Towards a Learning Theory of Cause-Effect Inference

David Lopez-Paz[1,2] DAVID@LOPEZPAZ.ORG
Krikamol Muandet[1] KRIKAMOL@TUEBINGEN.MPG.DE
Bernhard Schölkopf[1] BS@TUEBINGEN.MPG.DE
Ilya Tolstikhin[1] ILYA@TUEBINGEN.MPG.DE
[1]Max-Planck-Institute for Intelligent Systems
[2]University of Cambridge

Abstract

We pose causal inference as the problem of learning to classify probability distributions. In particular, we assume access to a collection $\{(S_i, l_i)\}_{i=1}^n$, where each S_i is a sample drawn from the probability distribution of $X_i \times Y_i$, and l_i is a binary label indicating whether "$X_i \to Y_i$" or "$X_i \leftarrow Y_i$". Given these data, we build a causal inference rule in two steps. First, we featurize each S_i using the kernel mean embedding associated with some characteristic kernel. Second, we train a binary classifier on such embeddings to distinguish between causal directions. We present generalization bounds showing the statistical consistency and learning rates of the proposed approach, and provide a simple implementation that achieves state-of-the-art cause-effect inference. Furthermore, we extend our ideas to infer causal relationships between more than two variables.

1. Introduction

The vast majority of statistical learning algorithms rely on the exploitation of associations between the variables under study. Given the argument that all associations arise from underlying causal structures (Reichenbach, 1956), and that different structures imply different influences between variables, the question of how to infer and use causal knowledge in learning acquires great importance (Pearl, 2000; Schölkopf et al., 2012). Traditionally, the most widely used strategy to infer the causal structure of a system is to perform interventions on some of its variables, while studying the response of some others. However, such interventions are in many situations unethical, expensive, or even impossible to realize. Consequently, we often face the need of causal inference purely from *observational data*. In these scenarios, one suffers, in the absence of strong assumptions, from the indistinguishability between latent confounding ($X \leftarrow Z \to Y$) and direct causation ($X \to Y$ or $X \leftarrow Y$). Nevertheless, disregarding the impossibility of the task, humans continuously learn from experience to accurately infer causality-revealing patterns. Inspired by this successful learning, and in contrast to prior work, this paper addresses causal inference by unveiling such causal patterns directly from data. In particular, we assume access to a set $\{(S_i, l_i)\}_{i=1}^n$, where each S_i is a sample set drawn from the probability distribution of $X_i \times Y_i$, and l_i is a binary label indicating whether "$X_i \to Y_i$" or "$X_i \leftarrow Y_i$". Using these data, we build a causal inference rule in two steps. First, we construct a suitable and nonparametric representation of each sample S_i. Second, we train a nonlinear binary classifier on such features to distinguish between causal directions. Building upon this framework, we derive theoretical guarantees regarding consistency and learning rates, extend inference to the multivariate case, propose approximations to scale learning to big data, and obtain state-of-the-art performance with a simple implementation.

Given the ubiquity of uncertainty in data, which may arise from noisy measurements or the existence of unobserved common causes, we adopt the probabilistic interpretation of causation from Pearl (2000). Under this interpretation, the causal structure underlying a set of random variables $X = (X_1, \ldots, X_d)$, with joint distribution P, is often described in terms of a Directed Acyclic Graph (DAG), denoted by $G = (V, E)$. In this graph, each vertex $V_i \in V$ is associated to the random variable $X_i \in X$, and an edge $E_{ji} \in E$ from V_j to V_i denotes the causal relationship "$X_i \leftarrow X_j$". More specifically, these causal relationships are defined by a *structural equation model*: each $X_i \leftarrow f_i(\mathrm{Pa}(X_i), N_i)$, where f_i is a function, $\mathrm{Pa}(X_i)$ is the parental set of $V_i \in V$, and N_i is some independent noise variable. Then, causal inference is the task of recovering G from $S \sim P^n$.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

1.1. Prior Art

We now briefly review the state-of-the-art on the inference of causal structures G from observational data $S \sim P^n$. For a more thorough exposition, see, e.g., Mooij et al. (2014).

One of the main strategies to recover G is through the exploration of conditional dependencies, together with some other technical assumptions such as the *Markov* and *faithfulness* relationships between P and G (Pearl, 2000). This is the case of the PC algorithm (Spirtes et al., 2000), which allows the recovery of the Markov equivalence class of G without placing any restrictions on the structural equation model specifying the random variables under study.

Causal inference algorithms that exploit conditional dependencies are unsuited for inference in the bivariate case. Consequently, a large body of work has been dedicated to the study of this scenario. First, the linear non-Gaussian causal model (Shimizu et al., 2006; 2011) recovers the true causal direction between two variables whenever their relationship is linear and polluted with additive and non-Gaussian noise. This model was later extended into nonlinear additive noise models (Hoyer et al., 2009; Zhang & Hyvärinen, 2009; Stegle et al., 2010; Kpotufe et al., 2013; Peters et al., 2014), which prefer the causal direction under which the alleged cause is independent from the additive residuals of some nonlinear fit to the alleged effect. Third, the information geometric causal inference framework (Daniusis et al., 2012; Janzing et al., 2014) assumes that the cause random variable is independently generated from some invertible and deterministic mapping to its effect; thus, it is unlikely to find dependencies between the density of the former and the slope of the latter, under the correct causal direction.

As it may be inferred from the previous exposition, there exists a large and heterogeneous array of causal inference algorithms, each of them working under a very specialized set of assumptions, which are sometimes difficult to test in practice. Therefore, there exists the need for a more flexible causal inference rule, capable of learning the relevant *causal footprints*, later used for inference, directly from data. Such a "data driven" approach would allow to deal with complex data-generating processes, and would greatly reduce the need of explicitly crafting identifiability conditions a-priori.

A preliminary step in this direction distilled from the competitions organized by Guyon (2013; 2014), which phrased causal inference as a learning problem. In these competitions, the participants were provided with a large collection of *cause-effect samples* $\{(S_i, l_i)\}_{i=1}^n$, where $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$ is drawn from the probability distribution of $X_i \times Y_i$, and l_i is a binary label indicating whether "$X_i \to Y_i$" or "$Y_i \to X_i$". Given these data, most participants adopted the strategy of i) crafting a vector of features from each S_i, and ii) training a binary classifier on top of the constructed features and paired labels. Although these "data-driven" methods achieved state-of-the-art performance (Guyon, 2013), the laborious task of hand-crafting features renders their theoretical analysis impossible.

In more specific terms, the approach described above is a learning problem with inputs being sample sets S_i, where each S_i contains samples drawn from the probability distribution $P_i(X_i, Y_i)$. In a separate strand of research, there has been several attempts to learn from probability distributions in a principled manner (Jebara et al., 2004; Hein & Bousquet, 2004; Cuturi et al., 2005; Martins et al., 2009; Muandet et al., 2012). Szabó et al. (2014) presented the first theoretical analysis of distributional learning based on kernel mean embedding (Smola et al., 2007), with focus on kernel ridge regression. Similarly, Muandet et al. (2012) studied the problem of classifying distributions, but their approach is constrained to kernel machines, and no guarantees regarding consistency or learning rates are provided.

1.2. Our Contribution

Inspired by Guyon's competitions, we pose causal inference as the problem of classifying probability measures on causally related pairs of random variables. Our contribution to this framework is the use of *kernel mean embeddings* to nonparametrically featurize each cause-effect sample S_i. The benefits of this approach are three-fold. First, this avoids the need of hand-engineering features from the samples S_i. Second, this enables a clean theoretical analysis, including provable learning rates and consistency results. Third, the kernel hyperparameters (that is, the data representation) can be jointly optimized with the classifier using cross-validation. Furthermore, we show how to extend these ideas to infer causal relationships between $d \geq 2$ variables, give theoretically sustained approximations to scale learning to big data, and provide the source code of a simple implementation that outperforms the state-of-the-art.

The rest of this article is organized as follows. Section 2 reviews the concept of kernel mean embeddings, the tool that will facilitate learning from distributions. Section 3 shows the consistency and learning rates of our kernel mean embedding classification approach to cause-effect inference. Section 4 extends the presented ideas to the multivariate causal inference case. Section 5 presents a variety of experiments displaying the state-of-the-art performance of a simple implementation of the proposed framework. For convenience, Table 1 summarizes our notations.

2. Kernel Mean Embeddings of Probability Measures

In order to later classify probability measures P according to their causal properties, we first need to featurize them into

$\mathbb{E}[\xi], \mathbb{V}[\xi]$	Expected value and variance of r.v. ξ
$\mathcal{Z}$	Domain of cause-effect pairs $Z = (X, Y)$
$\mathcal{P}$	Set of cause-effect measures P on $\mathcal{Z}$
$\mathcal{L}$	Set of labels $l_i \in \{-1, 1\}$
$\mathcal{M}$	Mother distribution over $\mathcal{P} \times \mathcal{L}$
$\{(P_i, l_i)\}_{i=1}^n$	Sample from $\mathcal{M}^n$
$S_i = \{Z_{ij}\}_{j=1}^{n_i}$	Sample from $P_i^{n_i}$
P_{S_i}	Empirical distribution of S_i
k	Kernel function from $\mathcal{Z} \times \mathcal{Z}$ to $\mathbb{R}$
$\mathcal{H}_k$	RKHS induced by k
$\mu_k(P)$	Kernel mean embedding of measure $P \in \mathcal{P}$
$\mu_k(P_{s_i})$	Empirical mean embedding of P_{s_i}
$\mu_k(\mathcal{P})$	The set $\{\mu_k(P) : P \in \mathcal{P}\}$
$\mathcal{M}_k$	Measure over $\mu_k(\mathcal{P}) \times \mathcal{L}$ induced by $\mathcal{M}$
$\mathcal{F}_k$	Class of functionals mapping $\mathcal{H}_k$ to $\mathbb{R}$
$R_n(\mathcal{F}_k)$	Rademacher complexity of class $\mathcal{F}_k$
φ, $R_\varphi(f)$	Cost and surrogate φ-risk of sign $\circ f$

Table 1. Table of notations

a suitable representation. To this end, we will rely on the concept of *kernel mean embeddings* (Berlinet & Thomas-Agnan, 2004; Smola et al., 2007).

In particular, let P be the probability distribution of some random variable Z taking values in the separable topological space $(\mathcal{Z}, \tau_z)$. Then, the *kernel mean embedding* of P associated with the continuous, bounded, and positive-definite kernel function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is

$$\mu_k(P) := \int_{\mathcal{Z}} k(z, \cdot)\, \mathrm{d}P(z), \tag{1}$$

which is an element in $\mathcal{H}_k$, the Reproducing Kernel Hilbert Space (RKHS) associated with k (Schölkopf & Smola, 2002). Interestingly, the mapping μ_k is injective if k is a characteristic kernel (Sriperumbudur et al., 2010), that is, $\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = 0 \Leftrightarrow P = Q$. Said differently, if using a characteristic kernel, we do not lose any information when embedding distributions. An example of characteristic kernel is the Gaussian kernel

$$k(z, z') = \exp\left(-\gamma \|z - z'\|_2^2\right), \ \gamma > 0, \tag{2}$$

which will be used throughout this paper.

In many practical situations, it is unrealistic to assume access to the true distribution P, and consequently to the true embedding $\mu_k(P)$. Instead, we often have access to a sample $S = \{z_i\}_{i=1}^n \sim P^n$, which can be used to construct the empirical distribution $P_S := \frac{1}{n} \sum_{z_i \in S} \delta_{(z_i)}$, where $\delta_{(z)}$ is the Dirac distribution centered at z. Using P_S, we can approximate (1) by the *empirical kernel mean embedding*

$$\mu_k(P_S) := \frac{1}{n} \sum_{i=1}^{n} k(z_i, \cdot) \in \mathcal{H}_k. \tag{3}$$

The following result is a slight modification of Theorem 27 from (Song, 2008). It establishes the convergence of the empirical mean embedding $\mu_k(P_S)$ to the embedding of its population counterpart $\mu_k(P)$, in RKHS norm:

Theorem 1. *Assume that $\|f\|_\infty \leq 1$ for all $f \in \mathcal{H}_k$ with $\|f\|_{\mathcal{H}_k} \leq 1$. Then with probability at least $1 - \delta$ we have*

$$\|\mu_k(P) - \mu_k(P_S)\|_{\mathcal{H}_k} \leq 2\sqrt{\frac{\mathbb{E}_{z \sim P}[k(z, z)]}{n}} + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}.$$

Proof. See Section B.1. □

3. A Theory of Causal Inference as Distribution Classification

This section phrases the inference of cause-effect relationships from probability measures as the classification of empirical kernel mean embeddings, and analyzes the learning rates and consistency of such approach. Throughout our exposition, the setup is as follows:

1. We assume the existence of some *Mother distribution* $\mathcal{M}$, defined on $\mathcal{P} \times \mathcal{L}$, where $\mathcal{P}$ is the set of all Borel probability measures on the space $\mathcal{Z}$ of two causally related random variables, and $\mathcal{L} = \{-1, +1\}$.

2. A set $\{(P_i, l_i)\}_{i=1}^n$ is sampled from $\mathcal{M}^n$. Each measure $P_i \in \mathcal{P}$ is the joint distribution of the causally related random variables $Z_i = (X_i, Y_i)$, and the label $l_i \in \mathcal{L}$ indicates whether "$X_i \to Y_i$" or "$X_i \leftarrow Y_i$".

3. In practice, we do not have access to the measures $\{P_i\}_{i=1}^n$. Instead, we observe samples $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i} \sim P_i^{n_i}$, for all $1 \leq i \leq n$. Using S_i, the data $\{(S_i, l_i)\}_{i=1}^n$ is provided to the learner.

4. We featurize every sample S_i into the empirical kernel mean embedding $\mu_k(P_{S_i})$ associated with some kernel function k (Equation 3). If k is a characteristic kernel, we incur no loss of information in this step.

Under this setup, we will use the set $\{(\mu_k(P_{S_i}), l_i)\}_{i=1}^n$ to train a binary classifier from $\mathcal{H}_k$ to $\mathcal{L}$, which will later be used to unveil the causal directions of new, unseen probability measures drawn from $\mathcal{M}$. Note that this framework can be straightforwardly extended to also infer the "confounding ($X \leftarrow Z \to Y$)" and "independent ($X \perp\!\!\!\perp Y$)" cases by adding two extra labels to $\mathcal{L}$.

Given the two nested levels of sampling (being the first one from the Mother distribution $\mathcal{M}$, and the second one from each of the drawn cause-effect measures P_i), it is not trivial to conclude whether this learning procedure is consistent, or how its learning rates depend on the sample sizes n and $\{n_i\}_{i=1}^n$. In the following, we will study the generalization performance of empirical risk minimization over this learning setup. Specifically, we are interested in

upper bounding the excess risk between the empirical risk minimizer and the best classifier from our hypothesis class, with respect to the Mother distribution $\mathcal{M}$.

We divide our analysis in three parts. First, §3.1 reviews the abstract setting of statistical learning theory and surrogate risk minimization. Second, §3.2 adapts these standard results to the case of empirical kernel mean embedding classification. Third, §3.3 considers theoretically sustained approximations to deal with big data.

3.1. Margin-based Risk Bounds in Learning Theory

Let $\mathbb{P}$ be some unknown probability measure defined on $\mathcal{Z} \times \mathcal{L}$, where $\mathcal{Z}$ is referred to as the *input space*, and $\mathcal{L} = \{-1, 1\}$ is referred to as the *output space*[1]. One of the main goals of statistical learning theory (Vapnik, 1998) is to find a classifier $h\colon \mathcal{Z} \to \mathcal{L}$ that minimizes the *expected risk*

$$R(h) = \mathop{\mathbb{E}}_{(z,l)\sim\mathbb{P}}\big[\ell\big(h(z), l\big)\big]$$

for a suitable *loss function* $\ell\colon \mathcal{L}\times\mathcal{L} \to \mathbb{R}^+$, which penalizes departures between predictions $h(z)$ and true labels l. For classification, one common choice of loss function is the *0-1 loss* $\ell_{01}(l, l') = |l - l'|$, for which the expected risk measures the probability of misclassification. Since $\mathbb{P}$ is unknown in natural situations, one usually resorts to the minimization of the *empirical risk* $\frac{1}{n}\sum_{i=1}^n \ell\big(h(z_i), l_i\big)$ over some fixed hypothesis class $\mathcal{H}$, for *the training set* $\{(z_i, l_i)\}_{i=1}^n \sim \mathbb{P}^n$. It is well known that this procedure is consistent under mild assumptions (Boucheron et al., 2005).

Unfortunately, the 0-1 loss function is not convex, which leads to empirical risk minimization being generally intractable. Instead, we will focus on the minimization of surrogate risk functions (Bartlett et al., 2006). In particular, we will consider the set of classifiers of the form $\mathcal{H} = \{\operatorname{sign}\circ f\colon f \in \mathcal{F}\}$ where $\mathcal{F}$ is some fixed set of real-valued functions $f\colon \mathcal{Z} \to \mathbb{R}$. Introduce a nonnegative *cost function* $\varphi\colon \mathbb{R} \to \mathbb{R}^+$ which is surrogate to the 0-1 loss, that is, $\varphi(\epsilon) \geq \mathbb{1}_{\epsilon>0}$. For any $f \in \mathcal{F}$ we define its expected and empirical φ-risks respectively as

$$R_\varphi(f) = \mathop{\mathbb{E}}_{(z,l)\sim\mathbb{P}}\big[\varphi\big(-f(z)l\big)\big], \tag{4}$$

$$\hat{R}_\varphi(f) = \frac{1}{n}\sum_{i=1}^n \varphi\big(-f(z_i)l_i\big). \tag{5}$$

Many natural choices of φ lead to tractable empirical risk minimization. Common examples of cost functions include the *hinge loss* $\varphi(\epsilon) = \max(0, 1+\epsilon)$ used in SVM, the *exponential loss* $\varphi(\epsilon) = \exp(\epsilon)$ used in Adaboost, and the *logistic loss* $\varphi(\epsilon) = \log_2\big(1 + e^\epsilon\big)$ used in logistic regression.

[1] Refer to Section A for considerations on measurability.

The misclassification error of $\operatorname{sign}\circ f$ is always upper bounded by $R_\varphi(f)$. The relationship between functions minimizing $R_\varphi(f)$ and functions minimizing $R(\operatorname{sign}\circ f)$ has been intensively studied in the literature (Steinwart & Christmann, 2008, Chapter 3). Given the high uncertainty associated with causal inferences, we argue that one is interested in predicting soft probabilities rather than hard labels, a fact that makes the study of margin-based classifiers well suited for our problem.

We now focus on the estimation of $f^* \in \mathcal{F}$, the function minimizing (4). However, since the distribution $\mathbb{P}$ is unknown, we can only hope to estimate $\hat{f}_n \in \mathcal{F}$, the function minimizing (5). Therefore, we are interested in high-probability upper bounds on the *excess φ-risk*

$$\mathcal{E}_\mathcal{F}(\hat{f}_n) = R_\varphi(\hat{f}_n) - R_\varphi(f^*), \tag{6}$$

w.r.t. the random training sample $\{(z_i, l_i)\}_{i=1}^n \sim \mathbb{P}^n$. The excess risk (6) can be upper bounded in the following way:

$$\begin{aligned}\mathcal{E}_\mathcal{F}(\hat{f}_n) &\leq R_\varphi(\hat{f}_n) - \hat{R}_\varphi(\hat{f}_n) + \hat{R}_\varphi(f^*) - R_\varphi(f^*)\\ &\leq 2\sup_{f\in\mathcal{F}} |R_\varphi(f) - \hat{R}_\varphi(f)|.\end{aligned} \tag{7}$$

While this upper bound leads to tight results for worst case analysis, it is well known (Bartlett et al., 2005; Boucheron et al., 2005; Koltchinskii, 2011) that tighter bounds can be achieved under additional assumptions on $\mathbb{P}$. However, we leave these analyses for future research.

The following result — in spirit of Koltchinskii & Panchenko (1999); Bartlett & Mendelson (2002) — can be found in Boucheron et al. (2005, Theorem 4.1).

Theorem 2. *Consider a class $\mathcal{F}$ of functions mapping $\mathcal{Z}$ to $\mathbb{R}$. Let $\varphi\colon \mathbb{R} \to \mathbb{R}^+$ be a L_φ-Lipschitz function such that $\varphi(\epsilon) \geq \mathbb{1}_{\epsilon>0}$. Let B be a uniform upper bound on $\varphi\big(-f(\epsilon)l\big)$. Let $\{(z_i, l_i)\}_{i=1}^n \sim \mathbb{P}$ and $\{\sigma_i\}_{i=1}^n$ be i.i.d. Rademacher random signs. Then, with prob. at least $1-\delta$,*

$$\begin{aligned}&\sup_{f\in\mathcal{F}} |R_\varphi(f) - \hat{R}_\varphi(f)|\\ &\quad\leq 2L_\varphi\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i f(z_i)\right|\right] + B\sqrt{\frac{\log(1/\delta)}{2n}},\end{aligned}$$

where the expectation is taken w.r.t. $\{\sigma_i, z_i\}_{i=1}^n$.

The expectation in the bound of Thm. 2 is known as *the Rademacher complexity* of $\mathcal{F}$, will be denoted by $R_n(\mathcal{F})$, and has a typical order of $O(n^{-1/2})$ (Koltchinskii, 2011).

3.2. From Classic to Distributional Learning Theory

Note that we can not directly apply the empirical risk minimization bounds discussed in the previous section to our learning setup. This is because instead of learning a classifier on the i.i.d. sample $\{\mu_k(P_i), l_i\}_{i=1}^n$, we have to learn

over the set $\{\mu_k(P_{S_i}), l_i\}_{i=1}^n$, where $S_i \sim P_i^{n_i}$. Said differently, our input feature vectors $\mu_k(P_{S_i})$ are "noisy": they exhibit an additional source of variation as any two different random samples $S_i, S_i' \sim P_i^{n_i}$ do. In the following, we study how to incorporate these nested sampling effects into an argument similar to Theorem 2.

We will now frame our problem within the abstract learning setting considered in the previous section. Recall that our learning setup initially considers some Mother distribution $\mathcal{M}$ over $\mathcal{P} \times \mathcal{L}$. Let $\mu_k(\mathcal{P}) = \{\mu_k(P) : P \in \mathcal{P}\} \subseteq \mathcal{H}_k$, $\mathcal{L} = \{-1, +1\}$, and $\mathcal{M}_k$ be a measure (guaranteed to exist by Lemma 2, Section A.1) on $\mu_k(\mathcal{P}) \times \mathcal{L}$ induced by $\mathcal{M}$. Specifically, we will consider $\mu_k(\mathcal{P}) \subseteq \mathcal{H}_k$ and $\mathcal{L}$ to be the input and output spaces of our learning problem, respectively. Let $\big\{\big(\mu_k(P_i), l_i\big)\big\}_{i=1}^n \sim \mathcal{M}_k^n$ be our training set. We will now work with the set of classifiers $\{\text{sign} \circ f \colon f \in \mathcal{F}_k\}$ for some fixed class $\mathcal{F}_k$ of functionals mapping from the RKHS $\mathcal{H}_k$ to $\mathbb{R}$.

As pointed out in the description of our learning setup, we do not have access to the distributions $\{P_i\}_{i=1}^n$ but to samples $S_i \sim P_i^{n_i}$, for all $1 \leq i \leq n$. Because of this reason, we define the *sample-based empirical φ-risk*

$$\tilde{R}_\varphi(f) = \frac{1}{n} \sum_{i=1}^n \varphi\big(-l_i f\big(\mu_k(P_{S_i})\big)\big),$$

which is the approximation to the empirical φ-risk $\hat{R}_\varphi(f)$ that results from substituting the embeddings $\mu_k(P_i)$ with their empirical counterparts $\mu_k(P_{S_i})$.

Our goal is again to find the function $f^* \in \mathcal{F}_k$ minimizing expected φ-risk $R_\varphi(f)$. Since $\mathcal{M}_k$ is unknown to us, and we have no access to the embeddings $\{\mu_k(P_i)\}_{i=1}^n$, we will instead use the minimizer of $\tilde{R}_\varphi(f)$ in $\mathcal{F}_k$:

$$\tilde{f}_n \in \arg\min_{f \in \mathcal{F}_k} \tilde{R}_\varphi(f). \tag{8}$$

To sum up, the excess risk (6) can now be reformulated as

$$R_\varphi(\tilde{f}_n) - R_\varphi(f^*). \tag{9}$$

Note that the estimation of f^* drinks from two nested sources of error, which are i) having only n training samples from the distribution $\mathcal{M}_k$, and ii) having only n_i samples from each measure P_i. Using a similar technique to (7), we can upper bound the excess risk as

$$R_\varphi(\tilde{f}_n) - R_\varphi(f^*) \leq 2 \sup_{f \in \mathcal{F}_k} |R_\varphi(f) - \hat{R}_\varphi(f)| \tag{10}$$

$$+ 2 \sup_{f \in \mathcal{F}_k} |\hat{R}_\varphi(f) - \tilde{R}_\varphi(f)|. \tag{11}$$

The term (10) is upper bounded by Theorem 2. On the other hand, to deal with (11), we will need to upper bound the deviations $\big|f\big(\mu_k(P_i)\big) - f\big(\mu_k(P_{S_i})\big)\big|$ in terms of the distances $\|\mu_k(P_i) - \mu_k(P_{S_i})\|_{\mathcal{H}_k}$, which are in turn upper bounded using Theorem 1. To this end, we will have to assume that the class $\mathcal{F}_k$ consists of functionals with uniformly bounded Lipschitz constants, such as the set of linear functionals with uniformly bounded operator norm.

We now present the main result of this section, which provides a high-probability bound on the excess risk (9). Importantly, this excess risk will translate into the expected causal inference accuracy of our distribution classifier.

Theorem 3. *Consider the RKHS $\mathcal{H}_k$ associated with some bounded, continuous kernel function k, such that $\sup_{z \in \mathcal{Z}} k(z,z) \leq 1$. Consider a class $\mathcal{F}_k$ of functionals mapping $\mathcal{H}_k$ to $\mathbb{R}$ with Lipschitz constants uniformly bounded by $L_{\mathcal{F}}$. Let $\varphi \colon \mathbb{R} \to \mathbb{R}^+$ be a L_φ-Lipschitz function such that $\phi(z) \geq \mathbb{1}_{z>0}$. Let $\varphi\big(-f(h)l\big) \leq B$ for every $f \in \mathcal{F}_k$, $h \in \mathcal{H}_k$, and $l \in \mathcal{L}$. Then, with probability not less than $1 - \delta$ (over all sources of randomness)*

$$R_\varphi(\tilde{f}_n) - R_\varphi(f^*) \leq 4L_\varphi R_n(\mathcal{F}_k) + 2B\sqrt{\frac{\log(2/\delta)}{2n}}$$

$$+ \frac{4L_\varphi L_{\mathcal{F}}}{n} \sum_{i=1}^n \left(\sqrt{\frac{\mathbb{E}_{z \sim P_i}[k(z,z)]}{n_i}} + \sqrt{\frac{\log(2n/\delta)}{2n_i}} \right).$$

Proof. See Section B.2. □

As mentioned in Section 3.1, the typical order of $R_n(\mathcal{F}_k)$ is $O(n^{-1/2})$. For a particular examples of classes of functionals with small Rademacher complexity we refer to Maurer (2006). In such cases, the upper bound in Theorem 3 converges to zero (meaning that our procedure is consistent) as both n and n_i tend to infinity, in such a way that[2] $\log n / n_i = o(1)$. The rate of convergence w.r.t. n can be improved up to $O(n^{-1})$ if placing additional assumptions on $\mathcal{M}$ (Bartlett et al., 2005). On the contrary, the rate w.r.t. n_i cannot be improved in general. Namely, the convergence rate $O(n^{-1/2})$ presented in the upper bound of Theorem 1 is tight, as shown in the following novel result.

Theorem 4. *Under the assumptions of Theorem 1 denote*

$$\sigma_{\mathcal{H}_k}^2 = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{V}_{z \sim P}[f(z)].$$

Then there exist universal constants c, C such that for every integer $n \geq 1/\sigma_{\mathcal{H}_k}^2$, and with probability at least c

$$\|\mu_k(P) - \mu_k(P_S)\|_{\mathcal{H}_k} \geq C \frac{\sigma_{\mathcal{H}_k}}{\sqrt{n}}.$$

Proof. See Section B.3. □

Finally, it is instructive to relate the notion of "identifiability" often considered in the causal inference community (Pearl,

[2] We conjecture that this constraint is an artifact from our proof.

2000) to the properties of the Mother distribution. Saying that the model is *identifiable* means that the label l of $P \in \mathcal{P}$ is assigned deterministically by $\mathcal{M}$. In this case, learning rates can become as fast as $O(n^{-1})$. On the other hand, as $\mathcal{M}(l|P)$ becomes nondeterministic, the problem becomes unidentifiable and learning rates slow down (for example, in the extreme case of cause-effect pairs related by linear functions polluted with additive Gaussian noise, $\mathcal{M}(l = +1|P) = \mathcal{M}(l = -1|P)$ almost surely). The investigation of these phenomena is left for future research.

3.3. Low Dimensional Embeddings for Large Data

For some kernel functions, the embeddings $\mu_k(P_S) \in \mathcal{H}_k$ are infinite dimensional. Because of this reason, one must resort to the use of dual optimization problems, and in particular, kernel matrices. The construction of these matrices requires at least $O(n^2)$ computational and memory requirements, prohibitive for large n. In this section, we show that the infinite-dimensional embeddings $\mu_k(P_S) \in \mathcal{H}_k$ can be approximated with easy to compute, low-dimensional representations (Rahimi & Recht, 2007; 2008). This will allow us to replace the infinite-dimensional minimization problem (8) with a low-dimensional one.

Assume that $\mathcal{Z} = \mathbb{R}^d$, and that the kernel function k is real-valued, and shift-invariant. Then, we can exploit Bochner's theorem (Rudin, 1962) to show that, for any $z, z' \in \mathcal{Z}$:

$$k(z, z') = 2C_k \mathop{\mathbb{E}}_{w,b} [\cos(\langle w, z\rangle + b) \cos(\langle w, z'\rangle + b)] , \quad (12)$$

where $w \sim \frac{1}{C_k} p_k$, $b \sim \mathcal{U}[0, 2\pi]$, $p_k \colon \mathcal{Z} \to \mathbb{R}$ is the positive and integrable Fourier transform of k, and $C_k = \int_{\mathcal{Z}} p_k(w) dw$. For example, the squared-exponential kernel (2) is shift-invariant, and its evaluations can be approximated by (12), if setting $p_k(w) = \mathcal{N}(w|0, 2\gamma I)$, and $C_k = 1$.

We now show that for any probability measure Q on $\mathcal{Z}$ and $z \in \mathcal{Z}$, the function $k(z, \cdot) \in \mathcal{H}_k \subseteq L_2(Q)$ can be approximated by a linear combination of randomly chosen elements from the Hilbert space $L_2(Q)$. Namely, consider the functions parametrised by $w, z \in \mathcal{Z}$ and $b \in [0, 2\pi]$:

$$g^z_{w,b}(\cdot) = 2C_k \cos(\langle w, z\rangle + b) \cos(\langle w, \cdot\rangle + b), \quad (13)$$

which belong to $L_2(Q)$, since they are bounded. If we sample $\{(w_j, b_j)\}_{j=1}^m$ i.i.d., as discussed above, the average

$$\hat{g}^z_m(\cdot) = \frac{1}{m} \sum_{i=1}^m g^z_{w_i,b_i}(\cdot)$$

can be viewed as an $L_2(Q)$-valued random variable. Moreover, (12) shows that $\mathbb{E}_{w,b}[\hat{g}^z_m(\cdot)] = k(z, \cdot)$. This enables us to invoke concentration inequalities for Hilbert spaces (Ledoux & Talagrand, 1991), to show the following result, which is in spirit to Rahimi & Recht (2008, Lemma 1).

Lemma 1. *Let $\mathcal{Z} = \mathbb{R}^d$. For any shift-invariant kernel k, s.t. $\sup_{z\in\mathcal{Z}} k(z, z) \leq 1$, any fixed $S = \{z_i\}_{i=1}^n \subset \mathcal{Z}$, any probability distribution Q on $\mathcal{Z}$, and any $\delta > 0$, we have*

$$\left\| \mu_k(P_S) - \frac{1}{n} \sum_{i=1}^n \hat{g}^{z_i}_m(\cdot) \right\|_{L_2(Q)} \leq \frac{2C_k}{\sqrt{m}} \left(1 + \sqrt{2 \log(n/\delta)}\right)$$

with probability larger than $1 - \delta$ over $\{(w_i, b_i)\}_{i=1}^m$.

Proof. See Section B.4. □

Once sampled, the parameters $\{(w_i, b_i)\}_{i=1}^m$ allow us to approximate the empirical kernel mean embeddings $\{\mu_k(P_{S_i})\}_{i=1}^n$ using elements from $\mathrm{span}(\{\cos(\langle w_i, \cdot\rangle + b_i)\}_{i=1}^m)$, which is a finite-dimensional subspace of $L_2(Q)$. Therefore, we propose to use $\{(\mu_{k,m}(P_{S_i}), l_i)\}_{i=1}^n$ as the training sample for our final empirical risk minimization problem, where

$$\mu_{k,m}(P_S) = \frac{2C_k}{|S|} \sum_{z\in S} \big(\cos(\langle w_j, z\rangle + b_j)\big)_{j=1}^m \in \mathbb{R}^m. \quad (14)$$

These feature vectors can be computed in $O(m)$ time and stored in $O(1)$ memory; importantly, they can be used off-the-shelf in conjunction with any learning algorithm.

For the precise excess risk bounds that take into account the use of these low-dimensional approximations, please refer to Theorem 6 from Section B.5.

4. Extensions to Multivariate Causal Inference

It is possible to extend our framework to infer causal relatonships between $d \geq 2$ variables $X = (X_1, \ldots, X_d)$. To this end, and as introduced in Section 1, assume the existence of a causal directed acyclic graph G which underlies the dependencies present in the probability distribution $P(X)$. Therefore, our task is to recover G from $S \sim P^n$.

Naïvely, one could extend the framework presented in Section 3 from the binary classification of 2-dimensional distributions to the multiclass classification of d-dimensional distributions. However, the number of possible DAGs (and therefore, the number of labels in our multiclass classification problem) grows super-exponentially in d.

An alternative approach is to consider the probabilities of the three labels "$X_i \to X_j$", "$X_i \leftarrow X_j$", and "$X_i \perp\!\!\!\perp X_j$" for each pair of variables $\{X_i, X_j\} \subseteq X$, when embedded along with every possible *context* $X_k \subseteq X \setminus \{X_i, X_j\}$. The intuition here is the same as in the PC algorithm of Spirtes et al. (2000): in order to decide the (absence of a) causal relationship between X_i and X_j, one must analyze the confounding effects of every $X_k \subseteq X \setminus \{X_i, X_j\}$.

5. Numerical Simulations

We conduct an array of experiments to test the effectiveness of a simple implementation of the presented causal learning framework. Given the use of random embeddings (14) in our classifier, we term our method the *Randomized Causation Coefficient* (RCC). Throughout our simulations, we featurize each sample $S = \{(x_i, y_i)\}_{i=1}^n$ as

$$\nu(S) = (\mu_{k,m}(P_{S_x}), \mu_{k,m}(P_{S_y}), \mu_{k,m}(P_{S_{xy}})), \quad (15)$$

where the three elements forming (15) stand for the low-dimensional representations (14) of the empirical kernel mean embeddings of $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$, and $\{(x_i, y_i)\}_{i=1}^n$, respectively. The representation (15) is motivated by the typical conjecture in causal inference about the existence of asymmetries between the marginal and conditional distributions of causally-related pairs of random variables (Schölkopf et al., 2012). Each of these three embeddings has random features sampled to approximate the sum of three Gaussian kernels (2) with hyper-parameters 0.1γ, γ, and 10γ, where γ is found using the median heuristic. In practice, we set $m = 1000$, and observe no significant improvements when using larger amounts of random features. To classify the embeddings (15) in each of the experiments, we use the random forest[3] implementation from Python's `sklearn-0.16-git`. The number of trees is chosen from $\{100, 250, 500, 1000, 5000\}$ via cross-validation.

Our experiments can be replicated using the source code at

https://github.com/lopezpaz/causation_learning_theory.

5.1. Classification of Tübingen Cause-Effect Pairs

The *Tübingen cause-effect pairs* is a collection of heterogeneous, hand-collected, real-world cause-effect samples (Zscheischler, 2014). Given the small size of this dataset, we resort to the synthesis of an artificial Mother distribution to sample our training data from. To this end, assume that sampling a synthetic cause-effect sample set $\hat{S}_i := \{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n \sim \mathcal{P}_\theta$ equals the following simple generative process:

1. A *cause* vector $(\hat{x}_{ij})_{j=1}^n$ is sampled from a mixture of Gaussians with c components. The mixture weights are sampled from $\mathcal{U}(0, 1)$, and normalized to sum to one. The mixture means and standard deviations are sampled from $\mathcal{N}(0, \sigma_1)$, and $\mathcal{N}(0, \sigma_2)$, respectively, accepting only positive standard deviations. The cause vector is standardized to zero mean and unit variance.
2. A *noise* vector $(\hat{\epsilon}_{ij})_{j=1}^n$ is sampled from a centered Gaussian, with variance sampled from $\mathcal{U}(0, \sigma_3)$.
3. A *mapping mechanism* $\hat{f}_i$ is conceived as a spline fitted using an uniform grid of d_f elements from $\min((\hat{x}_{ij})_{j=1}^n)$ to $\max((\hat{x}_{ij})_{j=1}^n)$ as inputs, and d_f normally distributed outputs.
4. An *effect* vector is built as $(\hat{y}_{ij} := \hat{f}_i(\hat{x}_{ij}) + \hat{\epsilon}_{ij})_{j=1}^n$, and standardized to zero mean and unit variance.
5. Return the cause-effect sample $\hat{S}_i := \{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n$.

To choose a $\theta = (c, \sigma_1, \sigma_2, \sigma_3, d_f)$ that best resembles the unlabeled test data, we minimize the distance between the embeddings of N synthetic pairs and the Tuebingen samples

$$\arg\min_\theta \sum_i \min_{1 \le j \le N} \|\nu(S_i) - \nu(\hat{S}_j)\|_2^2, \quad (16)$$

over $c, d_f \in \{1, \ldots, 10\}$, and σ_1, σ_2, and $\sigma_3 \in \{0, 0.5, 1, \ldots, 5\}$, where the $\hat{S}_j \sim \mathcal{P}_\theta$, the S_i are the Tübingen cause-effect pairs, and ν is as in (15). This strategy can be thought of as transductive learning, since we assume to know the test inputs prior to the training of our inference rule. We set $n = 1000$, and $N = 10,000$.

Using the generative process outlined above, we construct the synthetic training data

$$\{\{\nu(\{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n), +1)\}_{i=1}^N,$$
$$\{\nu(\{(\hat{y}_{ij}, \hat{x}_{ij})\}_{j=1}^n), -1)\}_{i=1}^N\},$$

where $\{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n \sim \mathcal{P}_\theta$, and train our classifier on it.

Figure 1 plots the classification accuracy of RCC, IGCI (Daniusis et al., 2012), and ANM (Mooij et al., 2014) versus the fraction of decissions that the algorithms are forced to make out of the 82 scalar Tüebingen cause-effect pairs. To compare these results to other lower-performing methods,

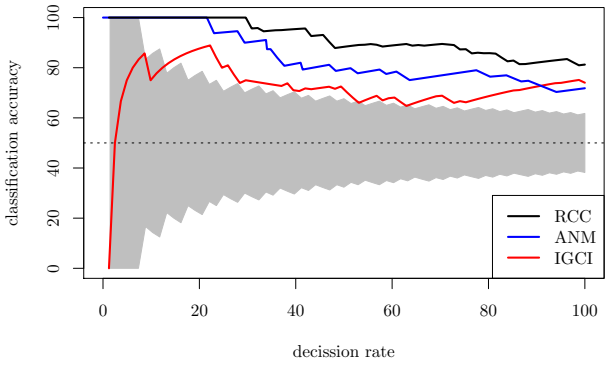

Figure 1. Accuracy of RCC, IGCI and ANM on the Tübingen cause-effect pairs, as a function of decision rate. The grey area depicts accuracies not statistically significant.

[3] Although random forests do not comply with Lipschitzness assumptions from Section 3, they showed the best empirical results. Compliant alternatives such as SVMs exhibited a typical drop in classification accuracy of 5%.

refer to Janzing et al. (2012). RCC surpasses the state-of-the-art with a classification accuracy of 81.61% when inferring the causal directions on all pairs. The confidence of RCC is computed using the classifier's output class probabilities. SVMs obtain a test accuracy of 77.2% in this same task.

5.2. Inferring the Arrow of Time

We test the effectiveness of our method to infer the arrow of time from causal time series. More specifically, we assume access to a set of time series $\{x_{ij}\}_{j=1}^{n_i}$, and our task is to infer, for each series, whether $X_i \to X_{i+1}$ or $X_i \leftarrow X_{i+1}$.

We compare our framework to the state-of-the-art of Peters et al. (2009), using the same electroencephalography signals (Blankertz, 2005) as in their original experiment. On the one hand, Peters et al. (2009) construct two Auto-Regressive Moving-Average (ARMA) models for each causal time series and time direction, and prefers the solution under which the model residuals are independent from the inferred cause. To this end, the method uses two parameters for which no estimation procedure is provided. On the other hand, our approach makes no assumptions whatsoever about the parametric model underlying the series, at the expense of requiring a disjoint set of $N = 10,000$ causal time series for training. Our method matches the best performance of Peters et al. (2009), with an accuracy of 82.66%.

5.3. ChaLearn's Challenge Data

The cause-effect challenges organized by Guyon (2014) provided $N = 16,199$ training causal samples S_i, each drawn from the distribution of $X_i \times Y_i$, and labeled either "$X_i \to Y_i$", "$X_i \leftarrow Y_i$", "$X_i \leftarrow Z_i \to Y_i$", or "$X_i \perp\!\!\!\perp Y_i$". The task of the competition was to develop a *causation coefficient* which would predict large positive values to causal samples following "$X_i \to Y_i$", large negative values to samples following "$X_i \leftarrow Y_i$", and zero otherwise. Using these data, our obtained a test *bidirectional area under the curve score* (Guyon, 2014) of 0.74 in one minute and a half, ranking third in the overall leaderboard. The winner of the competition obtained a score of 0.82 in thirty minutes, but resorted to several dozens of hand-crafted features.

Partitioning these same data in different ways, we learned two related but different binary classification tasks. First, we trained our classifier to *detect latent confounding*, and obtained a test classification accuracy of 80% on the task of distinguishing "$X \to Y$ or $X \leftarrow X$" from "$X \leftarrow Z \to Y$". Second, we trained our classifier to *measure dependence*, and obtained a test classification accuracy of 88% on the task of distinguishing between "$X \perp\!\!\!\perp Y$" and "else". We consider these results to be a promising direction to learn flexible hypothesis tests and dependence measures *directly from data*.

5.4. Reconstruction of Causal DAGs

We apply the strategy described in Section 4 to reconstruct the causal DAGs of two multivariate datasets: *autoMPG* and *abalone* (Lichman, 2013). Once again, we resort to synthetic training data, generated in a similar procedure to the one used in Section 5.1. Refer to Section C for details.

Regarding *autoMPG*, in Figure 2, 1) the release date of the vehicle (AGE) causes the miles per gallon consumption (MPG), acceleration capabilities (ACC) and horse-power (HP), 2) the weight of the vehicle (WEI) causes the horse-power and MPG, and that 3) other characteristics such as the engine displacement (DIS) and number of cylinders (CYL) cause the MPG. For *abalone*, in Figure 3, 1) the age of the snail causes all the other variables, 2) the overall weight of the snail (WEI) is caused by the partial weights of its meat (WEA), viscera (WEB), and shell (WEC), and 3) the height of the snail (HEI) is responsible for other phisicaly attributes such as its diameter (DIA) and length (LEN).

The target variable for each dataset is shaded in gray. Interstingly, our inference reveals that the *autoMPG* dataset is a *causal* prediction task (the features *cause* the target), and that the *abalone* dataset is an *anticausal* prediction task (the target *causes* the features). This distinction has implications when learning from these data (Schölkopf et al., 2012).

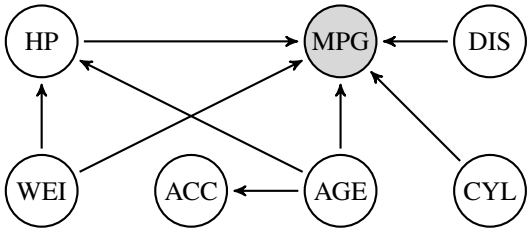

Figure 2. Causal DAG recovered from data *autoMPG*.

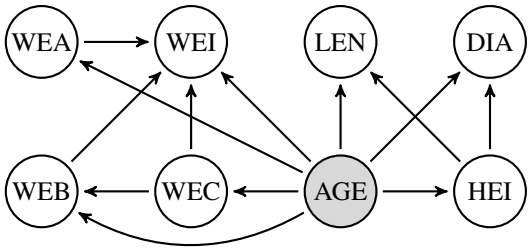

Figure 3. Causal DAG recovered from data *abalone*.

6. Future Work

Three research directions are in progress. First, to improve learning rates by using common assumptions from causal inference. Second, to further investigate methods to reconstruct multivariate DAGs. Third, to develop mechanisms to interpret the causal footprints learned by our classifiers.

References

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 2002.

Bartlett, P. L. and Mendelson, S. Empirical minimization. *Probability Theory and Related Fields*, 135(4), 2006.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4), 2005.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101, 2006.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

Blankertz, B. BCI Competition III data, experiment 4a, subject 3, 1000Hz, 2005. URL http://bbci.de/competition/iii/download/.

Boucheron, Stéphane, Lugosi, Gábor, and Bousquet, Olivier. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 2005.

Cuturi, M., Fukumizu, K., and Vert, J. P. Semigroup kernels on measures. *JMLR*, 2005.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. Inferring deterministic causal relations. *UAI*, 2012.

Guyon, I. Cause-effect pairs kaggle competition, 2013. URL https://www.kaggle.com/c/cause-effect-pairs/.

Guyon, I. Chalearn fast causation coefficient challenge, 2014. URL https://www.codalab.org/competitions/1381.

Hein, M. and Bousquet, O. Hilbertian metrics and positive definite kernels on probability measures. *AISTATS*, 2004.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J. R., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *NIPS*, 2009.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 2012.

Janzing, D., Steudel, B., Shajarisales, N., and Schölkopf, B. Justifying information-geometric causal inference. *arXiv*, 2014.

Jebara, T., Kondor, R., and Howard, A. Probability product kernels. *JMLR*, 2004.

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Ecole d'été de probabilités de Saint-Flour. Springer, 2011.

Koltchinskii, V. and Panchenko, D. Rademacher processes and bounding the risk of function learning. In E. Gine, D. and J.Wellner (eds.), *High Dimensional Probability, II*, pp. 443–457. Birkhauser, 1999.

Kpotufe, S., Sgouritsa, E., Janzing, D., and Schölkopf, B. Consistency of causal inference under the additive noise model. *ICML*, 2013.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. Nonextensive information theoretic kernels on measures. *JMLR*, 2009.

Maurer, A. The rademacher complexity of linear transformation classes. *COLT*, 2006.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. Learning from distributions via support measure machines. *NIPS*, 2012.

Pearl, J. *Causality: models, reasoning and inference*. Cambridge Univ Press, 2000.

Peters, J., Janzing, D., Gretton, A., and Schölkopf, B. Detecting the direction of causal time series. *ICML*, 2009.

Peters, J., M., Joris M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *JMLR*, 2014.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *NIPS*, 2007.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *NIPS*, 2008.

Reed, M. and Simon, B. *Methods of modern mathematical physics. vol. 1. Functional analysis*, volume 1. Academic press New York, 1972.

Reichenbach, H. *The Direction of Time*. Dover, 1956.

Rudin, W. *Fourier Analysis on Groups*. Wiley, 1962.

Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, 2002.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. *ICML*, 2012.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *ICML*, 2012.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *JMLR*, 2006.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *JMLR*, 2011.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *ALT*. Springer-Verlag, 2007.

Song, L. *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, The University of Sydney, 2008.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT Press, 2000.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. Hilbert space embeddings and metrics on probability measures. *JMLR*, 2010.

Stegle, O., Janzing, D., Zhang, K., Mooij, J. M., and Schölkopf, B. Probabilistic latent variable models for distinguishing between cause and effect. *NIPS*, 2010.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer, 2008.

Szabó, Z., Sriperumbudur, B. Póczos, B., and Gretton, A. Learning theory for distribution regression. *arXiv*, 2014.

Vapnik, V. N. *Statistical Learning Theory*. John Wiley & Sons, 1998.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. *UAI*, 2009.

Zscheischler, J. Benchmark data set for causal discovery algorithms, v0.8, 2014. URL `http://webdav.tuebingen.mpg.de/cause-effect/`.