlation **(c)** Wasserstein Dependency

Figure 16: IOSS can better distinguish entangled and disentangled representations than existing unsupervised disentanglement metrics on the smallnorb dataset.



(a) IOSS **(b)** Total Correlation **(c)** Wasserstein Dependency

Figure 17: IOSS is competitive in distinguishing entangled and disentangled representations compared with existing unsupervised disentanglement metrics on the cars3d dataset.



(a) IOSS regularization strength=0 **(b)** IOSS regularization strength=1000

Figure 18: Increasing regularization strengths of IOSS encourages the learned representation to have independent (i.e. hyperrectangular) support. (a) The pairwise scatter plot of the learned representations without IOSS regularization; some dimensions of the representation do not have independent support. (b) The pairwise scatter plot of the learned representations with IOSS regularization; most dimensions of the representation have independent support.