Semi-supervised learning, causality, and the conditional cluster assumption

Julius von Kügelgen[1,2]   Alexander Mey[3]   Marco Loog[3,4]   Bernhard Schölkopf [1]

[1] Max Planck Institute for Intelligent Systems Tübingen, Germany
[2] Department of Engineering, University of Cambridge, United Kingdom
[3] Delft University of Technology, The Netherlands
[4] University of Copenhagen, Denmark
{jvk, bs}@tuebingen.mpg.de,{a.mey, m.loog}@tudelft.nl

Abstract

While the success of semi-supervised learning (SSL) is still not fully understood, Schölkopf et al. (2012) have established a link to the principle of independent causal mechanisms. They conclude that SSL should be impossible when predicting a target variable from its causes, but possible when predicting it from its effects. Since both these cases are restrictive, we extend their work by considering classification using cause and effect features at the same time, such as predicting a disease from both risk factors and symptoms. While standard SSL exploits information contained in the marginal distribution of all inputs (to improve the estimate of the conditional distribution of the target given inputs), we argue that in our more general setting we should use information in the conditional distribution of effect features given causal features. We explore how this insight generalises the previous understanding, and how it relates to and can be exploited algorithmically for SSL.

1 INTRODUCTION

Due to the scarcity and often high acquisition cost of labelled data, machine learning methods that make effective use of large quantities of unlabelled data are crucial. One such method is semi-supervised learning (SSL) (Chapelle et al., 2010; Zhu, 2005) where, in addition to labelled data, possibly large numbers of unlabelled observations are available to the learner at training time. While positive results have been obtained on a range of problems, a shortcoming is that SSL can actually degrade performance if certain assumptions are not met (Chapelle et al., 2010). For example, Ben-David et al. (2008) show that the cluster assumption, commonly used in SSL settings, can lead to degraded performance even in simple cases, e.g., for binary classification with data generated from two uni-

modal Gaussians. Such examples make it clear that many aspects of SSL are, as of yet, not well understood.

Building on the *principle of independent causal mechanisms* (ICM) (Daniušis et al., 2010; Peters et al., 2017), Schölkopf et al. (2012) have pointed out a link between the possibility of SSL and the causal structure underlying a given learning problem. Specifically, they argue that SSL should be impossible when predicting a target variable from its causes (*causal learning*), but possible when predicting it from its effects (*anticausal learning*)—see §2 for details. Empirical evidence from a meta-analysis of various SSL scenarios supports this claim.

In this work, we extend the investigation of connections between SSL and causality to a more general setting. Rather than treating causal and anticausal learning in isolation, we consider *predicting a target variable from both causes and effects at the same time*. As an example, consider the setting of predicting disease from medical data where both types of features are commonly found: a patient's age, sex, medical family history, genetic information, diet, and other risk factors such as smoking all constitute (possible) causal features; examples of effect features, on the other hand, include the clinical symptoms exhibited by the patient, as well as results of medical tests such as imaging results, serum tests, or tissue samples.

As our main result, we show in §3 that for this setting of SSL with both cause and effect features, *the relevant information that additional unlabelled data may provide for prediction is contained in the conditional distribution of effect features given causal features*. This generalisation of Schölkopf et al. (2012) contains their results for causal and anticausal learning as special cases. We then use this new insight to reformulate classical SSL assumptions in §4, and propose algorithms based on these assumptions in §5. Results from evaluating our methods against well-established SSL algorithms on synthetic and medical datasets in §6 empirically support our analysis. We critically discuss our assumptions and results in §7 and conclude with an outlook on future work in §8.

2 BACKGROUND & RELATED WORK

Throughout, we use X to denote a random variable taking values in $\mathcal{X} \subseteq \mathbb{R}^d$. P denotes a probability measure and $P(X)$ the probability distribution of X with density p. We write $x \in \mathbb{R}$ for a scalar, $\mathbf{x} \in \mathbb{R}^d$ for a vector, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ for a matrix or collection of samples.

2.1 SEMI-SUPERVISED LEARNING (SSL)

SSL describes a learning setting where, in addition to a labelled sample $(\mathbf{X}^l, \mathbf{y}^l) = \{(\mathbf{x}^i, y^i)\}_{i=1}^{n_l}$, we have access to an unlabelled sample $\mathbf{X}^u = \{\mathbf{x}^i\}_{i=n_l+1}^{n_l+n_u}$ from the same distribution P at training time.[1] It is usually assumed that $n_l \ll n_u$. At test time, the task is to predict targets Y from inputs X. If predictions are made on the unlabelled training data only we speak of *transductive* learning (Vapnik, 1998). The aim and hope of SSL is that additional unlabelled data helps in making better predictions. $\mathbf{X}^u$ can improve the estimate of $P(X)$, but SSL aims at improving $P(Y|X)$. This can only work if there is a link between $P(X)$ and $P(Y|X)$. Indeed, many approaches to SSL establish such a link through additional assumptions (Chapelle et al., 2010; Mey and Loog, 2019; Zhu, 2005). Two common ones are the *cluster assumption*, positing that points in the same cluster of $P(X)$ have the same label Y; and *low-density separation*, stating that class boundaries of $P(Y|X)$ should lie in an area where $P(X)$ is small. For original references, as well as for a discussion of how these assumptions relate to various SSL methods refer to Chapelle et al. (2010).

We briefly mention four of the more common methods, starting with self-learning (sometimes also called the Yarowsky-algorithm). This is a wrapper algorithm that initialises the learner based on the labelled data, updates the labels for the unlabelled data, and then retrains based on all labelled data available, possibly iterating this procedure (Abney, 2004; Blum and Mitchell, 1998; Scudder, 1965). Secondly, generative model approaches maximise the likelihood of a generative model

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}^l, \mathbf{y}^l, \mathbf{X}^u | \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{y}^u \in \mathcal{Y}^{n_u}} p(\mathbf{X}^l, \mathbf{y}^l, \mathbf{X}^u, \mathbf{y}^u | \boldsymbol{\theta}). \end{aligned} \quad (1)$$

While this is a hard optimisation problem due to the latent variables $\mathbf{y}^u$, a local optimum can be found via the expectation maximisation algorithm (EM) (Dempster et al., 1977). The third class of common methods are the graph-based approaches. These construct a similarity-based graph representation of the data and propagate labels to neighbours in this graph (Zhou et al., 2004; Zhu and

Ghahramani, 2002; Zhu et al., 2003). Finally, transductive SVMs maximise a (soft) margin over labelled and unlabelled data while minimising a regularised risk on the labelled data (Joachims, 1999; Vapnik, 1998).

2.2 CAUSALITY

Despite data showing a positive correlation between chocolate consumption and the number of Nobel prizes per capita (Messerli, 2012), we would not expect that force-feeding the population with chocolate would result in higher research output. The correlation in this example may make chocolate consumption a useful predictor in an i.i.d. setting, but it does not allow one to answer interventional questions of the form "*what would happen if we actively changed some of the variables?*".

This notion of intervention is at the heart of the difference between correlation and causation. While much of machine learning is concerned with using correlations between variables to make predictions, Reichenbach (1956) has argued that such correlations always result from underlying causal relationships: statistical dependence is an epiphenomenon—a by-product of a causal process.[2]

Structural causal model (SCM) To reason about causality in SSL, we adopt the structural causal model (SCM) framework (Pearl, 2000) which defines a causal model over a set of observed variables $\{X_1, ..., X_d\}$ to consist of (i) a collection of structural assignments,

$$X_i := f_i(\mathrm{PA}_i, N_i) \qquad \text{for} \qquad i = 1, ..., d, \quad (2)$$

where f_i are deterministic functions computing X_i from its causal parents $\mathrm{PA}_i \subseteq \{X_1, ..., X_d\} \setminus X_i$ and an exogenous noise variable N_i; and (ii) a factorising joint distribution P over the unobserved noise variables,

$$P(N_1, ..., N_d) = \prod_{i=1}^{d} P(N_i).$$

This assumption of mutually independent noises entails that all causes common to any pair of observed variables are included in the model (i.e., there are no hidden confounders), and is referred to as *causal sufficiency*. Together, (i) and (ii) define a causal generative process and imply an observational joint distribution over $X_1, ..., X_d$ which factorises over the induced causal graph[3] $\mathcal{G}$ as:

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i | \mathrm{PA}_i). \quad (3)$$

---

[1]In the statistics literature this setting is also known as data *missing completely at random* (MCAR).

[2]For the given example, a possible explanation for the observed correlation is a healthy economy acting as common cause for both chocolate consumption and a good education system.

[3]The induced causal graph $\mathcal{G}$ is obtained by drawing a directed edge from each node in PA_i (i.e, the direct causes of X_i) to X_i for all i. We assume throughout that $\mathcal{G}$ is acyclic.
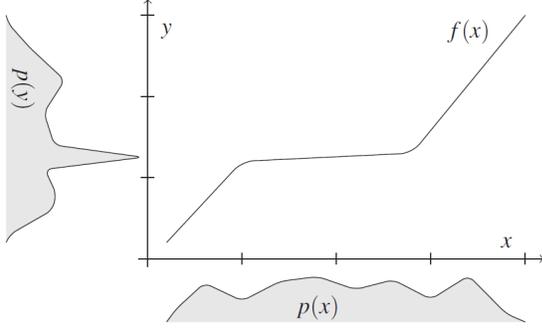
Figure 1: Illustration of the ICM principle for the setting $X \to Y$. If the distribution of the cause $p(x)$ is chosen independently of the mechanism $f : \mathcal{X} \to \mathcal{Y}$ representing $p(y|x)$, then such independence is violated in the backward (non-causal) direction: $p(y)$ has large density where f has small slope and thereby contains information about $f^{-1} = p(x|y)$. Figure from Janzing et al. (2012).

Principle of independent causal mechanisms (ICM) Motivated by viewing the f_i in (2) as independent physical mechanisms of nature, the principle of independent causal mechanisms (ICM) states that "*the causal generative process (...) is composed of independent and autonomous modules that do not inform or influence each other*" (Peters et al., 2017). In other words, the conditional distributions of each variable given its causal parents, $P(X_i|\text{PA}_i)$ in (3), are independent objects which do not share any information. Importantly, this notion of independence is different from *statistical* independence of random variables (indeed, the variables can still be statistically dependent). Instead, it should be understood as an *algorithmic* independence on the level of distributions.

Intuitively, two distributions are algorithmically independent if encoding them jointly does not admit a shorter description than describing each of them separately. In this case, we say that they do not share information. This notion has been formalised in terms of Kolmogorov complexity (or algorithmic information) $K(\cdot)$ by Janzing and Schölkopf (2010): two distributions $P(X)$ and $P(Y|X)$ are considered algorithmically independent if and only if

$$K\big(P(X,Y)\big) \overset{+}{=} K\big(P(X)\big) + K\big(P(Y|X)\big), \quad (4)$$

where the notation $\overset{+}{=}$ refers to a constant due to the choice of a Turing machine in the definition of algorithmic information. Note that the RHS of (4) is always greater or equal to the LHS, and equality indicates that there is no redundant information which could be compressed by describing the two distributions jointly.

Crucially, if the ICM principle holds, i.e., the conditionals in the causal factorisation (3) do not share information, this independence is generally violated for other non-causal factorisations. This is illustrated in Fig. 1 for the bivariate setting[4] of $X \to Y$ using the IGCI model of Janzing et al. (2012) in which a deterministic, invertible function f generates effect Y from cause X. If the input distribution of the cause, $p(x)$, is chosen independently from the mechanism $f(x)$ (or more generally, $p(y|x)$), then this independence is violated in the backward (non-causal) direction, since $p(y)$ has a large density where f has small slope (see Fig. 1), and $p(y)$ thereby contains information about $f^{-1}(y) = p(x|y)$.[5]

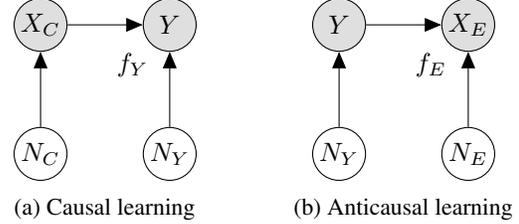(a) Causal learning   (b) Anticausal learning

Figure 2: Assuming ICM, in a causal learning setting (a) SSL should be impossible as $P(X_C)$ contains no information about $P(Y|X_C)$, whereas in an anticausal learning setting (b) $P(X_E)$ may contain information about $P(Y|X_E)$ and SSL is thus, in principle, possible.

2.3 SSL IN CAUSAL AND ANTICAUSAL LEARNING SETTINGS

For the task of predicting a target variable Y, Schölkopf et al. (2012) distinguish between *causal learning* (Fig. 2a) where all features X_C are causes of Y, i.e.,

$$X_C := N_C,$$
$$Y := f_Y(X_C, N_Y),$$

and *anticausal learning* (Fig. 2b) where all features X_E are effects of Y, i.e.,

$$Y := N_Y,$$
$$X_E := f_E(Y, N_E).$$

In a causal learning setting, it then follows from the ICM principle that $P(X_C)$ and $P(Y|X_C)$ are algorithmically independent, i.e., share no information. Recalling the goal of improving $P(Y|X)$ from $P(X)$, SSL should thus be impossible. In the anticausal direction, on the other hand, this algorithmic independence relation is between $P(Y)$ and $P(X_E|Y)$. Hence, $P(X_E)$ may (and in some cases provably will, see Daniušis et al., 2010) share information with $P(Y|X_E)$, and SSL is thus, in principle, possible.

---

[4]where ICM reduces to an *independence of cause and mechanism* (Daniušis et al., 2010; Lemeire and Dirkx, 2006)

[5]The amount of shared information can be quantified as $\text{cov}[\log g', p_Y] := \int_0^1 \log g'(y) p_Y(y) dy - \int_0^1 \log g'(y) dy \geq 0$, where $g = f^{-1}$, with equality iff. f is the identity (Janzing et al., 2012); note that, by assumption, $\text{cov}[\log f', p_X] = 0$.

3 SSL WITH CAUSE AND EFFECTS FEATURES

In this work, we consider a semi-supervised learning setting where both causes and effects of the target Y are available as features. Specifically, we assume that we are given a small labelled sample $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l) = \{(\mathbf{x}_c^i, y^i, \mathbf{x}_e^i)\}_{i=1}^{n_l}$ and a large unlabelled sample $(\mathbf{X}_C^u, \mathbf{X}_E^u) = \{(\mathbf{x}_c^i, \mathbf{x}_e^i)\}_{i=n_l+1}^{n_l+n_u}$ generated from an SCM of the form:

$$X_C := N_C, \tag{5}$$
$$Y := f_Y(X_C, N_Y), \tag{6}$$
$$X_E := f_E(X_C, Y, N_E). \tag{7}$$

This causal model is shown in Fig. 3. We refer to X_C as causal features and X_E as effect features and assume this partitioning to be known a priori (e.g., think of the medical example with risk factors and diagnostic tests).

This setting generalises the cases of causal and anticausal learning considered by Schölkopf et al. (2012) without positing any new statistical (conditional) independencies—recall that assumptions lie in the *missing* arrows, not in the *present* ones.[6] In particular, the set of causal features X_C may also contain "spouse features" $X_S \subseteq X_C$ which do not directly influence the target, $X_S \not\to Y$, but only its effects, $X_S \to X_E$. All features in the Markov blanket of Y can thus be expressed as members of either X_C or X_E, so that our setting remains widely applicable.

Analogous to (3), the SCM (5)–(7) induces an observational distribution which factorises over the causal graph:

$$P(X_C, Y, X_E) = P(X_C)P(Y|X_C)P(X_E|Y, X_C). \tag{8}$$

Following Schölkopf et al. (2012), we assume the principle of ICM, i.e., that the factors on the RHS of (8) constitute algorithmically-independent causal mechanisms.

Our goal is to predict the target Y from features (X_C, X_E), so we are interested in estimating

$$p(y|x_C, x_E) = \frac{p(y|x_C)p(x_E|y, x_C)}{\sum_{y' \in \mathcal{Y}} p(y'|x_C)p(x_E|y', x_C)} \tag{9}$$

while having additional information about $P(X_C, X_E)$ from unlabelled data.

In analogy to the case of causal learning (see §2.3), by the ICM principle, the distribution over causes $P(X_C)$ does not contain any information about $P(Y|X_C)$

Figure 3: In the SSL setting considered in this work, features can be partitioned into two disjoint sets: potential causes X_C and potential effects X_E of the target Y.

or $P(X_E|Y, X_C)$ (see RHS of (8)), and thereby also not about $P(Y|X_C, X_E)$ (see RHS of (9)). Indeed, $P(Y|X_C, X_E)$ is completely determined by the structural equations (6) for Y as a function of X_C, and (7) for X_E as a function of X_C and Y, and does not depend on (5), i.e., what distribution of causal features X_C is fed into this generative process.[7]

Having established that $P(X_C)$ does not contain useful information for our task,[8] we are left with $P(Y, X_E|X_C)$ which according to the chain rule of probability admits two possible factorisations,

$$P(Y, X_E|X_C) = P(Y|X_C)P(X_E|X_C, Y) \tag{10}$$
$$P(Y, X_E|X_C) = P(X_E|X_C)P(Y|X_C, X_E). \tag{11}$$

Equation (10) is a *causal* factorisation into independent mechanisms which do not share any information. Equation (11), however, corresponds to a *non-causal* factorization, implying that the factors on the RHS may share information. Since we care about estimating $P(Y|X_C, X_E)$ and we have additional information about $P(X_E|X_C)$ from unlabelled data, it is precisely this potential dependence or link between $P(Y|X_C, X_E)$ and $P(X_E|X_C)$ that SSL approaches should aim to exploit in our setting (Fig. 3). We formulate this result as follows.

Main insight. When learning with both causes and effects of a target as captured by the causal model in (5)–(7), $P(X_E|X_C)$ contains all relevant information provided by additional unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$ about $P(Y|X_C, X_E)$. Therefore, SSL approaches for such a setting should aim at exploiting this information by linking these two conditional distributions via suitable additional assumptions.[9]

---

[6]For example, omitting the link $X_C \to X_E$ renders the two feature sets conditionally independent given Y (von Kügelgen et al., 2019), which is restrictive for realistic scenarios, and can already be well addressed by approaches like co-training (Blum and Mitchell, 1998) which are tailored to such assumptions.
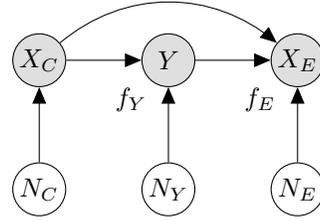
[7] This suggests the possibility to perform SSL with covariate-shifted causal features, see, e.g., von Kügelgen et al. (2019).

[8]Since we generally aim to minimise a risk, or *expected* loss, $P(X_C)$ can still help get a better estimate of the expectation operator (Peters et al., 2017). By useful information here we mean information about $P(Y|X_C, X_E)$, see §7 for further discussion.

[9]Our statement is about how SSL should proceed on the level of distributions, i.e., in the general case; to develop specific SSL algorithms, more concrete assumptions are necessary.

We remark that this contains previous results for SSL in causal and anticausal learning settings as special cases: in absence of causal features (i.e., for anticausal learning) $P(X_E|X_C)$ reduces to the known setting of $P(X_E)$ containing information about $P(Y|X_E)$, whereas in absence of effect features (i.e., for causal learning) $P(X_E|X_C)$ becomes meaningless, and SSL thereby impossible, both consistent with the findings of Schölkopf et al. (2012).

However, our result goes further than this: having additional unlabelled data of both cause and effect features can be strictly more informative than having only unlabelled effects. To illustrate this point, consider the following thought experiment. Suppose that X_E is a noisy copy or proxy label of Y, i.e., $X_E := Y + N_E$. In this case, unlabelled data contains information which is very similar to the information contained in the labelled data. Learning to predict X_E from X_C requires predicting Y from X_C and can thus be very helpful to solve the problem.

4 NEW ASSUMPTIONS FOR SSL

We now use our insight to reformulate, or refine, standard SSL assumptions (see §2.1) for the setting of Fig. 3 where both X_C and X_E are observed. Our aim is to adapt these assumptions such that they make use of potential information shared between $P(Y|X_C, X_E)$ and the conditional $P(X_E|X_C)$—as opposed to the marginal $P(X_C, X_E)$.

While the previous analysis (§3) applies to general prediction tasks including regression, we now focus on classification. For conceptual simplicity and ease of illustration, we will assume binary classification in what follows, but extensions to the multi-class setting are straightforward. For binary $Y \in \{0, 1\}$, we can rewrite (6) and (7) as:

$$Y := \mathbb{I}\{g(X_C) > U\} \tag{12}$$
$$X_E := Y f_1(X_C, N_E) + (1 - Y) f_0(X_C, N_E) \tag{13}$$

where $\mathbb{I}$ is the indicator function and U is a uniform random variable on $[0, 1]$, so that $g(X_C)$ computes $P(Y = 1|X_C)$. Allowing arbitrary g, f_0, f_1 and N_E, this comes without loss of generality.

4.1 CONDITIONAL CLUSTER ASSUMPTION

While the standard cluster assumption advocates for sharing labels within clusters in the marginal distribution of all features, in view of the above we postulate that *points in the same cluster of $X_E|X_C$ share the same label Y*. We refer to this as the *conditional cluster assumption*.

One can think of clusters of $X_E|X_C$ by considering the space $\mathcal{F}$ of functions $f : \mathcal{X}_C \to \mathcal{X}_E$ computing effects from causes. Different functions in this space correspond, e.g., to different choices of Y and N_E in (13). The conditional cluster assumption can then be understood as
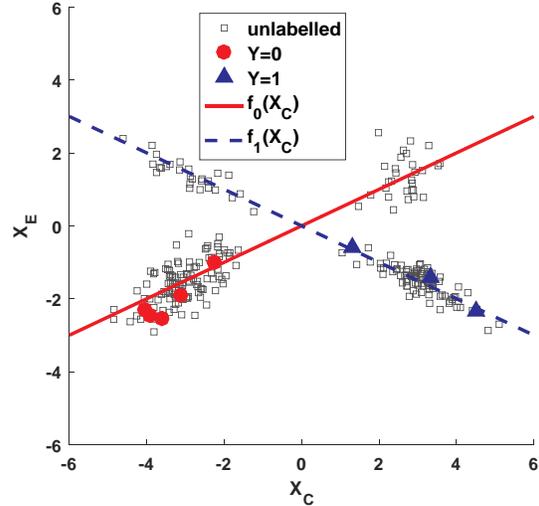

Figure 4: An example dataset arising from our setting with linear f_0 and f_1 in (13). The conditional cluster assumption links class labels to membership in clusters of $X_E|X_C$, suggesting to classify unlabelled points $(\mathbf{x}_C, \mathbf{x}_E)$ according to whether $\mathbf{x}_E$ is better explained by $f_0(\mathbf{x}_C)$ (solid red line) or $f_1(\mathbf{x}_C)$ (dashed blue line). Conventional SSL methods can easily fail in this case: label propagation or maximum margin would both classify by $\text{sgn}(x_C)$ yielding $\approx 50\%$ error.

saying that the class-dependent mechanisms f_0 and f_1 correspond to cluster centroids in $\mathcal{F}$.[10]

This idea is illustrated in Fig. 4 for linear f_0 and f_1. In this simple example, knowing that X_E causally depends on X_C (combined with assumptions about the type of dependence, see §5.2) can help identify the true mechanisms (solid red and dashed blue lines), and therefore the correct labelling. Standard SSL approaches (see §2.1) agnostic to the causal structure, on the other hand, can easily fail in this situation: large-margin methods (TSVMs) learn to classify Y by the sign of X_C (i.e., using the boundary $X_C = 0$), while (unconditional) clustering or graph-based approaches yield a similar classification due to measuring similarity in the joint feature space (X_C, X_E), rather than in terms of the conditional $X_E|X_C$;[11] in either case this leads to an error rate of almost 50%.

More generally, knowing the causal partitioning of features into X_C and X_E introduces an asymmetry between features that is not present in standard SSL methods. In

---

[10]Note that due to the general form of (13), it is possible to have more than one cluster per class in $\mathcal{F}$. For handwritten digits, for example, N_E could act as a switch between 7s with and without the horizontal stroke.

[11]E.g., the points $(-4, -2)$ and $(4, 2)$ are far away in terms of marginal distance, but close in terms of the conditional as measured, e.g., by $\|x_E - f_0(x_C)\|$ with $f_0(x) = 0.5x$, see §5.2.
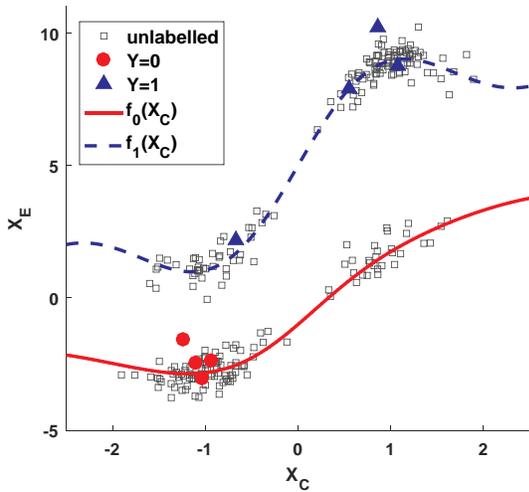
Figure 5: An example with non-linear mechanisms f_0 and f_1 in (13). Under additional assumptions such as additive unimodal noise, the causal structure induces an asymmetry between features which imposes constraints on the cluster assignments. E.g., the two clusters around $X_C = 1$ cannot both be explained by f_1, though this would be possible with the roles of X_C and X_E interchanged (i.e., if the figure were rotated by $90°$).

particular, the fact that functions are not allowed to be *one-to-many* provides a constraint. This becomes apparent for non-linear mechanisms f_0, f_1 as shown in Fig. 5. There, the roles of X_C and X_E cannot be exchanged (i.e., the figure rotated by $90°$) since f_0 and f_1 are not invertible, thus restricting the possibilities of clustering the data.

4.2 LOW-CONDITIONAL-DENSITY SEPARATION ASSUMPTION

In a similar vein to §4.1, we can also adapt the low-density separation assumption to our setting. While in its original form, low-density separation is a statement about the joint density of all features, we have argued that (subject to the ICM principle) $P(X_C)$ contains no information about $P(Y|X_C, X_E)$, but that the conditional $P(X_E|X_C)$ may do so. We therefore propose that a more justified notion of separation is that *class boundaries of $P(Y|X_C, X_E)$ should lie in regions where $P(X_E|X_C)$ is small*. We refer to this is as *low-conditional-density separation*.

5 ALGORITHMS

While the main contribution of this work is conceptual, it is illustrative to discuss the implication of our assumptions from §4 for some of the standard approaches to SSL introduced in §2.1 and propose variations thereof which explicitly aim to make use of the information shared between the two conditionals $P(X_E|X_C)$ and $P(Y|X_C, X_E)$.

5.1 SEMI-GENERATIVE MODELS

First, we reconsider the generative approach to SSL. While a naïve model would unnecessarily fit the full distribution—including the uninformative part $P(X_C)$—this approach to SSL is easily adapted to our new assumptions by only modelling the informative part of the generative process, $P(Y, X_E|X_C)$. This type of *semi-generative model* has been introduced by von Kügelgen et al. (2019) in the context of domain adaptation, and under the restrictive assumption of conditionally independent feature sets, $X_C \perp\!\!\!\perp X_E|Y$. Here, we relax this assumption, and propose a new approach to fit the resulting model.

Given a semi-generative model parameterised by $\boldsymbol{\theta}$, a maximum likelihood approach similar to (1) then yields

$$\arg\max_{\boldsymbol{\theta}} p(\mathbf{y}^l, \mathbf{X}_E^l|\mathbf{X}_C^l; \boldsymbol{\theta}) \sum_{\mathbf{y}^u} p(\mathbf{y}^u, \mathbf{X}_E^u|\mathbf{X}_C^u; \boldsymbol{\theta}).$$

Equivalently, we minimise the negative log-likelihood (NLL) which *for fixed labels* decomposes according to (8) into two separate terms which can be optimised independently for $\boldsymbol{\theta}_Y$ and $\boldsymbol{\theta}_E$:

$$\mathrm{NLL}(\boldsymbol{\theta}|\mathbf{X}_C, \mathbf{y}, \mathbf{X}_E) := -\log p(\mathbf{y}, \mathbf{X}_E|\mathbf{X}_C; \boldsymbol{\theta})$$
$$= -\log p(\mathbf{y}|\mathbf{X}_C; \boldsymbol{\theta}_Y) - \log p(\mathbf{X}_E|\mathbf{X}_C, \mathbf{y}; \boldsymbol{\theta}_E)$$

This separation leads us to an EM-like approach (Dempster et al., 1977) to find a local optimum of the NLL by iteratively computing the expected label given the current parameters (E-step), and then minimising the NLL w.r.t. to the parameters keeping the labels fixed (M-step). For the setting of hard labels (i.e., $y \in \{0, 1\}$), this procedure is summarised in Algorithm 1 where for brevity we omit explicitly conditioning the NLL on $\mathbf{X}_C$ and $\mathbf{X}_E$, and where *converged* means $\boldsymbol{y}^{(t)} = \boldsymbol{y}^{(t-1)}$. For the specific case of logistic regression for $P(Y|X_C)$ and a class-dependent linear Gaussian model for $P(X_E|X_C, Y)$ we provide a more detailed procedure for both soft and hard labels in Algorithm 3 in Appendix A.

5.2 CONDITIONAL SELF-LEARNING

The second algorithm we propose aims to use the conditional cluster assumption without postulating a generative model in parametric form. It is loosely related to the ideas of label propagation and self-learning (Scudder, 1965; Zhu and Ghahramani, 2002). However, instead of propagating labels based on similarities between points computed in the joint feature space (X_C, X_E) as in the conventional approach, we instead focus on extracting information contained in $P(X_E|X_C)$. To this end, we need to place restrictions on the class of functions f_0, f_1 allowed in (13). In the following, we therefore assume an *additive noise model* (Hoyer et al., 2009) defined by

$$f_i(X_C, N_E) = f_i(X_C) + N_{E,i} \quad \text{for} \quad i = 0, 1. \quad (14)$$

Algorithm 1: EM-like algorithm for fitting a semi-generative model by maximum likelihood

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; parametric models $p(y|\mathbf{x}_C; \boldsymbol{\theta}_Y)$ and $p(\mathbf{x}_E|\mathbf{x}_C, y; \boldsymbol{\theta}_E)$

Output: fitted labels $\mathbf{y}^u$; estimates $\hat{\boldsymbol{\theta}}_Y, \hat{\boldsymbol{\theta}}_E$

1 $t \leftarrow 0$
2 $\hat{\boldsymbol{\theta}}_Y^{(0)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_Y|\mathbf{y}^l)$
3 $\hat{\boldsymbol{\theta}}_E^{(0)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_E|\mathbf{y}^l)$
4 **while** *not converged* **do**
5 $\quad \mathbf{y}^{(t)} \leftarrow \mathbb{I}\{p(\mathbf{y}|\mathbf{X}_C^u, \mathbf{X}_E^u; \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}) > 0.5\}$
6 $\quad \hat{\boldsymbol{\theta}}_Y^{(t+1)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_Y|\mathbf{y}^l, \mathbf{y}^{(t)})$
7 $\quad \hat{\boldsymbol{\theta}}_E^{(t+1)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_E|\mathbf{y}^l, \mathbf{y}^{(t)})$
8 $\quad t \leftarrow t+1$
9 **end**
10 **return** $\mathbf{y}^{(t-1)}, \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}$

Algorithm 2: Conditional self-learning

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data; $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; `regress()` method

Output: fitted labels $\mathbf{y}^u$; functions $\hat{f}_0, \hat{f}_1$

1 $t \leftarrow 0$
2 **while** *unlabelled data left* **do**
3 $\quad$ **for** $i = 0, 1$ **do**
4 $\quad\quad \hat{f}_i^{(t)} \leftarrow \texttt{regress}(\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l)$
5 $\quad\quad \mathbf{r}_i \leftarrow ||\mathbf{X}_E^u - \hat{f}_i^{(t)}(\mathbf{X}_C^u)||^2$
6 $\quad$ **end**
7 $\quad (i,j) \leftarrow \arg\min\{\mathbf{r}_{i,j} : i = 0, 1; j = 1, ..., n_u\}$
8 $\quad y^{n_l+j} \leftarrow i$
9 $\quad \mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l \leftarrow \texttt{append}(\mathbf{x}_E^{n_l+j}, \mathbf{x}_C^{n_l+j})$
10 $\quad t \leftarrow t+1$
11 **end**
12 **return** $\mathbf{y}^u, \hat{f}_0^{(t-1)}, \hat{f}_1^{(t-1)}$

Fig. 5 shows an example of data generated in this way.[12] Note that, unlike in the probabilistic approach of §5.1, we do not make additional assumptions about the exact noise distribution, such as Gaussianity. We do, however, assume that the noise has zero mean and is unimodal, so that there is one function from X_C to X_E for each label.

Our approach then aims at learning these functions, and can be summarised as follows. First, we initialise $\hat{f}_0$ and $\hat{f}_1$ from labelled data by regressing X_E on X_C. Next, we compute the predictions of the $\hat{f}_i$ on the unlabelled sample, label the point with smallest prediction error as the respective class, and use it to update the corresponding $\hat{f}_i$. This procedure is repeated until all initially unlabelled points are labelled. We refer to this approach as *conditional self-learning*, and summarise it in Algorithm 2. The notation $\mathbf{X}_{C,i}$ and $\mathbf{X}_{E,i}$ refers to those samples of X_C and X_E with label $Y = i$, $\mathbf{r}_i$ in step 5 denotes the vector of residuals from regressing $\mathbf{X}_E^u$ on $\mathbf{X}_C^u$ using $\hat{f}_i$, and $\mathbf{r}_{i,j}$ in step 7 denotes the jth such residual.

Connection to the probabilistic approach It is also possible to use the above approach with soft labels (as often done in conventional label propagation (Zhou et al., 2004; Zhu et al., 2003)) by using a weighted regression scheme. This requires a method of computing regression weights from prediction errors of $\hat{f}_0$ and $\hat{f}_1$, though, and therefore needs additional assumptions or heuristics. We note that choosing a particular distribution for N_E and using $P(Y|X_C)$ as a class prior leads to a soft-label EM approach similar to §5.1, see Algorithm 3 in Appendix A.

---

[12]Other choices are, of course, possible: another interesting and more flexible option is the post-nonlinear model of Zhang and Hyvärinen (2009), $g_i(f_i(X_C) + N_{E,i})$.

We therefore presently restrict ourselves to hard labels.

Connection to competition of experts While it is conceptually based on the ICM assumption and an analysis of the causal structure among the feature set, the conditional self-learning approach is linked to a number of known methods, including not only self-learning, but also methods building on *competition of experts*, as recently applied to the problem of learning causal mechanisms. In this work, the functions f_0, f_1 are generative models competing for data that has undergone unknown transformations, and eventually each specialising on how to invert one of those transformations (Parascandolo et al., 2018).

6 EXPERIMENTAL RESULTS

To corroborate our analysis with empirical evidence, we evaluate our algorithms from §5 on synthetic data as well as on two medical datasets from the UCI repository. We compare with T-SVMs (Joachims, 1999; Vapnik, 1998) with linear and RBF kernels using the `q3svm` implementation (Gieseke et al., 2014); and with label propagation (Zhou et al., 2004; Zhu et al., 2003) using the implementation in `scikit-learn` (Pedregosa et al., 2011). We use the default hyper-parameters in all cases. For our conditional self-learning algorithm, we use linear ridge regression with default regularisation strength 1, and for the EM algorithms we use logistic regression for $Y|X_C$ and linear, class-dependent Gaussians for $X_E|X_C, Y$, see Algorithm 3 and Appendix B.1 for details.

Synthetic data As controlled environments, we generate various synthetic datasets with cause and effect features of three different types (S1, S2, S3): S1 represents linearly-separable data, S2 corresponds to datasets with

Table 1: Average accuracies on unlabelled data (higher is better) $\pm$ one standard deviation across 100 runs, each time randomly drawing 10 (for S1, S2, S3) or 20 (for Pima Diabetes and Heart Disease) new labelled and 200 new unlabelled samples. Results refer to transductive evaluation for ease of comparison with other methods. The three best methods for each dataset are highlighted in bold. The last four rows are our causally-motivated methods. The "-" indicates that label propagation did not converge on S3, and (sup.) indicates purely supervised baselines.

Method	S1 (linear)	S2 (non-linear)	S3 (multi-dim.)	Pima Diabetes	Heart Disease
Lin. log. reg. (sup.)	.968 $\pm$ .023	.823 $\pm$ .080	.945 $\pm$ .039	.626 $\pm$ .058	.526 $\pm$ .066
Lin. T-SVM	.865 $\pm$ .093	.878 $\pm$ .074	.822 $\pm$ .117	.602 $\pm$ .065	**.746 $\pm$ .060**
RBF T-SVM	.863 $\pm$ .094	.876 $\pm$ .075	.821 $\pm$ .116	.601 $\pm$ .064	**.745 $\pm$ .060**
RBF label propag.	.924 $\pm$ .082	.909 $\pm$ .065	-	.650 $\pm$ .030	.528 $\pm$ .068
Semi-gen. (sup.)	.968 $\pm$ .076	**.935 $\pm$ .074**	.949 $\pm$ .082	**.669 $\pm$ .064**	.550 $\pm$ .096
Semi-gen.+soft EM	**.986 $\pm$ .081**	**.989 $\pm$ .024**	**.991 $\pm$ .067**	**.661 $\pm$ .063**	.518 $\pm$ .050
Semi-gen.+hard EM	**.985 $\pm$ .079**	**.972 $\pm$ .058**	**.987 $\pm$ .076**	**.695 $\pm$ .064**	.518 $\pm$ .050
Cond. self-learning	**.980 $\pm$ .052**	.923 $\pm$ .090	**.961 $\pm$ .069**	.659 $\pm$ .079	**.719 $\pm$ .076**

a non-linear decision boundary of which the data shown in Figure 5 is an example, and S3 is a version of S2 with multi-dimensional features. Details of how exactly synthetic data is generated are provided in Appendix B.1.

Medical data As real-world data, due to the fact that both plausibly contain cause and effect features, we choose the two medical datasets *Pima Indians Diabetes* (Smith et al., 1988) and *Heart Disease* (Detrano et al., 1989). We select those features which are most strongly correlated with the target variable ($p < 0.01$), and categorise them into cause and effect features to the best of our knowledge, see Appendix B.2 for details.

Results The results of our experiments are summarised in Table 1, see the table caption for simulation details. On the synthetic datasets, our causally-motivated methods outperform the purely supervised logistic regression baseline as well as the other SSL approaches, which in the case of S1 and S3 even lead to decreased accuracy. The probabilistic approaches perform particularly well on the synthetic datasets, which was expected since the generative model for these cases was specified by us, and its correct form thus known. Conversely, on the Heart Disease dataset they degraded performance, presumably due to model misspecification.

Conditional self-learning, based on the weaker assumption of additive noise, performs competitively, in particular on the real data. Notably, it is the only method that improves upon the supervised logistic-regression baseline (i.e., achieves SSL) for all five datasets considered.

As a sanity check, we also ran our causally inspired methods with the roles of cause and effect features exchanged which led to deteriorated performance (results not shown). This observation provides additional support for our anal-

ysis that algorithmic information is shared only between the conditionals $P(Y|X_C, X_E)$ and $P(X_E|X_C)$.

7 DISCUSSION

The present paper looks at SSL from the perspective of causal modelling. We argue that if we know that the input X can be partitioned into cause and effect features (X_C and X_E), then this has surprising theoretical implications for how SSL should utilise unlabelled data: rather than simply exploiting links between the marginal $P(X)$ and the conditional $P(Y|X)$ (as formalised, e.g., in the standard SSL cluster assumption), one should exploit links between two conditional distributions, $P(X_E|X_C)$ and $P(Y|X_C, X_E)$. In other words, SSL should not blindly be applied to the joint feature set, but instead only to effect (or confounded, see below) features while conditioning on purely causal ones.[13] Note that we view this not as a contradiction to the usual cluster assumption, but rather as an explication or refinement thereof, taking into account the causal structure; indeed, it subsumes SSL in the anticausal setting as a special case. It does not subsume SSL in the causal setting, but, as argued by Schölkopf et al. (2012), SSL is futile in this case.

Impossibility of SSL in the causal direction For certain settings, SSL yields provable improvements over a purely-supervised solution, irrespective of the causal structure, in terms of the surrogate loss in a transductive setting (Krijthe and Loog, 2017, 2018; Loog, 2015). For linear discriminant analysis, for instance, such improvements are obtained almost surely. This may appear at

---

[13]Interestingly, such conditioning may allow to extend SSL to settings where the missingness mechanism depends on X_C, i.e., covariate-shifted or missing-at-random (MAR) data (Mohan et al., 2013; Sugiyama and Kawanabe, 2012).

odds with the result that SSL in the causal direction is impossible. The apparent contradiction can, however, be explained as follows. The claimed lack of useful information (subject to ICM) contained in $P(X_C)$, as used in the analysis of Schölkopf et al. (2012) and in the present work, refers to information useful *for improving our estimate of* $P(Y|X)$. However, the success of SSL is often measured instead in terms of the *risk*, i.e., the expected loss w.r.t. the *joint* distribution. Thus, even though knowing $P(X_C)$ in a causal learning setting does not help make better predictions *for any given* x, it may still help to achieve a lower risk. In a sense, knowing $P(X_C)$ helps decide which parts of the support are most important to get right. This, in turn, can be particularly helpful when dealing with model misspecification (Urner et al., 2011). For further discussion of this point, we refer to Peters et al. (2017, Section 5.1.2). Moreover, we remark that most settings for which SSL has proven particularly useful empirically fall in the category of anticausal learning problems, e.g, recognising digit identity Y from its handwritten realisation X_E (e.g., Kingma et al., 2014) or classifying document topics Y from the contained words X_E (e.g., Nigam et al., 2006).

Validity of ICM and the role of latent confounding
Our work uses the assumption of independent causal mechanisms (ICM) generating the data. While well-motivated from the independence of separate physical processes, in practice, the ICM principle may not hold *exactly*: for example, some mechanisms may have co-evolved to be algorithmically dependent (Peters et al., 2017). While latent confounding does not *a priori* invalidate our assumption—recall that ICM is not a statement about *statistical* but *algorithmic* independence on the level of mechanisms—a confounder H of X_C and Y can nonetheless introduce such dependence due to its shared influence on the two distributions:

$$P(X_C) = \int P(X_C|H)\,dP(H),$$
$$P(Y|X_C) = \int P(Y|X_C, H)\,dP(H).$$

Methods for detecting hidden confounding (e.g., Hoyer et al., 2008; Janzing and Schölkopf, 2018) can potentially help identify such confounded causal features, which should then also be treated as containing useful information. In weakly-confounded settings, however, ICM can still be a useful principle, and has proven to be so in applications, e.g., in cause-effect inference. Moreover, many causal features such as age or sex in our medical example can reasonably be assumed to not be caused by other variables, and can therefore safely be treated as unconfounded. We also note that since SSL is about prediction in an i.i.d. setting, latent confounding is generally less problematic than for classical causal inference tasks, such as predicting effects of interventions. Considering features which are neither causes nor effects, but are correlated to the target by unobserved confounders, constitutes

an interesting extension for future work.

Known causal partitioning Another important assumption for our analysis is knowledge about the causal structure. While we do not address the separate problem of causal discovery, we argue that often sufficient domain knowledge is available to identify maybe not a fine-grained causal graph, but at least a partitioning of features into disjoint sets of potential causes and potential effects. The medical setting described in the introduction constitutes such a real-world example. We also note that under certain assumptions (e.g., additive noise, as for the version of conditional self-learning presented here) causal direction can be identified from purely observational data (Hoyer et al., 2009). Such techniques applied to $\{X_C, X_E\}$ may help reveal the causal nature of features, even with only few labels.

Connections to domain adaptation Since our proposed approach is robust to (domain-induced) changes in $P(X_C)$, provided that $P(Y|X_C)$ and $P(X_E|Y, X_C)$ remain invariant, exploring it further in the context of domain adaptation constitutes an interesting future direction (see also footnotes 7 and 13). Data from different environments may provide valuable causal clues that can potentially be leveraged by constraint- (Huang et al., 2019; Magliacane et al., 2018), invariance- (Rojas-Carulla et al., 2018), or variance-penalty–based approaches (Heinze-Deml and Meinshausen, 2017) to learn the underlying causal structure and identify subsets of features for which $P(Y|X_C)$ and $P(X_E|Y, X_C)$ remain stable.

8   CONCLUSION

While the present analysis is intriguing and points out a previously unexplored link between two conditional distributions, the jury is still out on how in general to best exploit unlabelled data in machine learning. The present insight is but one step, and in particular, while encouraging, the algorithms and experiments based upon it can only be a starting point (see Appendix C for proposed extensions). They may lead to new approaches that make explicit use of causal structure and exploit the conditional cluster assumption (and potentially the low-conditional-density separation assumption) in more elegant and effective ways. Ultimately, the value of novel assumptions and conceptual models lies in whether they provide a fertile basis to inspire further algorithm development and theoretical understanding. We hope that the present ideas and analysis will constitute such a contribution.

References

Abney, S. (2004). "Understanding the Yarowsky Algorithm". In: *Comput. Linguist.* 30.3, pp. 365–395.

Ben-David, S., T. Lu, and D. Pál (2008). "Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning". In: *Proceedings of the 21st Annual Conference on Learning Theory*. Helsinki, Finland.

Blum, A. and T. Mitchell (1998). "Combining labeled and unlabeled data with co-training, semi-supervised". In: *Proceedings of the 11th Annual Conference on Computational Learning Theory*. ACM, pp. 92–100.

Chapelle, O., B. Schölkopf, and A. Zien (2010). *Semi-Supervised Learning*. MIT Press.

Daniušis, P., D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf (2010). "Inferring deterministic causal relations". In: *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pp. 143–150.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Detrano, R. et al. (1989). "International application of a new probability algorithm for the diagnosis of coronary artery disease". In: *The American Journal of Cardiology* 64.5, pp. 304–310.

Feldstein, A. C., G. A. Nichols, D. H. Smith, and et al. (2008). "Weight change and glycemic control after diagnosis of type 2 diabetes". In: *J Gen Intern Med* 23, pp. 1339–1345.

Gieseke, F., A. Airola, T. Pahikkala, and O. Kramer (2014). "Fast and simple gradient-based optimization for semi-supervised support vector machines". In: *Neurocomputing* 123, pp. 23–32.

Heinze-Deml, C. and N. Meinshausen (2017). "Conditional variance penalties and domain shift robustness". In: *arXiv preprint arXiv:1710.11469*.

Hoyer, P. O., D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf (2009). "Nonlinear causal discovery with additive noise models". In: *Advances in Neural Information Processing Systems 21*, pp. 689–696.

Hoyer, P. O., S. Shimizu, A. J. Kerminen, and M. Palviainen (2008). "Estimation of causal effects using linear non-Gaussian causal models with hidden variables". In: *International Journal of Approximate Reasoning* 49.2, pp. 362–378.

Huang, B., K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf (2019). "Causal Discovery from Heterogeneous/Nonstationary Data". In: *arXiv preprint arXiv:1903.01672*.

Janzing, D. and B. Schölkopf (2010). "Causal inference using the algorithmic Markov condition". In: *IEEE Transactions on Information Theory* 56.10, pp. 5168–5194.

— (2018). "Detecting non-causal artifacts in multivariate linear regression models". In: *35th International Conference on Machine Learning (ICML)*. PMLR, pp. 2245–2253.

Janzing, D., J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf (2012). "Information-geometric approach to inferring causal directions". In: *Artificial Intelligence* 182, pp. 1–31.

Joachims, T. (1999). "Transductive inference for text classification using support vector machines". In: *International Conference on Machine Learning*. Vol. 99, pp. 200–209.

Kingma, D. P., S. Mohamed, D. J. Rezende, and M. Welling (2014). "Semi-supervised learning with deep generative models". In: *Advances in Neural Information Processing Systems*, pp. 3581–3589.

Krijthe, J. H. and M. Loog (2017). "Projected estimators for robust semi-supervised classification". In: *Machine Learning* 106.7, pp. 993–1008.

— (2018). "The Pessimistic Limits and Possibilities of Margin-based Losses in Semi-supervised Learning". In: *Advances in Neural Information Processing Systems*, pp. 1790–1799.

Lemeire, J. and E. Dirkx (2006). *Causal models as minimal descriptions of multivariate systems*.

Loog, M. (2015). "Contrastive pessimistic likelihood estimation for semi-supervised classification". In: *IEEE transactions on pattern analysis and machine intelligence* 38.3, pp. 462–475.

Magliacane, S., T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij (2018). "Domain adaptation by using causal inference to predict invariant conditional distributions". In: *Advances in Neural Information Processing Systems*, pp. 10846–10856.

Messerli, F. H. (2012). "Chocolate Consumption, Cognitive Function, and Nobel Laureates". In: *The New England Journal of Medicine* 367.16, pp. 1562–1564.

Mey, A. and M. Loog (2019). "Improvability Through Semi-Supervised Learning: A Survey of Theoretical Results". In: *arXiv preprint arXiv:1908.09574*.

Mohan, K., J. Pearl, and J. Tian (2013). "Graphical models for inference with missing data". In: *Advances in Neural Information Processing systems*, pp. 1277–1285.

Nigam, K., A. McCallum, and T. Mitchell (2006). "Semi-supervised text classification using EM". In: *Semi-Supervised Learning*, pp. 33–56.

Parascandolo, G., N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf (2018). "Learning Independent Causal Mechanisms". In: *International Conference on Machine Learning*, pp. 4033–4041.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press.

Reichenbach, H. (1956). *The direction of time*. Vol. 65. Univ of California Press.

Rojas-Carulla, M., B. Schölkopf, R. Turner, and J. Peters (2018). "Invariant models for causal transfer learning". In: *The Journal of Machine Learning Research* 19.1, pp. 1309–1342.

Schölkopf, B., D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij (2012). "On causal and anticausal learning". In: *Proceedings of the 29th International Conference on Machine Learning*.

Scudder, H. J. (1965). "Probability of Error of Some Adaptive Pattern-Recognition Machines". In: *IEEE Transactions on Information Theory* 11, pp. 363–371.

Smith, J. W., J. Everhart, W. Dickson, W. Knowler, and R. Johannes (1988). "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus". In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, p. 261.

Sugiyama, M. and M. Kawanabe (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Urner, R., S. Shalev-Shwartz, and S. Ben-David (2011). "Access to unlabeled data can speed up prediction time". In: *Proceedings of the 28th International Conference on Machine Learning*, pp. 641–648.

Vapnik, V (1998). *Statistical learning theory*. Wiley, New York.

von Kügelgen, J., A. Mey, and M. Loog (2019). "Semi-generative modelling: covariate-shift adaptation with cause and effect features". In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1361–1369.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning*. 3. MIT press Cambridge, MA.

Zhang, K. and A. Hyvärinen (2009). "On the identifiability of the post-nonlinear causal model". In: *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 647–655.

Zhou, D., O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf (2004). "Learning with local and global consistency". In: *Advances in Neural Information Processing Systems*, pp. 321–328.

Zhu, X. (2005). *Semi-Supervised Learning Literature Survey*. Tech. rep. 1530. University of Wisconsin-Madison.

Zhu, X. and Z. Ghahramani (2002). *Learning from labeled and unlabeled data with label propagation*. Tech. rep. Carnegie Mellon University.

Zhu, X., Z. Ghahramani, and J. D. Lafferty (2003). "Semi-supervised learning using Gaussian fields and harmonic functions". In: *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919.

A Algorithms

Algorithm 3 describes concrete soft and hard labelling versions of the EM approach proposed in Algorithm 1 for the model assumption of a logistic regression for $Y|X_C$ and linear Gaussian distributions for $X_E|X_C, Y = 0$ and $X_E|X_C, Y = 1$ in some more detail. It was used for our experiments in §6.

B Experimental details

B.1 Synthetic datasets

The synthetic datasets used in our experiments were generated as follows. First, we draw $X_C \in \mathbb{R}^{d_C}$ from a mixture of m d_C-dimensional Gaussians. Next, we draw $Y \in \{0, 1\}$ and $X_E \in \mathbb{R}^{d_E}$ according to the SCM

$$Y := \mathbb{I}[\sigma(\mathbf{a}^T X_C + b) > N_Y],$$

$$X_E := \begin{cases} \mathbf{A}_0 X_C + \mathbf{b}_0 + \mathbf{D}_0 N_E & \text{if} \quad Y = 0, \\ \mathbf{A}_1 X_C + \mathbf{b}_1 + \mathbf{D}_1 N_E & \text{if} \quad Y = 1, \end{cases}$$

with $N_Y \sim U[0, 1]$, $N_E \sim \mathcal{N}_{d_E}(0, I)$, $\mathbf{a} \in \mathbb{R}^{d_C}$; $b \in \mathbb{R}$; $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{R}^{d_E \times d_C}$; and $\mathbf{D}_0, \mathbf{D}_1 \in \mathbb{R}^{d_E \times d_E}$ are diagonal matrices of standard deviations. $\sigma(x) = (1 + e^{-x})^{-1}$ denotes the logistic sigmoid function.

This induces the following distributions:

$$Y|X_C \sim \text{Bernoulli}(\mathbf{a}^T X_C + b),$$

$$X_E|(X_C, Y = i) \sim \mathcal{N}(\mathbf{A}_i X_C + \mathbf{b}_i, D_i^2) \quad \text{for} \quad i = 0, 1.$$

For experiments on synthetic data we draw a new dataset according to the above generative process in each run, keeping parameters fixed as follows.

S1: Linear synthetic dataset

- feature dimensions: $d_C = d_E = 1$
- X_C: $m = 3$ components with weights $w = [0.3, 0.4, 0.3]$, means $\mu_C = [-5, 0, 5]$ and standard deviations $\sigma = [0.5, 0.5, 0.5]$
- Y: $a = 0.5, b = 0$
- X_E: $A_0 = A_1 = 1, b_0 = -b_1 = 2, D_0 = D_1 = 0.25$

S2: Nonlinear synthetic dataset

- feature dimensions: $d_C = d_E = 1$
- X_C: $m = 2$ components with weights $w = [0.5, 0.5]$, means $\mu_C = [-3, 3]$ and variances $\sigma^2 = [0.5, 0.5]$

- Y: $a = 0.5, b = 0$
- X_E: $A_0 = -A_1 = 0.5, b_0 = b_1 = 0, D_0 = D_1 = 0.25$

S3: Nonlinear multidimensional synthetic dataset

- feature dimensions: $d_C = d_E = 2$
- X_C: $m = 2$ components with weights $w = [0.5, 0.5]$, means $[-3, -3], [3, 3]$ and covariances $\Sigma = \text{diag}([0.5, 0.5])$
- Y: $a = 0.5, b = 0$
- X_E: $a_0 = -a_1 = 0.5, b_0 = b_1 = 0, D_0 = D_1 = \text{diag}(0.25, 0.25)$

B.2 Medical datasets

For the Pima Indians Diabetes dataset we use the partitioning $X_C = \{$DiabetesPedigreeFunction, Pregnancies, BMI$\}$ and $X_E = \{$Glucose$\}$. DiabetesPedigreeFunction is a measure of the family history of diabetes and BMI stands for body mass index.[14]

For the (Coronary) Heart Disease dataset we used the partitioning $X_C = \{$sex, ca, thal$\}$ and $X_E = \{$chest pain$\}$. Here, "ca" refers to the number of major vessels (0-3) that contained calcium (colored by flouroscopy), and "thal" to thallium scintigraphy results, a nuclear medicine test that images the blood supply to the muscles of the heart.

C Non-linear mechanisms

For simplicity, we have focused on linear mechanisms in our experiments. More complex choices are, of course, possible, and—depending on the setting—necessary.

Figure 5 shows an example of a non-linear additive noise model (Hoyer et al., 2009) for f_0 and f_1. To address this case, one may simply choose a different regression method in our conditional self-learning approach (Algorithm 2 in the main paper); kernel-ridge or Gaussian-process regression (Williams and Rasmussen, 2006) are two obvious choices.

For high-dimensional or structured data such as natural images or text, more complex non-additive noise models can be used for $f_i(X_C, N_E)$, such as the conditional versions of GANs or VAEs. The work of Parascandolo et al. (2018) mentioned at the end of §5 constitutes an example of the former.

---

[14]The relationship between BMI and Diabetes may actually be cyclic (Feldstein et al., 2008), though it is believed that the causal link BMI $\rightarrow$ Diabetes is the stronger one.

Algorithm 3: Soft and hard label EM algorithms for a semi-generative model with logistic regression for $P(Y|X_C)$ and class-dependent linear Gaussian models for $P(X_E|Y, X_C)$.

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; regularisation strength λ for ridge regression; labelling type (hard/soft)

Output: fitted labels $\mathbf{y}^u$; parameters $\boldsymbol{\theta}_Y, \boldsymbol{\theta}_E$

1   $t \leftarrow 0$

    // initialise parameter estimates using only the labelled sample

2   $\boldsymbol{\theta}_Y^{(0)} \leftarrow \texttt{LogisticRegresssion}(\mathbf{X}_C^l, \mathbf{y}^l; \boldsymbol{\theta}_Y)$

3   $\boldsymbol{\theta}_{E,0}^{(0)} \leftarrow ((\mathbf{X}_{C,0}^l)^T \mathbf{X}_{C,0}^l + \lambda I)^{-1}(\mathbf{X}_{C,0}^l)^T \mathbf{X}_{E,0}^l$

4   $\boldsymbol{\theta}_{E,1}^{(0)} \leftarrow ((\mathbf{X}_{C,1}^l)^T \mathbf{X}_{C,1}^l + \lambda I)^{-1}(\mathbf{X}_{C,1}^l)^T \mathbf{X}_{E,1}^l$

5   **while** *not converged* **do**

      // compute soft (class probabilities) and hard labels (E-step)

6    $\mathbf{q}^{(t)} \leftarrow p(\mathbf{y}|\mathbf{X}_C^u, \mathbf{X}_E^u; \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_{E,0}^{(t)}, \boldsymbol{\theta}_{E,1}^{(t)})$

7    $\mathbf{y}^{(t)} \leftarrow \mathbb{I}\{\mathbf{q}^{(t)} > 0.5\}$

      // compute weights (1 for hard labels, class prob. for soft labels)

8   **if** *hard labelling* **then**

9     $\mathbf{w}^{(t)} \leftarrow \mathbf{1}$

10   **else**

11     $\mathbf{w}^{(t)} \leftarrow [\mathbf{1}; \mathbf{y}^{(t)} \odot \mathbf{q}^{(t)} + (1 - \mathbf{y}^{(t)}) \odot (1 - \mathbf{q}^{(t)})]$

12   **end**

13   $\mathbf{W}_1^{(t)} \leftarrow \text{diag}([\mathbf{y}^l; \mathbf{q}^{(t)}])$

14   $\mathbf{W}_0^{(t)} \leftarrow \text{diag}([\mathbf{1} - \mathbf{y}^l; \mathbf{1} - \mathbf{q}^{(t)}])$

      // Update parameter estimates keeping estimated labels fixed (M-step)

15   $\boldsymbol{\theta}_Y^{(t)} \leftarrow \texttt{WeightedLog. Reg.} \ ([\mathbf{X}_C^l; \mathbf{X}_C^u], [\mathbf{y}^l; \mathbf{y}^{(t)}], \mathbf{w}^{(t)}; \boldsymbol{\theta}_Y)$

16   $\boldsymbol{\theta}_{E,1}^{(t)} \leftarrow (\mathbf{X}_C^T \mathbf{W}_1^{(t)} \mathbf{X}_C + \lambda I)^{-1} \mathbf{X}_C^T \mathbf{W}_1^{(t)} \mathbf{X}_E$

17   $\boldsymbol{\theta}_{E,0}^{(t)} \leftarrow (\mathbf{X}_C^T \mathbf{W}_0^{(t)} \mathbf{X}_C + \lambda I)^{-1} \mathbf{X}_C^T \mathbf{W}_0^{(t)} \mathbf{X}_E$

18   $t \leftarrow t + 1$

19   **end**

20   **return** $\mathbf{y}^{(t-1)}, \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}$