Causal Representation Learning

Sanae Lotfi, Taro Makino, Lily Zhang
Inference & Representation Final Report

Abstract

Although the field of causality has developed separately from machine learning, it has proven in recent years to be a useful tool for addressing many fundamental questions in machine learning, from robustness to distribution shift, explainability to fairness. In this review, we discuss existing work in causal representation learning and structure learning. First, we define a causal representation as one which satisfies properties of causal variables and survey existing works to build causal representations. We describe methods for building causal representations which utilize data from multiple environments, as well as methods which only assume access to samples from a single environment. We then describe approaches for learning the structure of a causal graph. Finally, we provide closing thoughts on potential future directions based on existing works.

1 Introduction and motivation

Causal representation learning has attracted increasing interest from the machine learning community in recent years. By combining ideas in representation learning with causal inference, causal representation learning offers the potential for representations which are robust and transferrable, interpretable as cause-effect mechanisms, and sample-efficient [Schölkopf et al., 2021].

Machine learning models have achieved impressive success on a wide range of tasks. However, most statistical machine learning methods rely on the assumption that training data and future test data are independent and identically distributed (IID). When the IID assumption is violated (e.g. due to natural distribution shift in the world), such methods can show a dramatic degradation in performance. In contrast, predictions which take into account the underlying causal mechanism behind the observed data have the potential to generalize to new settings. Below, we discuss particular instantiations of this idea via a variety of methods which combine machine learning and causality towards building models which generalize to settings beyond the training distribution.

2 Background

2.1 The goal of causal representation learning

Statistical models can be learned from unstructured data such as images and text but cannot predict reliably under real-world data distribution shifts. In contrast, causal models can predict reliably under data distribution shifts but are traditionally used in conjunction with structured (rather than unstructured) data provided in the form of causal variables. Causal representation learning aims to incorporate ideas from both representation learning and causal inference in order to learn models from unstructured data which have desirable properties of causal models, such as robustness to data distribution shifts.

2.2 From statistical to causal modeling

The notion of an intervention is a defining characteristic of causal modeling that differentiates it from statistical modeling. Consider a pair of variables X and Y, where X causes Y, denoted $X \to Y$.

If we intervene on X, the resulting intervention distribution, $P(Y \mid do(X = x))$, is the population distribution of Y if we fix everyone in the population's X value to x. In contrast, the conditional probability $P(Y \mid X = x)$ is the distribution of Y in the subset of the population where X was x. In general, $P(Y \mid do(X = x))$ does not equal $P(Y \mid X = x)$, and using $P(Y \mid X = x)$ naively to represent $P(Y \mid do(X = x))$ can lead to problems [Schölkopf, 2019]. Incorporating causality into our ML methods could potentially prevent us from conflating correlation with causation.

2.3 Branches of causal modeling

Causal modeling contains a pair of inverse problems: causal reasoning and causal discovery. In causal reasoning, a causal model is given and used to compute various interventional and counterfactual probabilities. In causal discovery, the structure of the causal graph is learned from a dataset of causal variables. We can divide existing work in causal representation learning via these two branches: causal (or causally-inspired) representation learning focuses primarily on causal reasoning in that the goal is to compute quantities that are counterfactual in nature; and structure learning is akin to causal discovery, as the goal is to learn a causal graph.

2.4 Structural causal models

In structural causal models (SCMs), we assume that the observables $X_1, \ldots, X_n$ are random variables that are associated with the vertices of a directed acyclic graph G. Each observable X_i is assigned as follows: $X_i := f_i(\mathbf{PA}_i, U_i)$, $i = 1, \ldots, n$, where f_i is a deterministic function and $U_1, \ldots, U_n$ are jointly independent stochastic variables. The graph G is a *causal graph* since the parents of X_i affect its assignment directly. Using this framework, we can formalize **interventions** as modifications in a subset of assignments, e.g. by changing U_i or f_i. More importantly, this framework gives rise to a causal or distangled factorization of the joint distribution as follows:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\mathbf{PA}_i). \tag{1}$$

This factorization is the only one that decomposes the joint distribution into factors that correspond to the structural assignments defined by our SCM. From the other side, many entangled factorizations are possible, e.g.,

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|X_{i+1}, \ldots, X_n). \tag{2}$$

One of the main differences between the two factorizations is what we call the Independent Causal Mechanisms (ICM) principle, which we will discuss in the next section.

2.5 Independent Causal Mechanisms

A crucial assumption in causal inference called the *Independent Causal Mechanisms Principle* can be stated as follows: *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other* [Schölkopf et al., 2012, Peters et al., 2017, Schölkopf, 2019].

Two consequences follow from this principle with respect to the causal factorization in Eq. (1):

- No flow of influence: intervening upon one mechanism $p(X_i|\mathbf{PA}_i)$ does not change the other mechanisms $p(X_j|\mathbf{PA}_j)$, $i \neq j$.
- No flow of information: knowing a mechanism $p(X_i|\mathbf{PA}_i)$ does not give us information about another mechanism $p(X_j|\mathbf{PA}_j)$, $i \neq j$.

Notice that the first consequence of the ICM principle does not necessarily apply to the entangled factorization in Eq. (2). Consider for example the relationship between altitude A and temperature T [Peters et al., 2017]. The joint distribution can be written as $p(T, A) = p(T|A)p(A)$ or $p(T, A) = p(A|T)p(T)$. Since $p(T|A)$ describes the physical mechanisms that generate temperature from altitude, it will not be influenced by a change in the distribution $p(A)$ over altitudes. Therefore, the first factorization is a causal factorization that obeys the two consequences of the ICM principle. We

cannot say the same thing about the second entangled factorization, where $p(A|T)$ can be effected by an intervention upon $p(T)$. As a concrete example, arbitrarily heating up a subset of cities will change the conditional distribution $p(A|T)$. It is this robustness under intervention and the ability of causal mechanisms to generalize across different datasets that motivates our next section on causation as invariance.

2.6 Causation as invariance

Causation can also be seen through the lens of invariance, since the Independent Causal Mechanisms principle specifies **invariance under intervention**. As described previously, the causal factorization in Eq. (1) guarantees that intervening upon one mechanism $p(X_i|\mathbf{PA}_i)$ leaves mechanisms $p(X_j|\mathbf{PA}_j)$ with $i \neq j$ **invariant**. This view of causality inspired many philosophical works [Skyrms, 1980, Cartwright, 2003, Woodward, 2005, Cheng and Lu, 2017]. In machine learning, it particularly inspired two lines of research: 1) invariant causal prediction and invariant risk minimization (section 3), and 2) learning causal graphs from interventions (section 4).

2.7 Structure of the paper

The paper is structured as follows: in section 3, we give an overview of recent works that aim to achieve causal representation learning and eliminate dependence on spurious correlations, either by harnessing access to multiple environments or knowledge of the graph. In section 4, we discuss learning causal structures where we try to uncover the true parent-child relationship for a set of variables. Finally, we offer closing thoughts in the conclusion. See Fig. 2 for a visual overview.

3 Learning causal representations

In this section, when we refer to a causal representation, we mean a representation that satisfies properties we would expect causal variables to satisfy. For instance, we think of causal representations as those whose relationship with the label is consistent, just as we expect causal relationships to stay the same across time and geography even when other relationships do not. We refer to these other relationships which do not persist as *spurious correlations*.

Consider the image classification task of classifying cows vs. penguins. While the relationship between the animal shape and label will persist, the relationship between the background of the image and the animal classification is not guaranteed to stay the same. For instance, in images from natural habitats, cows appear more frequently in front of grassy backgrounds while penguins appear more frequently in front of snow. In images taken in human-constructed environments (e.g. farms, zoos) on the other hand, the relationship between the background and the animal label may be different. We will use this running example throughout the section for illustration.

There have been many causally-inspired approaches to representation learning. We discuss each below, dividing them between methods which do and do not assume access to multiple environments.

3.1 Learning Representations Using Multiple Environments

The link between causality and invariance was established by many works in the machine learning field and has been used to avoid learning spurious correlations. As defined previously, a correlation is called spurious when it is not maintained across different settings. Therefore, the goal is to learn correlations that remain invariant across different training environments. In this section, we consider the setting where we have multiple training environments $e \in \mathcal{E}_{\text{tr}}$ that correspond to datasets $D_e = \{(x_i^e, y_i^e)\}_{i=1}^{n_e} = (X^e, Y^e)$. We additionally assume that these environments were generated by unknown and not precisely controlled interventions.

Invariant causal prediction (ICP) Peters et al. [2016] exploit the link between causality and invariance to find the **direct** causes of a target variable in the framework of structural causal models described in section 2.4. Formally, for each subset of features X^e, we consider the linear model $Y^e = \mu + X^e\gamma^* + \epsilon^e$, where μ is a constant intercept term, γ^* denotes the coefficients corresponding to X^e and $\epsilon^e \sim F_\epsilon$ is the noise variable distributed identically for all environments. If this linear model does not change across multiple training environments, then the corresponding subset of

features gives raise to an invariant predictor, which makes them *plausible causes* of the target variable Y. We iterate over subsets of features combinatorially and find multiple sets of plausible causes. Finally, the true direct causes of the target variable Y correspond to the intersection of the sets of plausible causes. Notice that for this method to work, there should be no unobserved counfounders between the target variable and the features. Otherwise, the objective should be adjusted to take these confounders into account. The authors prove that this method recovers the true set of causal predictors when the data is Gaussian and satisfies a linear SCM under certain types of interventions, namely do-interventions, noise interventions and simultaneous noise interventions (interventions that change the noise for different variables at once).

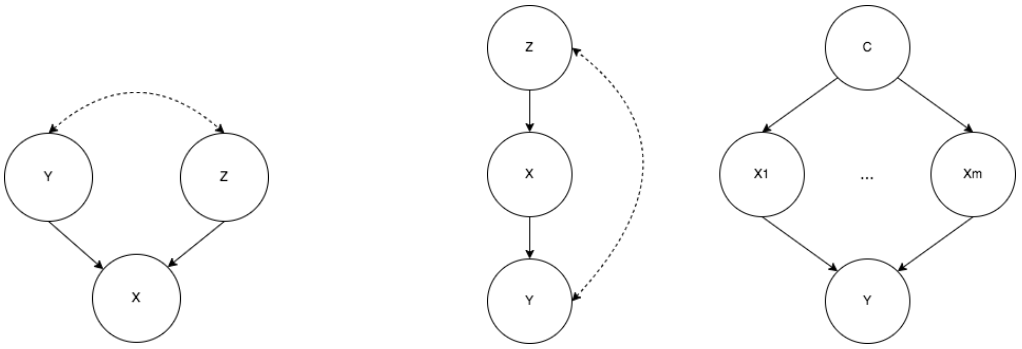
Invariant risk minimization (IRM) One of the main drawbacks of ICP proposed by Peters et al. [2016] is the difficulty of encoding the causal relationships between observed variables, or the assumption that a meaningful causal graph relating them exists. For example what is the causal relationship between individual pixels and a target prediction? Moreover, ICP only applies to linear models and its complexity scales exponentially with the number of variables. Arjovsky et al. [2019] depart from these observations to propose *invariant risk minimization*, a learning method that exploits multiple environments to achieve invariance *without* attempting to directly learn the causal parents of the target variables. Formally, a data representation $\Phi : \mathcal{X} \to \mathcal{H}$ is said to elicit an *invariant predictor* if there is a classified $w : \mathcal{H} \to \mathcal{Y}$ that is simultaneously optimal on all environments. This is a bi-level constrained optimization problem, where we want to: i, find an optimal data representation Φ, and ii, find the optimal induced classifier w across multiple training environments. The authors propose a practical IRM formulation where they introduce a fixed dummy classifier and add a penalty term that penalizes high gradients in each environment:

$$\min_{\Phi:\mathcal{X}\to\mathcal{H}} \sum_{e\mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=0.1} R^e(w \cdot \Phi)\|^2,$$

where R^e is the risk under environment e. Intuitively speaking, the regularization term penalizes solutions that can be improved with more training using the high gradient signal. The authors prove that linear IRM, where both data representations and classifiers are linear, learns the target invariance even when the three conditions needed for ICP's theoretical results do not hold. However, no theoretical guarantees for the non-linear case were provided due to the challenging bi-level nature of the optimization problem.

Does invariant risk minimization really work? Rosenfeld et al. [2021] provide theoretical evidence that if i, a function of invariant and spurious independent Gaussian factors is used to generate the features, ii, the factors have linear correlations with the label, and iii, the dimensionality of invariant features is greater than the number of domains, then the IRM framework would lead to learning the spurious representations. Guo et al. [2021] show that deep neural networks can learn non-overlapping spurious correlations across environments without hurting the IRM performance on the training environment and thus perform poorly on the new test environment. They show that this is particularly true when there is a *strong triangle spuriousness*, i.e., when there is a stronger correlation among spurious features, the domain variable, and the class variable than invariant correlation among invariant features and the class label.

Extending IRM to the nonlinear case Lu et al. [2021] extend the theoretical guarantees of IRM to the case of nonlinear representations and nonlinear classifiers using a novel approach: invariant Causal Representation Learning (iCaRL). More precisely, they consider a variational autoencoder where X^e stands for the observed variables, Z for the latent variables, and Y^e for the target variables given environment e. Following Khemakhem et al. [2020], it is possible to obtain an identifiable model, i.e., where two different choices of model parameters lead to the same model, by placing a conditionally factorized prior distribution over the latent variables $p_\theta(Z|U)$, where U is an additional observed variable, e.g., target or environment label. Lu et al. [2021] use this result and place a more flexible prior over the latent variables, namely general exponential family distributions. This guarantees in theory that all the direct causes of the target can be fully discovered, which is done in practice by observing that the dependency between two causes of Y, Z_i and Z_j, increases after conditioning on Y, and testing this conditional dependency for all pairs. Finally, discovering the direct causes of Y simplifies the IRM bi-level optimization problem to two simpler independent optimization problems that can be performed separately to find the data representation and the optimal classifier across different environments.



(a) Anti-causal (b) Causal

Figure 1: Anti-causal vs. causal graphs. While CSR and NuRD assume an anti-causal relationship (a), CR-PNS assumes a causal relationship (b, right). C-Inv considers both (a; b, left).

3.2 Learning Representations Using a Single Environment

Several recent papers describe methods for handling spurious correlations assuming only one environment, the training data distribution. While methods such as IRM assume access to samples from multiple environments, methods which assume access to only a single environment make alternative assumptions to identify a suitable representation for generalization—for instance, knowledge of the causal graph and the covariate that is spuriously correlated. We walk through each of the following four methods below, highlighting the similarities and differences in their assumptions and proposed methods:

- Causally-motivated Shortcut Removal (CSR) [Makar et al., 2021]
- Nuisance-randomized Distillation (NuRD) [Puli et al., 2021]
- Counterfactual Invariance (C-Inv) [Veitch et al., 2021]
- Causal Representation Learning via Probability of Sufficiency and Necessity (CR-PNS) [Wang and Jordan, 2021]

First, we detail the data generating processes assumed in these works. Makar et al. [2021], Puli et al. [2021], Veitch et al. [2021] assume that there exists a label Y, input X, and a variable Z which is spuriously correlated with Y in the training distribution. Puli et al. [2021] call Z the nuisance variable, terminology we will use in the remainder of the section. Using the running cows vs. penguins example, X denotes the image, Y the animal label, and Z the background. Wang and Jordan [2021] do not assume the existence of an additional variable Z but rather that deterministic functions of X can be spuriously correlated with the label.

It is possible to consider both a causal and anticausal relationship between input X and label Y. Puli et al. [2021], Makar et al. [2021] assume an anti-causal relationship—that is, Y causes X—while Wang and Jordan [2021] assume a causal relationship. Veitch et al. [2021] consider both scenarios in their approach. Puli et al. [2021], Makar et al. [2021] describe their setup in a non-causal interpretation for the anticausal assumption using the notion of a family of distributions. In this family, the marginal label distribution $p(Y)$ and the conditional input distribution $p(X|Y, Z)$ remain the same across all members, but the conditional distribution of Z given the label changes, i.e. $p_D(Z|Y) \neq p_{D'}(Z|Y)$. See Fig.1 for a comparison of the structural causal graph assumptions of these methods. We highlight additional assumptions in the exposition of the specific methods.

Next, the works propose different methods for building predictive models based on representation learning. We first focus on the anti-causal scenario considered in Makar et al. [2021], Puli et al. [2021], Veitch et al. [2021].

Assuming an anti-causal relationship. In CSR, Makar et al. [2021] assume that the label Y affects X only through a sufficient statistic $X^* = f(X)$. In contrast, the theory in NuRD and C-Inv do not make such an assumption and specifically consider that parts of X can be a function of both

	CSR	NuRD	C-Inv	CR-PNS
anti-causal	✗	✗	✗	
causal			✗	✗
Z	✗	✗	✗	
discrete Z	✗		✗	
suff. stat X^*	✗			

Table 1: Distinguishing single-environment methods by their assumptions.

Y and Z. Given this assumption in CSR, Makar et al. [2021] show that breaking the dependence between label Y and nuisance Z in the training data via reweighting is sufficient to ensure that the resulting Bayes optimal predictor is only a function of X^*—and thus also the optimal predictor across all the distributions governed by the family of distributions described above. They additionally include a marginal independence constraint under the reweighted distribution (which we call $p_\perp$) as a regularizer: $f(X) \perp\!\!\!\perp_{p_\perp} Z$. Even though it is not necessary to obtain the result above, they show that this additional constraint helps with sample efficiency in learning. They actualize the regularizer constraint via Maximum Mean Discrepancy (MMD).

Puli et al. [2021] also utilize reweighting as part of the solution in NuRD. However, as X can contain mixed functions (i.e. using both Y and Z), Puli et al. [2021] introduce an additional constraint which ensures that even after conditioning on the representation $f(X)$, the label and nuisance remain independent: $Y \perp\!\!\!\perp_{p_\perp} Z|f(X)$. They call such a representation $f(X)$ an uncorrelating representation. Puli et al. [2021] show that a representation within the uncorrelating set that is simultaneously optimal (across the family) can be found, assuming existence, by maximizing mutual information with the label. To avoid local maxima which complicate optimization, NuRD also incorporates the marginal constraint $f(X) \perp\!\!\!\perp_{p_\perp} Z$ which combined with the uncorrelating constraint yields the following joint independence constraint: $(Y, f(X)) \perp\!\!\!\perp_{p_\perp} Z$.

In C-Inv, Veitch et al. [2021] design predictors which do not change under arbitrary changes to Z. They do so by incorporating the constraint $f(X) \perp\!\!\!\perp Z|Y$ via MMD. Puli et al. [2021] show that this constraint also implies $(Y, f(X)) \perp\!\!\!\perp_{p_\perp} Z$, which is more restrictive than the uncorrelating assumption. Veitch et al. [2021] note that C-Inv will not produce predictors which are minimax optimal across causally-relevant distributions in general, except in the case where the parts of X which mix Y and Z do not contain useful information.

CSR and C-Inv both assume a discrete nuisance, while NuRD works with general nuisances, including high-dimensional ones (e.g. patches of image pixels).

Assuming a causal relationship Under the assumed causal graph, Veitch et al. [2021] note that a representation which is independent of Z will be counterfactually invariant: $f(X) \perp\!\!\!\perp Z$. The same implementation ideas and theoretical results noted in the anti-causal setting also apply here.

Wang and Jordan [2021] define a causally-inspired objective to optimize directly when learning a representation. Like in Veitch et al. [2021], they define a counterfactual property or properties they desire in a causal feature. In particular, they describe probability of sufficiency and necessity: turning on a feature counterfactually (i.e. setting it to a value when it is not that value) should likely turn on the label given that it is not on, while turning off a feature should turn off the label given it is on. They provide a single metric which jointly optimizes probability of sufficiency and necessity and describe the assumptions needed to enable identifiability of the relevant interventional quantities from observational data. They then optimize a lower bound of this objective to build causal representation. They do not directly consider theoretical optimality guarantees of predictive models built on such representations.

4 Learning causal structures

The goal of causal representation learning is to learn from unstructured observational data a latent representation that satisfies properties of a causal variables. In this section, we discuss causal structure learning, the setting where the causal variables are given to us, and our goal is to learn their parent-child relationships. There are two main difficulties with this. The first is that DAG structure learning is a difficult constrained combinatorial optimization problem, since the number

of possible graphs scales super-exponentially with the number of variables, and the graphs must be acyclic. The second is that the ground-truth graph is unidentifiable from purely observational data without further assumptions, since many factorizations of the joint distribution are consistent with the same observational distribution. Unidentifiability is due to the fact that there are multiple causal explanations which can explain observational data equally well, which is problematic.

We examine a recent line of work which turns the combinatorial optimization problem into a continuous one [Zheng et al., 2018], eliminating the need for specialized algorithms. Others then followed up to incorporate the idea of interventions in order to alleviate the issue of unidentifiability [Brouillard et al., 2020, Ke et al., 2019, Scherrer et al., 2021]. With interventional data, the ground-truth DAG is identifiable up to a smaller equivalence class given purely observational data.

Zheng et al. [2018] DAGs with NO TEARS: Continuous Optimization for Structure Learning

The main contribution of this paper is to introduce a smooth and differentiable function $h : \mathbb{R}^{d\times d} \to \mathbb{R}$ such that a graph with adjacency matrix $W \in \mathbb{R}^{d\times d}$ is acyclic if and only if $h(W) = 0$. The function is given by

$$h(W) = \operatorname{tr}(\exp(W \odot W)), \tag{3}$$

where $\odot$ is the Hadamard product. Note that $\operatorname{tr} B^k$ is the number of length-k closed walks in a directed graph with adjacency matrix $B \in \{0, 1\}^d$. Therefore, an acyclic graph will have $\operatorname{tr} B^k$ for all $k \geq 1$. Guided by this insight, the authors arrive at Eq. (3) by also considering numerical stability.

Ke et al. [2019] Learning Neural Causal Models from Unknown Interventions

While Zheng et al. [2018] opened the door for continuous optimization in structure learning, it only utilized purely observational data. This is problematic, as the graph is only identifiable up to a Markov equivalence class. The authors of Ke et al. [2019] address this by extending Zheng et al. [2018] to allow for interventional data. There are two sets of parameters: θ encodes the conditional distributions $P(X_i \mid \mathbf{PA}_i)$, while γ parameterizes the adjacency matrix, where $\sigma(\gamma_{i,j})$ represents the probability that node X_j is a parent of X_i. Their method follows a three-stage procedure. In the first stage, θ is optimized to maximize the log-likelihood of the observational data under the model. In the second stage, data is sampled from the SCM, and one of the variables X_i is intervened on by changing $P(X_i \mid \mathbf{PA}_i)$. Then, the log-likelihood of the sampled data is maximized after omitting the $\log P(X_i \mid \mathbf{PA}_i)$ corresponding to the intervened variable. This evaluation instantiates the idea of the principle of independent mechanisms, since we wish to learn $P(X_j \mid \mathbf{PA}_j)$'s which are invariant to changes in $P(X_i \mid \mathbf{PA}_i)$, where $i \neq j$. In the third and final stage, γ is updated w.r.t. the omitted-data log-likelihood using a gradient estimator.

Brouillard et al. [2020] Differentiable Causal Discovery from Interventional Data

Brouillard et al. [2020] is another extension of Zheng et al. [2018] to include interventional data. In contrast to Ke et al. [2019], Brouillard et al. [2020] assume continuous data and is theoretically-grounded. Their theoretical results state that their optimization objective yields the ground-truth DAG up to an $\mathcal{I}$-Markov equivalence [Yang et al., 2018], which is more restrictive than Markov equivalence.

Their method maximizes the log-likelihood of interventional data in a way that encourages the $P(X_i \mid \mathbf{PA}_i)$'s to be invariant to one another. The authors instantiate the principle of independent mechanisms albeit in a different way than Ke et al. [2019]. The method encodes the interventional family with a binary matrix $R \in \{0, 1\}^{K\times d}$. The k'th row of this matrix is a d-dimensional mask which specifies the variables to be targeted in the k'th intervention. The joint density for the k'th intervention is given by

$$f^{(k)}(x; M, R, \phi) \coloneqq \prod_{j=1}^{d} f(x_j; \text{NN}(M_j \odot x; \phi_j^{(1)}))^{1-R_{k,j}} f(x_j; \text{NN}(M_j \odot x; \phi_j^{(k)}))^{R_{k,j}}. \tag{4}$$

This expression evaluates to the first term in the product if the j'th node is not being intervened on in the k'th intervention, and the second term otherwise. M_j is a binary mask which selects the parents

of the j'th node, and NN$(PA_j; \phi_j^{(k)})$ returns the parameters of $P(X_j \mid \mathbf{PA}_j)$. The optimization objective is to maximize this joint density over all $k = 1, \ldots, K$ interventions, or

$$\mathcal{S}(\mathcal{G}) \coloneqq \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}}[\log f^{(k)}(X; M, R, \phi)] - \lambda |\mathcal{G}|. \tag{5}$$

This objective encourages $P(X_i \mid \mathbf{PA}_i)$ to be invariant across interventions that do not target X_i, which is the principle of independent mechanisms.

Connection to IRM [Arjovsky et al., 2019] In the original formulation, we are given data from multiple intervention distributions $\{p^{(k)}\}_{k=1}^{K}$, and we also know the ground-truth intervention targets $R_{k,j}$ for all k, j. The authors also propose a way to learn the intervention targets in the event that the ground-truth assignments are unknown. They accomplish this in a relatively simple way, by adding a regularizer to the loss to favor sparse interventions. The setting of Brouillard et al. [2020], where the data comes grouped from multiple intervention distributions, but the identities of the intervened variables is unknown, mirrors that of IRM [Arjovsky et al., 2019]. The different intervention distributions are akin to environments, whose labels are known. In IRM, we know that the data comes from different environments, but we do not know what causal variables were intervened on in order to create the different environments.

Scherrer et al. [2021] Learning Neural Causal Models from Active Interventions

In Ke et al. [2019], the optimization objective for learning the causal graph is obtained by randomly sampling a set of intervention targets on each iteration. Similarly, in Brouillard et al. [2020], the optimization objective involves summing over all possible intervention sets $k = 1, \ldots, K$. The authors of Scherrer et al. [2021] realized that this is inefficient, and proposed a more sample efficient way of selecting interventions. Instead of selecting them randomly or evaluating all possible combinations, they propose selecting a subset of intervention targets which maximizes the variation between data sampled from a set of graph configurations following a given intervention. In order to compute the discrepancy score for a given intervention, they first sample a set of causal graphs that are consistent with the intervention. Then, the score for the k'th intervention target is defined as

$$D_k \coloneqq \frac{\sum_i (\mu_i^k - \mu^k)^2}{\sum_i \sum_j (S_j^{k,i} - \mu_i^k)^2}, \tag{6}$$

where i is the index of the causal graph, j is the index of the data sample. μ_i^k is the mean of the data sampled from graph i, μ^k is the mean of the μ_i^k's across all graphs, and $S_j^{k,i}$ is the j'th sample from the i'th graph. In other words, this score is the ratio between the variance between groups, and the variance within each group. This score is used to find the intervention targets that provide the strongest signal for finding the correct graph structure.

5 Conclusion

This review surveys existing work in causal representation and structure learning. We note that work in causal representation learning mirrors causal reasoning in traditional causal inference, while structure learning is related to the problem of causal discovery. The works discussed in this review generally fall under the umbrella of incorporating causality into machine learning [Schölkopf, 2019].

There are many future directions in this space. First, just as causal inference relies on a careful specification of assumptions, future work on causal representation learning could consider alternative assumptions. Such assumptions include other causal graphs, access to various sources of knowledge/information, or further assumptions on the relationship between environments. Additionally, various machine learning-specific advances could also help in this space; concretely, the theory behind most of these works assumes the exact distributions of interest, which means that better modeling could help bring the empirical results closer to their analytical counterparts. Finally, for causal structure learning, additional constraints or information to enable sample-efficient learning could yield benefit.

References

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *NeurIPS*, 2020.

Nancy Cartwright. Two theorems on invariance and causality. *Philosophy of Science*, 70(1):203–224, 2003.

Patricia W Cheng and Hongjing Lu. Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing. *The Oxford handbook of causal reasoning*, page 65, 2017.

Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kıcıman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *ArXiv*, abs/2101.07732, 2021.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, H. Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *ArXiv*, abs/1910.01075, 2019.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv*, 2102.12353, 2021.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally-motivated shortcut removal using auxiliary labels. *arXiv*, 2105.06422, 2021.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Aahlad Manas Puli, Lily H. Zhang, Eric K. Oermann, and Rajesh Ranganath. Predictive modeling in the presence of nuisance-induced spurious correlations. *arXiv*, 2107.00520, 2021.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *ArXiv*, abs/2010.05761, 2021.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions. *ArXiv*, 2021.

Bernhard Schölkopf. Causality for machine learning. *arXiv*, 1911.10500, 2019.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv*, 2102.11107, 2021.

Brian Skyrms. Causal necessity: a pragmatic investigation of the necessity of laws. 1980.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv*, 2106.00545, 2021.

Yixin Wang and Michael I. Jordan. Desiderata for representation learning: A causal perspective. *arXiv*, 2109.03795, 2021.

James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

Karren D. Yang, Abigail Katoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *ICML*, 2018.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with NO TEARS: continuous optimization for structure learning. In *NeurIPS*, 2018.

A Appendix

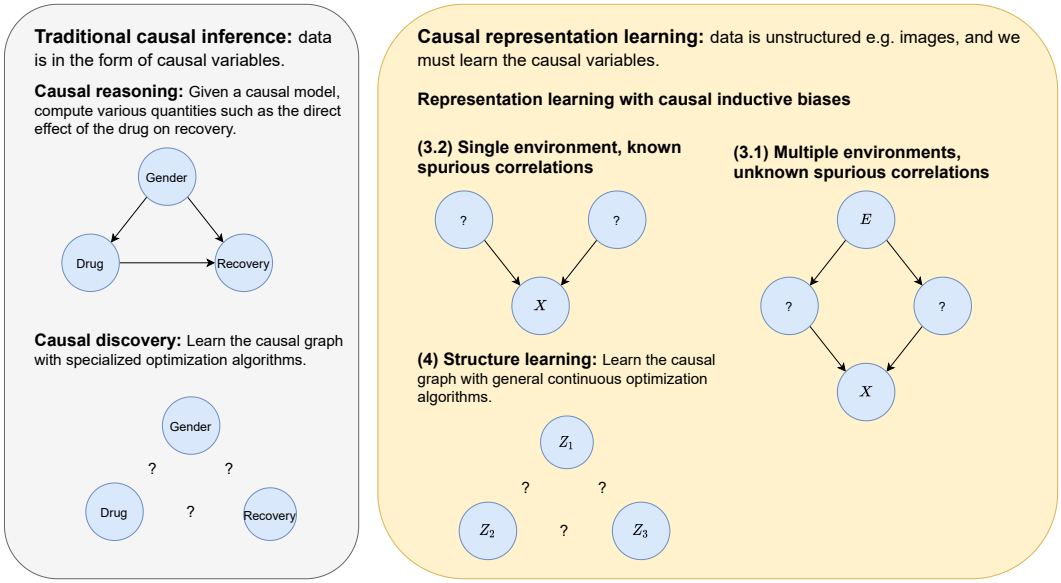


Figure 2: This figure summarizes the difference between traditional causal inference (left) and causal representation learning (right). In traditional causal inference, data is given to us in the form of causal variables, such as gender, drug, and recovery. Two main activities in causal inference are causal reasoning and causal discovery. In causal reasoning, a causal model is provided, and we use it to compute various interventional and counterfactual probabilities. In causal discovery, the causal relationships between the variables is not known, and the goal is the learn the structure of the causal graph. Meanwhile, in causal representation learning, the data is in an unstructured form, such as images and raw text. In representation learning with causal inductive biases, the goal is to learn a latent representation of the data that satisfy properties of causal variables. One example of this is robustness to spurious correlations. In Section 3.2, we discuss methods where the data is assumed to come from a single environment, and we may or may not know the identify of the spurious correlations. In Section 3.1, the data is assumed to come from multiple environments, whose labels are known. The objective is then to learn a predictor that is invariant to changes in the environment. Finally, Section 4, the objective is to learn causal structures with continuous optimization algorithms that can potentially scale to a large number of causal variables.