First Principle Methods for Causal Discovery

This report is submitted as part of the requirement
for the M.Sc. Degree in Data Science at the
École polytechnique fédérale de Lausanne by

Ignacio Sukarno Alemán

Supervisors :

Prof Martin Jaggi,
Sai Praneeth,
Reddy Karimireddy

Lausanne, EPFL, July, 2020

If a man will begin with certainties,
he shall end in doubts;
but if he will be content to begin with doubts,
he shall end in certainties.
— Bacon

To my family...

Acknowledgements

TODO

Lausanne, July 28, 2020 Ignacio S. Aleman

Abstract

Understanding the relationships among objects is in some sense half of the study of science; the other half is to define and discover these objects. The question of causality is more straightforward if we are able to intervene on a system – this is indeed how we learn as we grow up. What happens when we are unable to perform any interventions and only make observations? When are we able to make any sort of causal inference? If so, how? We will explore and answer some of the questions for the most basic setting, that is, the bivariate setting with two random variables X and Y. When can we deduce that X causes Y? We will review classical methods such as the additive noise model (ANM), and more recent ones such as the causal generative neural networks (CGNN). We will also review some theory that tells us when causal is possible in the bivariate setting under some assumptions about the causal model (how Y depends on X). We will conclude by showing a new type of method – inspired by the ANM type scoring methodology. Instead of testing for independence between residuals and inputs we will test the consistency of the residual.

Contents

1 Introduction

1.1 Problem and Motivation

Suppose we are given samples of data say X and Y, s.t.

$$X = x_1, ..., x_n$$
$$Y = y_1, ..., y_n$$

For example, we may be measuring the blood pressure and heart rate of Alice at time k, say x_k and y_k respectively. Further, suppose we are unaware of her context, for example, Bob hacked into Alice's apple watch and so can only read X and Y – he has no idea of anything she might be up to.

Bob then observes the following trend:

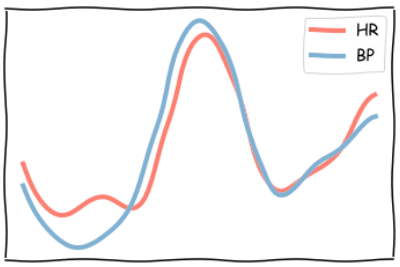


Figure 1.1 – Heart rate (HR) and Blood pressure (BP) of Alice.

Bob, having studied data science, is well aware of the fallacy of the law of small numbers[1]. He

[1] The law of small numbers is the error of concluding too much from too few data.

therefore checks again the data the next day at a slightly different time. He again observes a similar trend, and is now more confident in the existance of a causal relation and – having neglected biology – makes the conjecture that either blood pressure causes heart rate, or perhaps the other way around.

Given this strong correlation, Bob asserts that he may either model X as a function of Y or the other way around. He proceeds to find some f s.t. $f(X) \approx Y$. The next day, to his dismay, he notices that his model has terrible performance when evaluated on new data. He then proceeds to see what is going on, and observes the following:

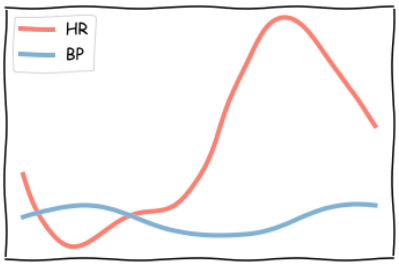


Figure 1.2 – Heart rate (HR) and Blood pressure (BP) of Alice.

As it turns out, in the last few days, Alice was working hard on finishing her thesis and the deadline had been the previous day. But how, Bob wondered, could this have changed the relationship between BP and HR? Finally, admitting to himself that machine learning alone is not enough to understand the world; Bob spends some time learning about the heart. It turns out, that fear triggers a "flight or fight" response that increases both the heart rate and blood pressure; Interestingly your heart rate and blood pressure won't always rise and fall in sync.

So what did Bob learn[2]?

1. When we train a model with some data, when we use it on some newly aquired data, we might face a **covariate-shift** – that is, the distribution might change due to the context changing.

2. When we see correlation it might be spurious due to a **confounder** – fear was the **confounder** of the heart rate and blood pressure.

3. His degree in Data Science is worth less than he thought; **machine learning is in fact not a panacea**, contrary to common culture. However, applied with domain knowledge and causal reasoning it may be useful.

[2] Note that heart rate and blood pressure are intimately linked, and the story between them is more complicated. The plots were randomly generated using a gaussian process, however they do resemble some real examples that can be found in the internet.

If Bob was able to incorporate these notions into his machine learning models, then it might have been more robust to the covariate-shift. To give a more concrete example, there is a "neural net tank urban legend"[3] , where a neural network accuratly predicts if there is a tank or not in an image, but it turns out it uses the weather as a predictor. From this it is clear that the model will preform badly under covariate shift, and indeed it makes the case that incorporating causality to a model should make it more robust as Schölkopf (2019) argues. Note that this is in effect the issue with generalisation in machine learning: how can we ensure that we learn *meaningful* representations (features about the tanks) rather than just correlations (the weather) useful for train accuracy.

As for confounders, it is impossible to say anything in general[4]. We must therefore specify a causal model, and then see what gurantees we can give under what assumptions. Even in the absence of confounders it is highly non trivial to determine causality.

As this simple example illustrates, causality is related to many interesting questions; perhaps, one of the most simple questions we can ask – and the one that we will explore – is, given that either X causes Y, or Y causes X (we assume no confounders) then, when can we predict the direction of causality? If yes, how?

In the figure bellow (figure 1.3) can you tell if X causes Y? Or perhaps it is the other way around? The right answer is that X causes Y, and we will show algorithms that can accuretly predict causality in such settings with as few as 75 samples.

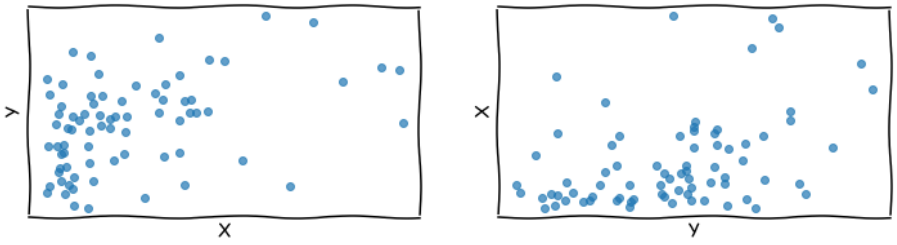


Figure 1.3 – 75 samples of data X, Y. The samples are generated independently as follows: $y_i = f(x_i) + n_i$ where x_i is drawn from an expoential distribution and n_i is drawn independently from a gaussian one and $f(x) = 10\tanh(x) + 4\sin(x) + x + x^2$

[3] More about this story here: https://www.gwern.net/Tanks.

[4] For most of the 20th century, a huge debate took place to determine the question of whether or not smoking caused cancer. A clever argument against a causal relation was that there existed a gene that made a person both want to smoke and more prone to cancer; even the father of modern statistics himself thought this explanation more plausible (For a good read on how science is and was used for wrong see the excellent book Oreskes (2011)).

1.2 Causality

1.2.1 Causal models: FCM

We can model any causal model by using a **Function Causal Model (FCM)** which can be constructed as follows:

We *generate* a random vector $X = (X_1, ..., X_d)$ by using a graph $\mathcal{G}$ (encoding the relationships), a set of functions $f = (f_1, ..., f_d)$ (encoding the type of relationship) and a noise distribution $\mathcal{E}$ (the randomness generator).

For each $i = 1, ..., d$

$$X_i \leftarrow f_i(X_{\mathrm{Pa}(i)}, E_i), \quad E_i \sim \mathcal{E}$$

where $\mathrm{Pa}(i)$ is the set of parents of i; and so $X_{\mathrm{Pa}(i)}$ is the set of random parent random variables of X_i. For example in Figure 1.4, $\mathrm{Pa}(4) = \{1,2\}$, and so $X_{\mathrm{Pa}(4)} = X_1, X_2$.

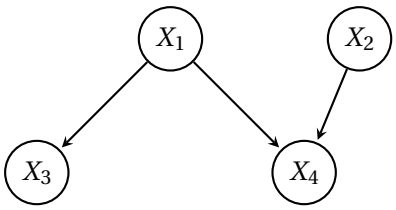


Figure 1.4 – Example for FCM with $X = X_1, ...X4$, with $E_i \sim \mathcal{E}$, with $X_1 = f_1(E_1)$, $X_2 = f_2(E_2)$, $X_3 = f_3(X_1, E_3)$ and $X_4 = f_4(X_1, X_2, E_4)$

Note that causal relations can also be *cyclic*, i.e. X causes Y which in turn causes X ad infinitum. While this deserves consideration as many systems have feedback loops we will for simplicty not look at such settings.

1.2.2 Interventions

To make precise the meaning of causality, suppose that we are given two random variables X, Y with joint distribution $p_{x,y}$. Intuitively we would say that X causes Y, or $X \rightarrow Y$, if we intervene on X and then see an effect on Y. In particular we will denote $\mathrm{do}(x)$ – short for $\mathrm{do}(X = x)$ – as an intervention that forces the variable X to have the value x, and leaves the rest of the system untouched. Following the convention inspired by Pearl (2000), we define the resulting distribution as $p_{y|do(x)}$.

This motivates the following definition:

Definition 1 *We say that X **causes** Y if $p_{y|do(x)} \neq p_{y|do(x')}$ for some x, x'*

When we talk about $p_{y|x}$, we often say, "The chance of y given that x happened". This sounds similar to $p_{y|do(x)}$; note however that "x happened" and "force $X = x$" are very different. Imagine that there indeed was a gene that made people both prone to smoking and cancer; then if we forced someone at random to smoke, he would on average be less likely to have cancer than someone who smoked because he wanted to. This also illustrates one of the limitations of causality: some interventions are not possbile due to ethical issues.

You might have heard about randmoized trials or A/B testing, these are both ways to estimate $p_{y|do(x)}$. For example, when developing cures, the idea of a random trial is to give experimental drugs to participants at random. When designing new UIs to maximise user participation in apps, developers implement A/B testing, they assing new versions to people at random to estimate engagement. Note that in both of these, we are able to avoid a potential confounder by using picking x's at random, and "forcing" them to "do(x)".

Since we will restrict ourselves to the observational setting, we will not be able to perform any interventions, which would allow us to estimate $p_{y|do(x)}$. In this setting however, in order to perform any meaningful inference, we will need to make concessions; in particular, we will make some assumptions about the causal structure. If and when we are able to that we can infer the causal structure in such a setting, we shall call it **identifiable**.

We will restrict ourselves also to the bivariate case; one big difference worth noting is that in the multivariate setting we can test conditional independence. Uusing conditional independence tests is a very powerful method for causal inference. Suppose we have random variables X, Y and Z, then if we can estimate that $X \perp\!\!\!\perp Y|Z$ then it must be that all information between X and Y must flow through Z (See Figure 1.5). With one test we were able to pinpoint the causal graph $\mathcal{G}$!

One can in fact generalise the conditional independence such that X and the other variables are a collection of random variables, which gives a lot of flexiblity to devise clever algorithms. The theory comes from graphical models, which tries to understand the relationship between distributions and their graphical counterparts, such as Figure 1.5. The key difference is that in graphical models we do not care about the causal direction. Since we will not be using any of this theory, we will not go into any detail.

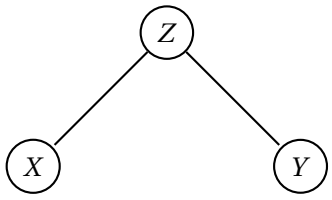


Figure 1.5 – An example of FCM with random variables X, Z and Y; we leave it undirected

In some sense the two variable case is hard because we cannot use conditional independence. As we will see, it is not possible to distinguish causality in the general bivariate setting when only observing observational data; we will thus need to restrict the class of such models. The

underlying structure behind such causal models is what is known as Strucutual Equation Models (SEM). Essentially it is a model specification; and the key insight is that it should not be reversible.

1.3 Proposed Methods

We propose two methods to deal with the ANM – one less practical, but with a nice theoretical analysis; and another more practical, but perhaps with a less beutiful analysis. However both are based on a first principle approach with known asymptotics in mind. TODO DESCRIBE IN MORE DETAIL

1.4 Outline

The first part is dedicated to covering background material. The scond part will cover the proposed methods in more details as well as showing experimental results TODO TALK ABOUT MY

Part I Preliminaries

2 Causal Inference

2.1 Bivariate causal model

2.1.1 ANM

We will now introduce[1] the bivariate causal model and define the particular subset of such models that we will work on. For the bivariate case, if we have $Y \in \mathbb{R}$ as a direct cause of $X \in \mathbb{R}$, then we can model the relationship as follows:

$$\begin{cases} Y = f(X, Z) \\ X \perp\!\!\!\perp Z, \quad X \sim p_X, \quad Z \sim p_Z \end{cases} \tag{2.1}$$

where p_X is the density of the cause and p_Z that of the latent variable; $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a borel measurable function (w.r.t. to the Borel sets of $\mathbb{R} \times \mathbb{R}$ and $\mathbb{R}$).

If we assume that there is no confounder, no sampling bias[2] and no cycles then it is natural to assume that $X \perp\!\!\!\perp Z$.

While we can model Z as either a scalar or a vector, we can without loss of generality assume it to be a scalar. It can be shown that if Z is a vector, then one can construct a simpler model with scalar noise, which has the same observational and interventional distribution Mooij et al. (2016).

An important remark is that given the direct model in equation 2.1, we can find some $\tilde{f}$ and $\tilde{Z}$

[1] Note that this introduction follows closely that of Mooij et al. (2016), and we encourage the reader to have a look to fill in details that have been omitted here.

[2] If you consider the example given in the introduction, then if Bob always sampled when Alice was excercising, then this would have lead to sampling bias, as during excercise both heart rate and blood pressure tend to increase.

such that

$$\begin{cases} X = \tilde{f}(Y, \tilde{Z}) \\ Y \perp\!\!\!\perp \tilde{Z}, \quad Y \sim p_Y, \quad \tilde{Z} \sim p_{\tilde{Z}} \end{cases} \tag{2.2}$$

with the important property that it induces an equivalent observational distribution $p_{X,Y}$ as that of equation 2.1. However in general the interventional distribution will differ. In particular this means that with observational data alone we are not able identify the right causal direction.

We must therefore make further assumptions on f that break this symmetry and allow us to make causal inference on observational data alone. In particular we will consider the following class of models. Note that these models are a subset of those that we just introduced (equation 2.1).

Definition 2 *Given a triplet (p_X, p_Z, f), consisting of two finite mean densities and a Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$, we can define a **bivariate Additive Noise Model (ANM)** $X \to Y$*

$$\begin{cases} Y = f(X) + Z \\ X \perp\!\!\!\perp Z, \quad X \sim p_X, \quad Z \sim p_Z \end{cases}$$

If the induced joint density of X and Y, $p_{X,Y}$ has a density with respect to Lebesgue measure, we say that $p_{X,Y}$ satisfies the ANM $X \to Y$.

Given such a model, we are interested in the cases when the observational distribution $p_{X,Y}$ can only lead to omly one causal explenation; this motivates the following definition:

Definition 3 *If the joint density $p_{X,Y}$ satisfies an ANM $X \to Y$, but does not satisfy ANM $Y \to X$, then we call the ANM $X \to Y$ **identifiable**.*

Intuitively, non-linearities due to f will break the symmetry needed to make a reverse ANM. This is what Hoyer et al. (2009) and friends explore; they show that for the a triplet (p_X, p_Z, f) to generate a non-identifiable ANM, it needs to satisfy a particular differential equation. Loosely speaking cannot happen in the generic case: in other words the forward model $X \to Y$ cannot be inverted.

If f is linear, then one can give a much more precise statement about identifiability:

Theorem 1 *Let X and Y be random variables, such that*

$$Y = aX + Z, \quad X \perp\!\!\!\perp Z, \; a \neq 0$$

Then we can reverse the process, i.e. there exists $\tilde{a} \in \mathbb{R}$ and a noise $\tilde{Z}$ such that

$$X = \tilde{a}Y + \tilde{Z}, \quad Y \perp\!\!\!\perp \tilde{Z}$$

if and only if $X, Y, Z, \tilde{Z}$ are Gaussian distributed.

The proof is a simple appilcation of the Darmois-Skitovich Theorem[3]; it states the following:

Theorem 2 (Darmois-Skitovich) *Let X_i, $i \in [n]$ be independent random variables, and let α_i, β_j be non zero constants. Then, if the random variables*

$$L_1 = \sum_{i \in [n]} \alpha_i X_i$$

$$L_2 = \sum_{i \in [n]} \beta_i X_i$$

are independent, i.e. $L_1 \perp\!\!\!\perp L_2$; then all the random variables X_i are gaussian.

We now prove Theorem 1 using the Darmois-Skitovich Theorem.

Proof: For the "only if" part, note that by simple manipulation, we have the following:

$$\begin{bmatrix} Y \\ \tilde{Z} \end{bmatrix} = \begin{bmatrix} a & 1 \\ 1 - \tilde{a}a & -\tilde{a} \end{bmatrix} \cdot \begin{bmatrix} X \\ Z \end{bmatrix}$$

If $\tilde{a} \neq 0$ and $1 - \tilde{a}a \neq 0$ then by Darmois-Skitovich, the result follows.

We will next show that both of these conditions must be true for the process to be reversible:

1. If $\tilde{a} = 0$ then $\tilde{Z} = X$, but then $X \perp\!\!\!\perp Y$, a contradiction[4].
2. Finally, if $1 - \tilde{a}a \neq 0$ then $\tilde{Z} = -\tilde{a}Z$, and thus $-\tilde{a}Z \perp\!\!\!\perp Y$, a contradiction.

We have thus show the "only if" part.

We next show the "if" part; first assume that X and Z are Gaussian random variables. It is easy to verify that: $\mathrm{Cov}(Y, \tilde{Z}) = a(1 - a\tilde{a})\mathrm{Var}(X) - \tilde{a}\mathrm{Var}(Z)$. Thus if we set $\tilde{a} = \frac{a\mathrm{Var}(X)}{a^2\mathrm{Var}(X) + \mathrm{Var}(Z)}$

[3] The theorem also plays an imporant role in independent component analys (ICA), in short it deals with source separation. It turns out that if we have the linear multivariate causal setting, then we can cast it as a source separation problem and use ICA to solve it. Note that there too gaussianty makes or breaks the method.

[4] Note that $X \perp\!\!\!\perp aX + Z$ is trivially false in the discrete case, but if both are continuous then we need to be a bit more careful as is the case with degenerate random variables; but essentially the same holds (see Peters (2008)).

we get that $\text{Cov}(Y, \tilde{Z}) = 0$, and since they are gaussian random variables we get that they are also independent.

□

These results show that indeed, in most cases, we should be able to perform causal inference as most additive models should be identifiable. Interestingly, the guaussian setting provides difficulty due to the symmetry of the gaussian (in the linear case). In general the gaussian is our friend, but not today. We will next explote some methods for causal inference.

2.2 ANM Methods

The additive noise model (ANM) methods are a very simple score based family of methods for causal inference. The following lemma[5] motivates the method:

Lemma 1 *Given a joint density $p_{X,Y}$ of two random variables X, Y s.t. the conditional expectation $\mathbb{E}(Y|X = x)$ is well-defined for all x and measurable. Then, $p_{X,Y}$ satisfies a bivariate Additive Noise Model $X \to Y$ if and only if $E_Y := Y - \mathbb{E}(Y|X)$ has finite mean and is independent of X.*

In practice we get some data from $p_{X,Y}$; say $\mathcal{D}_N = \{(x_i, y_i)\}_{i\in[N]}$. We can then either split it into a test/train in order to first fit a regression which we then evaluate using the test set. If the data is scarce we may alternatively recycle the data – i.e. reuse it for both training and evaluation.

First we estimate through regression the function $x \mapsto \mathbb{E}(Y|X = x)$, say $\hat{f}$; we then compute the estimated residual $\hat{e} = \hat{f}(X) - Y$. Next, we estimate the dependence between $\hat{e}$ and X using a score function C; i.e. C could be the empirical mutural information betheen them. Thus a low score would be evidance for an ANM in that direction; we can compute the score for the reverse model by switching the roles of X and Y. We can then compare the scores and use this as a criteria for inference. We write down this idea more explictly here Algorithm 1.

In order to show that such a procedure is consistent we need 3 things:

1. $p_{X,Y}$ satisfies either $X \to Y$ or $Y \to X$, but not both.
2. The regression method should be **suitable** for regressing Y on X.
3. If $X \to Y$, then asymptotically $\hat{C}_{X\to Y} < \hat{C}_{Y\to X}$

We will take point 1 as an assumption as there is currently no theoretical result that allows for a consistent test to check if $p_{X,Y}$ satisfies an ANM $X \to Y$.

[5] A simple proof can be found here Mooij et al. (2016)

In point 2, by **suitable** regression we mean that $\|\hat{\mathbf{e}} - \mathbf{e}\| \to 0$. Essentially we require the regression used on the data to have 0 mean square error in expectation.

More precisely, given two real-valued random variable X, Y, with joint distribution $p_{X,Y}$. If we are given two sequences – for training and test – say $D_N = X_1, ..., X_N$ and $D'_N = X'_1, ..., X'_N$. We say that a regression method is **suitable** for regressing Y on X satisfies

$$\lim_{N\to\infty} \mathbb{E}_{\mathcal{D}_N, \mathcal{D}'_N} \left(\frac{1}{N} \sum_{n=1}^{N} \left| \hat{f}_Y \left(X'_n; \mathcal{D}_N \right) - \mathbb{E}\left(Y \mid X = X'_n \right) \right|^2 \right) = 0$$

Algorithm 1 General procedure to decide whether $p_{X,Y}$ satisfies and ANM $X \to Y$ or $Y \to X$

Input:

1. I.i.d samples $\mathcal{D}_N = \{(x_i, y_i)\}_{i\in[N]}$ of X and Y
2. Regression method
3. Score estimator $\hat{C} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

Output: $\hat{C}_{X\to Y}$, $\hat{C}_{Y\to X}$, dir

1. Split the $\mathcal{D}_N$ in half randomly to obtain $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$
2. Use the regression method on the training data $\mathcal{D}_{train}$:
   - $\hat{f}_X$ of the regression function $x \mapsto \mathbb{E}(Y|X = x)$
   - $\hat{f}_Y$ of the regression function $y \mapsto \mathbb{E}(X|Y = y)$
3. Estimate residuals using the predicted regressions on the test data $\mathcal{D}_{test}$:
   - $\hat{\mathbf{e}}_\mathbf{Y} := \mathbf{y} - \hat{f}_Y(\mathbf{x})$
   - $\hat{\mathbf{e}}_\mathbf{X} := \mathbf{x} - \hat{f}_X(\mathbf{y})$
4. Compute scores to measure dependence between inputs and estimated residuals based on the test data $\mathcal{D}_{test}$
   - $\hat{C}_{X\to Y} := \hat{C}(\mathbf{x}, \hat{\mathbf{e}}_\mathbf{Y})$
   - $\hat{C}_{Y\to X} := \hat{C}(\mathbf{y}, \hat{\mathbf{e}}_\mathbf{X})$
5. Output $\hat{C}_{X\to Y}$, $\hat{C}_{Y\to X}$, and

$$\text{dir} := \begin{cases} X \to Y & \text{if } \hat{C}_{X\to Y} \le \hat{C}_{Y\to X} \\ Y \to X & \text{otherwise} \end{cases}$$

We will now overview some score functions, note that the list is far from exhaustive and other methods can be found in Mooij et al. (2016). We will end the chapter by presenting to

alternative methods.

2.2.1 HSIC score

First considered by Hoyer et al. (2009), is the Hilbert-Schmidt independence Criterion (HSIC) for testing the independence between the residuals and the input.

$$\hat{C}(u,v) := \widehat{\mathrm{HSIC}}_{k,l}(u,v)$$

We will give a formal description of the HSIC in the next chapter, where we will develop some of the background to get an intuition on this measure. For now, you can think of the HSIC as a metric that computes the distance between the the product distribution and the joint (similar to the Mutual information); the key difference is that we provide two kernels, k and l which will transform the distributions to a different space. For specific kernel choices the HSIC indeed becomes a metric.

Note that one can approximate the distribution of $\widehat{\mathrm{HSIC}}$, and use hypothesis testing for the independence test.

2.2.2 Entropy score

Another type of score function looks at differential entropies instead of directly testing for independence. These ideas stem from Kpotufe et al. (2014) and Nowzohour and Bühlmann (2016); The following lemma shows how this might be used in practice:

Lemma 2 *Consider random variables X and Y, with joint density $p_{X,Y}$. For any functions $f, g : \mathbb{R} \to \mathbb{R}$ we have:*

$$H(X) + H(Y - f(X)) = H(Y) + H(X - g(X)) - I(X - g(X), Y) + I(Y - f(X), X)$$

where $H(\cdot)$ denotes the differential entropy and $I(\cdot,\cdot)$ denotes the differential mutual information (Cover (1999)).

The proof is a simple application of the chain rule. Note that if $X \to Y$ then $I(Y - f(X), X) = 0$; since $I(X,Y) \geq 0$ for any X, Y it follows that:

$$H(X) + H(Y - f(X)) \leq H(Y) + H(X - g(X))$$

Which motivates the score function

$$C(U,V) = H(U) + H(V)$$

This approach to estimate the direction of the ANM is consistent under certain assumptions as is shown by Kpotufe et al. (2014) and Nowzohour and Bühlmann (2016). One of the main drawbacks of using differential entropy is that we need to go through discretization, which can lead to undesired effects. TODO give example

2.2.3 Other methods

2.2.4 IGCI

Probably will skip this?

2.2.5 CGNN: Causal Generative Neural Nets

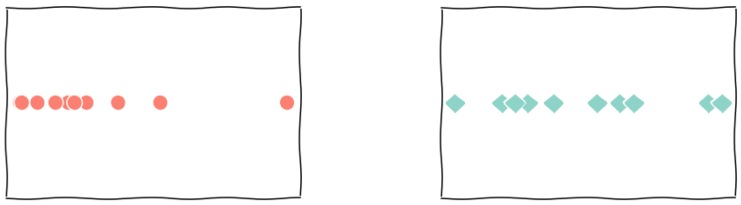
In the work of Goudet et al. (2017), they estimate a *generative* model by approximating the FCM structure using neural networks given some data from $p_{X,Y}$. Using the same notation as in the introduction on FCM, the idea is to estimate each f_i by a neural network, and the search through the DAG space. Since the DAG space is super exponetial in the number of variables, they apply a greedy procedure to decide whether or not to include an edge $X_i \to X_j$. In some sense this is similar to greedy methods used in model selection. The caveat here is that using the generative model they can backpropagate to learn all the f_i simulatneously.

More specifically, given the current graph estimate $\mathcal{G}$, they can generate some $\hat{P}$ from the current f_i and some noise $\mathcal{E}$. Then they can train the model by using the MMD[6] as a loss function between $\hat{P}$ and P (the data distribution).

So far the CGNN appears to have the best performance on the various benchmarks.

[6] We present and give some background on the MMD in the next chapter

3 Statistical distance



$X \sim P$ $Y \sim Q$

Figure 3.1 – Samples from two different sources, X and Y, how can we tell if they come from the same distribution?

Suppose that we are given samples from two unkown distributions P and Q, an important question to ask is: are P and Q equal?

The Integral Probability Metric (IPM) and f-divergence are two very rich and well studied families of measures of "distance" between probability measures.

We start by introducing the Reproducing Kernel Hilbert Spaces (RKHS), which will serve as a building block for the maximum mean discrepancy, an important instance of IPM.

3.1 Reproducing Kernel Hilbert Space

We will begin by defining the kernel,

3.1.1 Kernels

Definition 4 *Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if*

1. *k is symmetric: $k(x,y) = k(y,x)$.*
2. *k is positive semi-definite, i.e. $\forall x_1,...,x_n \in \mathcal{X}$, the "Gram Matrix" K, defined by $K_{ij} = k(x_i,x_j)$ is positive semi-definite* [1].

It is easy construct new kernels since they are preserved under addition, multiplication and other operations. (See for example Gretton (2019)).

One example of a kernel – and one of the most popular ones – is the Gaussian Kernel defined on $\mathbb{R}^d$:

$$k(x,y) = \exp(-\gamma^{-2} \left\| x-y \right\|^2)$$

3.1.2 Constructing the Reproducing Kernel Hilbert Space

Let $\mathcal{X}$ be an arbitrary set and $\mathcal{H}$ a Hilbert space of real valued functions on $\mathcal{X}$. As per general convention, addition and multiplication are define pointwise:

$$\begin{array}{lll} (\lambda \cdot f)(x) & := \lambda \cdot f(x) & \forall \lambda \in \mathbb{R}, \forall f \in \mathcal{H} \text{ and } \forall x \in \mathcal{X} \\ (f+g)(x) & := f(x) + g(x) & \forall f \in \mathcal{H}, \forall g \in \mathcal{H} \text{ and } \forall x \in \mathcal{X} \end{array} \tag{3.1}$$

We will now take a look at Hilbert spaces whose structure is highly linked with a kernel. Note that if we pick some $x \in \mathcal{X}$, then $k(x,.)$ is a function from $\mathcal{X}$ to $\mathbb{R}$.

Definition 5 *Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space (RKHS) if there is a kernel k such that*

1. $k(x,\cdot) \in \mathcal{H} \quad \forall x \in \mathcal{X}$
2. $\left\langle f, k(x,\cdot) \right\rangle = f(x) \quad \forall f \in \mathcal{H}$

Given the kernel k it is convinient to define the feature map $\phi : \mathcal{X} \to \mathcal{H}$ as:

$$\phi(x) = k(x,\cdot)$$

The intuition is that in this space, we can view functions as linear combinations[2] of features:

$$f(x) = \left\langle f, k(x,\cdot) \right\rangle = \left\langle f, \phi(x) \right\rangle$$

[1] A matrix $M \in \mathbb{R}^{n \times n}$ is positive semi-definite if $\forall a \in \mathbb{R}^n$, $a^\intercal M a \geq 0$

[2] Note that if $f(x)$ is an element of $\mathcal{H}$, then we write f as the coefficients for the feature representation.

The power of this setup – which is known as the kernel trick – is that inner products between features (which can live in infinite spaces) are simple function evaluations; indeed by letting $f(x) = k(x, x\prime)$ we get

$$\left\langle k(x', \cdot), k(x, \cdot) \right\rangle = k(x, x')$$

Observe that both conditions imply that k spans $\mathcal{H}$, i.e.

$$\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}} \tag{3.2}$$

Indeed is is possible to go the other way around[3] and first define the following vector space

$$\text{span}(\{\phi(x) : x \in \mathcal{X}\}) = \left\{ f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : n \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R} \right\} \tag{3.3}$$

We can then equip this space with an inner product and to show that it is complete in order to create a Hilbert Space (at which point we will have created a RKHS).

3.1.3 The kernel trick in action

We will now show an application to illustrate both the power of the RHKS and to refine our intuition of it. Suppose that we have some data say $\{x_i, y_i\}_{i \in [n]}$; we belive for example y to be a smooth function of x and we expect some independent aditive noise.

We can estimate f as follows [4] , pick an RHKS $\mathcal{H}$ with a gaussian kernel:

$$f^* = \arg\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \left\langle f, \phi(x_i) \right\rangle_{\mathcal{H}} \right)^2 + \Omega \left\| f \right\|_{\mathcal{H}}^2 \right) \tag{3.4}$$

An amazing result is that an optimisation of the above form will always admit a representation of the form:

$$f^* = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

where $\alpha_i \in \mathbb{R}$ for all $1 \leq i \leq n$

[3] See the excellent lecture notes on RKHS Bartlett (2008) for more details.

[4] Note that it is not obvious how to implement the optimisation as $\mathcal{H}$ may be infinite. However, this setup with a gaussian kernel is in fact equivalent to a Gaussian Processes, which can be easily implemented in practice (see Jordan (2004)).

This is known as the Representer Theorem (Schölkopf et al. (2001)); all it requires is that we be in the usual RHKS setup, and that the regularisation be a strictly increasing[5] real valued function. If we wish to approximate a prediction for some new sample x, we can do so as follows:

$$f^*(x) = \left\langle f^*, \phi(x) \right\rangle = \sum_{i=1}^{n} \alpha_i \left\langle \phi(x_i), \phi(x) \right\rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

It is precisely because the solution is of this form, that we may exploit the kernel trick. We can also quickly see what the role of the kernel is. If for example, k is the Gaussian Kernel, then the solution will be a linear combination of scaled gaussians centered at the data points [6].

As a final remark we will explain the role of the penalty $\Omega \left\| f \right\|_{\mathcal{H}}^2$; from statistical models, we now that this kind of term is known as regularisation and is supposed to help choose a "simpler" model. As we will now show, this is also the case here.

To see this, we will use Mercer's Theorem – a Generalisation of the spectral theorem for positive-semidefinite matrices[7].

Theorem 3 (Mercer's) *Suppose k is a continuous positive semi-definite kernel on a compact set $\mathcal{X}$, then if, $\forall f \in L_2(\mathcal{X})$*

$$\int_{\mathcal{X}} k(u,v) f(u) f(v) du dv \geq 0$$

then k has the following decomposition

$$k(u,v) = \sum_{i=1}^{\infty} \lambda_i \psi_i(u) \psi_i(v) \tag{3.5}$$

Where $\{\psi_i\}$ forms an orthonormal basis of $L_2(\mathcal{X})$, such that the corresponding sequence of eigenvalues $\{\lambda_i\}$ are non-negative.

Where the convergence is absolute and uniform, that is,

$$\lim_{n \to \infty} \sup_{u,v} \left| k(u,v) - \sum_{i=1}^{n} \lambda_i \psi_i(u) \psi_i(v) \right| = 0$$

We can now use this decomposition of the Kernel to get further insight, using Mercer's theorem we can thus write – assuming the conditions are met:

[5] In our case regularisation is linear, we thus simply need to pick $\Omega \geq 0$.

[6] In fact this will always be the case when we can write $k(x_i, x) = \tilde{k}(x_i - x)$

[7] Recall that our Kernel k is a generalisation of a positive-semidefinite Matrix

$$k(x,x') = \sum_{i=1}^{\infty} \underbrace{\left[\sqrt{\lambda_i}\psi_i(x)\right]}_{\phi_i(x)} \underbrace{\left[\sqrt{\lambda_i}\psi_i(x')\right]}_{\phi_i(x')}$$

We can thus rewrite the solution as follows

$$f^*(x) = \sum_{i=1}^{n} \alpha_i k(x_i,x) = \sum_{i=1}^{\infty} \phi_i(x) \sum_{j=1}^{n} \alpha_j \phi_i(x_j) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(x) f_i^*$$

Note that due to the $\Omega \|f\|_{\mathcal{H}}^2$ penalty, f_i^* must decay for higher values of i. Note that for example in the Fourier Transform , in the basis $\{\psi_i\}$, higher values of i correspond to higher frequency functions; similarly, for the Gaussian Kernel, higher indicies basis functions correspond to higher frequecies[8]. Thus, a higher Ω will force a faster decay on f_i and thus result in smoother functions – in principle, this will reduce overfitting.

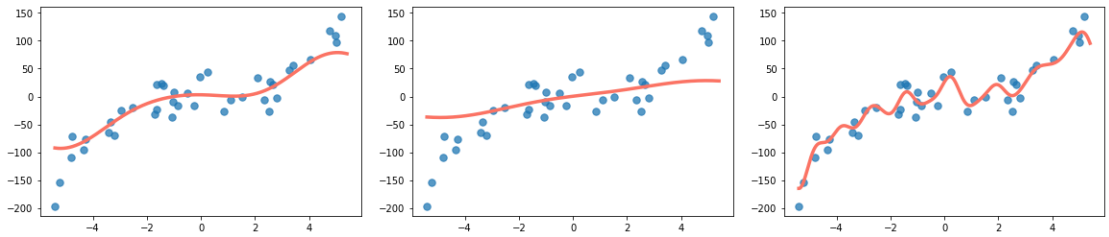


Figure 3.2 – Small RKHS norm results in smooth functions. From left to right $\Omega = 2$, $\Omega = 14$, $\Omega = 0.2$, we fix the Gaussian kernel with $\gamma = 0.6$

3.2 Integral Probability Metric

3.2.1 Introduction

We now turn to the question of statistical distance, i.e. given samples of P and Q, how can we determine if $P = Q$?

Observe that if two random variables X, Y share the same distribution, then

$$\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$$

for any continuous and bounded function $g : \mathbb{R} \to \mathbb{R}$. It turns out that the reciprocal statement holds. (See Gretton et al. (2012))

This motivates the following construction

[8] In the fourier space, we have the following basis $\psi_\omega = \exp(2\pi i x\omega)$

$$D_{\mathcal{F}}(P,Q) = \sup_{g\in\mathcal{F}} | \mathop{\mathbb{E}}_{X\sim P} g(X) - \mathop{\mathbb{E}}_{Y\sim Q} g(Y) |$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions.

This defines a rich class of distance measures known as integral probability metrics (IPMs) (see Müller (1997)). Depending on how we choose $\mathcal{F}$ we may end up with different popular distance measures, such as the Wasserstein distance or the Total variation distance to name a few.

The goal is to craft an $\mathcal{F}$ that is "expressive" enough so that the IPM vanishes iff $P = Q$, and on the other hand, we need $\mathcal{F}$ to be "restrictive" enough so as to have fast and reliable guarantees of the empirical estimate of the IPM (Gretton et al. (2012).)

3.2.2 MMD

Consider $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, this is known as the maximum mean discrepancy (MMD). Where $\mathcal{H}$, is a reproducing kernel Hilbert space (RHKS) with k as its reproducing kernel.

We will next extend the notion of the feature map to the **embedding of probability distributions**. Recall that if ϕ is the associated feature map to the kernel k from RKHS $\mathcal{H}$ then we have $g(x) = \langle g, \phi(x)\rangle$.

We define $\mu_P \in \mathcal{H}$, s.t. $\forall g \in \mathcal{H}$, we have that $\mathbb{E}_X\, g(X) = \langle g, \mu_P\rangle$. We will now show under which conditions μ_P exits.

Lemma 3 *If k is measurable and $\mathbb{E}_X \sqrt{k(X,X)} < \infty$ then $\mu_P \in \mathcal{H}$*

Proof:

$$\begin{aligned}
\left|\mathbb{E}_X\, g(X)\right| &\leq \mathbb{E}_X\, |g(X)| \\
&= \mathbb{E}_X \left|\langle g, \phi(X)\rangle_{\mathcal{H}}\right| \\
&\leq \mathbb{E}_X \left\|g\right\|_{\mathcal{H}} \left\|\phi(X)\right\|_{\mathcal{H}} \\
&= \left\|g\right\|_{\mathcal{H}} \mathbb{E}_X \sqrt{k(X,X)}
\end{aligned}$$

Thus $E_X g(X)$ is a bounded linear operator $\forall g \in \mathcal{F}$, and by the Riesz representer theorem it follows that there exists a $\mu_P \in \mathcal{H}$ s.t. $\mathbb{E}_X\, g(X) = \langle g, \mu_P\rangle$. □

We can also see that the mean embedding of the distribution P is the expectation under P of the feature map ϕ.

$$\mathop{\mathbb{E}}_{X\sim P} g(X) = \left\langle g, \mathop{\mathbb{E}}_{X\sim P} \phi(X) \right\rangle = \langle g, \mu_P \rangle$$

Asumming Lemma 3 – and using Cauchy-Schwartz, we can explicitly solve the MMD in terms of the mean embeddings:

$$\begin{aligned}
\mathrm{MMD}_{\mathcal{F}}(P,Q) &= \sup_{g\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{X\sim P} g(X) - \mathop{\mathbb{E}}_{Y\sim Q} g(Y) \right| \\
&= \sup_{g\in\mathcal{F}} \left| \langle g, \mu_P - \mu_Q \rangle \right| \\
&= \left\| \mu_P - \mu_Q \right\|_{\mathcal{H}}
\end{aligned}$$

We can therefore see the MMD as the feature mean difference of the distributions; we can further expand this expression to get the result as a function of the kernel.

$$\begin{aligned}
\mathrm{MMD}^2_{\mathcal{F}}(P,Q) &= \left\| \mathop{\mathbb{E}}_{X\sim P} \phi(X) - \mathop{\mathbb{E}}_{Y\sim Q} \phi(Y) \right\|^2_{\mathcal{H}} \\
&= \mathop{\mathbb{E}}_{X\sim P} \mathop{\mathbb{E}}_{X'\sim P} \langle \phi(X), \phi(X') \rangle - 2 \mathop{\mathbb{E}}_{X\sim P} \mathop{\mathbb{E}}_{Y\sim Q} \langle \phi(X), \phi(Y) \rangle + \mathop{\mathbb{E}}_{Y\sim Q} \mathop{\mathbb{E}}_{Y'\sim Q} \langle \phi(Y), \phi(Y') \rangle \\
&= \mathop{\mathbb{E}}_{X\sim P} \mathop{\mathbb{E}}_{X'\sim P} k\left(X, X'\right) - 2 \mathop{\mathbb{E}}_{X\sim P} \mathop{\mathbb{E}}_{Y\sim Q} k(X,Y) + \mathop{\mathbb{E}}_{Y\sim Q} \mathop{\mathbb{E}}_{Y'\sim Q} k\left(Y, Y'\right)
\end{aligned}$$

Note that we can straightforwardly estimate with samples the above expression; all the we require is to specify a kernel: *so how do we choose a kernel*?

We need to ensure that $\mathrm{MMD}(P,Q) = 0$ iff $P = Q$, in other words, μ_P needs to be injective as a function of P. Intuitively this means that $\mathcal{F}$ needs to be expressive enough to reproduce enough continuous functions. One can show that to check if the resulting embedding μ_P is injective, we may check either of these sufficient conditions (Sriperumbudur et al. (2008)) on the Kernel k:

1. k is a universal kernel.
2. k is a convolution kernel on $\mathbb{R}^n$, for which the Radon-Nikodym derivative of its inverse Fourier transform is supported almost everywhere.

The first condition is basically what we knew intuitively: If we consider a compact metric space, say $(\mathcal{X}, d)$, then a Kernel k on $\mathcal{X}$ is called universe if the corresponding RKHS is dense in the space $C(\mathcal{X})$ of all continuous functions. The drawback is that the input space $\mathcal{X}$ needs to be compact – which excludes $\mathbb{R}^n$; this means that we cannot use universality to check our

gaussian kernel. Luckly the second condition is enough.

Assuming k is a bounded continuous positive definite function, then if we can write $k(x, y) = \psi(x - y)$ we say that k is a convolution kernel.

From inspection it is clear that the gussian kernel is convolutional

$$k(x, y) = \exp(-\gamma^{-2} \left\| x - y \right\|^2)$$

Recall that the fixed point of convolution is the gaussian, which trivially implies that the inverse Fourier transform of a gaussian is supported everywhere. This means that the gaussian kernel satisfies the second condition, and it therefore generates an injective embedding μ_P.

We note that HSIC is to MMD, what the Mutual Information is to the Kullback–Leibler divergence (in the sense that measure the distance between the joint and product distributions to test for independence).

In practice the gaussian kernel is very popular, it is used in the HSIC test when used as a score function by Mooij et al. (2016); but how do we find the parameter – sometimes referred to as the bandwith – of the kernel?

One approach is the median heuristic (Schölkopf et al. (2002)):

$$\hat{\gamma}(u) := \text{median}\left\{ \left\| u_i - u_j \right\| : i < j, \left\| u_i - u_j \right\| \neq 0 \right\}$$

3.2.3 The case for MMD

In their study, Sriperumbudur et al. (2009)) argue that the "IPM is much simpler than estimating f-divergences, and that the estimators are strongly consistent while exhibiting good rates of conver- gence. IPMs also account for the properties of the underlying space M through the Kernel in case of MMD. This is especially useful when considering disjoint supports between P and Q"

Another argument for the MMD, is that we only need to choose a kernel; in contrast, when applying the f-divergence in practie we need to quantize in order to get an emprical distribution. While both can be seen as a hyperparameter, the effect of discretisation is not as obvious as that of choosing a kernel.

3.3 f-divergence

The f-divergence is another family of probability measures, and more simple than IPM.

Definition 6 (f-divergence) *Let P and Q be two probability distributions over a space Ω, such that P is absolutely continuous with respect to Q; and let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. The f divergence of P from Q is defined as*

$$D_f(P \parallel Q) := \int_\Omega f\left(\frac{dP}{dQ}\right) dQ$$

We first show why divergence has some desirable properties for a probability measure:

$$\begin{aligned} D_f(P \parallel Q) &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \\ &\geq f\left(\mathbb{E}_Q\left[\frac{dP}{dQ}\right]\right) \\ &= f\left(\int_\Omega \frac{dP}{dQ} dQ\right) \\ &= f(1) \\ &= 0 \end{aligned}$$

The inequality follows from the convexity of f, this tells us that $D_f(P \parallel Q) \geq 0$. From the definition it is clear that $D_f(P \parallel P) = 0$; further, if f is *strictly* convex at 1, then we have that $D_f(P \parallel Q) = 0$ iff $P = Q$.

Therefore roughly speaking, all f-divergences define a way to measure similiarities between distributions.

However, in general it is not symmetric in P and Q, so it is not a metric.

The following are some examples of f-divergences:

- **Kullback-Leibler (KL) divergence**: $f(x) = x\log(x)$
- **Total Variation (TV)**: $f(x) = \frac{1}{2}|x-1|$, note that in this case we have

$$D_f(P \parallel Q) = \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{dP}{dQ} - 1\right|\right] = \frac{1}{2}\int_\Omega |dP - dQ|$$

  Note that the TV is also a metric on the space of probability distributions.

3.4 Independence tests

It is rather straightforward to come up with independence tests once we are we are able to test for the distance between distributions. Say we are given two random variables X and Y,

with values over the product space $\mathcal{X} \times \mathcal{Y}$. If their joint distribution is $p_{X,Y}$, and their marginal distributions are p_X and p_Y. Then to check if $X \perp\!\!\!\perp Y$ we need to verify if $p_{X,Y} = p_X \otimes p_Y$.

If we want to create an independence test from an f-divergence, say $f(x) = xlog(x)$, then we can do as follows:

$$I(X;Y) := D_f(p_{X,Y} \parallel p_X \otimes p_Y)$$

This in fact the well known Mutual Information from Information Theory!

Recall that for the MMD we need to provide a kernel, since we are now in a product space, we need to provide a product kernel on the space $(\mathcal{X}, \mathcal{Y})$:

$$\begin{aligned} \mathcal{X} \times \mathcal{Y} \quad &\to \qquad \mathbb{R} \\ ((x,y)(\tilde{x},\tilde{y})) &\mapsto k(x,\tilde{x}) \cdot l(y,\tilde{y}) \end{aligned}$$

where k and l are kernels on $\mathcal{X}$ and $\mathcal{Y}$ respectively. We can then define the MMD to test independence as follows:

$$\mathrm{MMD}\left(p_{X,Y}, p_X \otimes p_Y\right)$$

Observe that

$$\mathrm{MMD}\left(p_{X,Y}, p_X \otimes p_Y\right)^2 = \mathrm{HSIC}(p_{X,Y})$$

So the HSIC is bassically an MMD distance between the joint and product distribution.

Proposed methods Part II

4 First principle methods

4.1 The twin test

Suppose we are given samples $\mathcal{D} = \{x_i, y_i\}_{i\in[n]}$ from an ANM $X \to Y$, which recall has the form

$$\begin{cases} Y = f(X) + Z \\ X \perp\!\!\!\perp Z, \quad X \sim p_X, \quad Z \sim p_Z \end{cases}$$

The main strategy of most causal inference methods has been to estimate f, and then to compute the estimated residual $\hat{e} = \hat{f}(x) - y$; the final step is to test the independence between $\hat{e}$ and X. This exploits the assumption that $X \perp\!\!\!\perp Z$. In practice we often have that the noise is *independent*, i.e. we produce a sequence $Y_1, ..., Y_n$, where $Z_i \perp\!\!\!\perp Z_j \ \forall i \neq j$.

By directly exploiting the IID noise assumption, we will circumvent the need for an independence test. The idea is to to partition the data – for simplicty you can think about splitting it around the median; we then estimate the residuals for each partition, and we then test if the IID noise assumption holds by comparing the residuals of each partition. We can apply this procedure to both directions and then we will call the direction causal if it's residuals are more similar.

We will explain this idea in more detail by following an example: we are given samples $\mathcal{D} = \{x_i, y_i\}_{i\in[n]}$ from an ANM $X \to Y$; We first visualise the data $\{x_i, y_i\}_{i\in[n]}$, by plotting X to Y, and viceversa (see Figure 4.1).

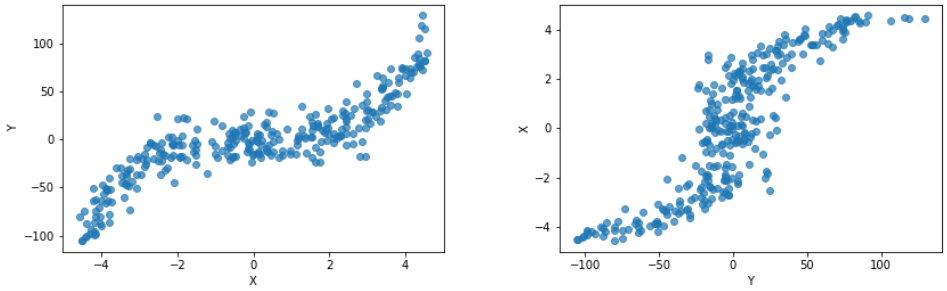
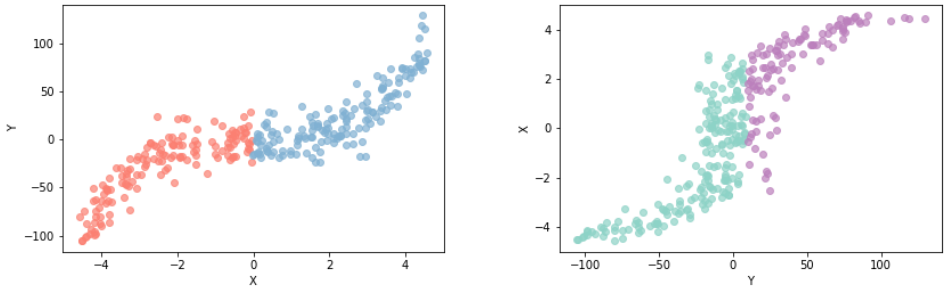


Figure 4.1 – 300 samples of data X, Y. With **ANM**:
$f(x) = \tanh(x) + 2\sin(2x) + x^3$
$X \sim \mathcal{U}_{[-a,a]}$ and $Z \sim \mathcal{N}(0, \sigma^2)$

For simplicty assume $X \sim \mathcal{U}_{[-a,a]}$ (i.e. X is uniformly disributed), then we can split the data in two, say D_1 and D_2, where we place all samples with $x_i < 0$ into D_1, and the rest into D_2. To be more precise, $\mathcal{D}_1 = \{(x_i, y_i) : x_i < 0\}$ and $\mathcal{D}_2 = \mathcal{D} \backslash \mathcal{D}_1$. We also do the same procedure for the reverse set up, i.e. we reverse the roles of x and y , $\tilde{\mathcal{D}} = \{y_i, x_i\}_{i \in [n]}$ and by the same procedure we obtain $\tilde{\mathcal{D}}_1$ and $\tilde{\mathcal{D}}_2$. We can visualise this parition bellow in Figure 4.2.



Partitions D_1 and D_2 Partitions $\tilde{\mathcal{D}}_1$ and $\tilde{\mathcal{D}}_2$

Figure 4.2 – We highlight each partition in a different color. On the left we have D_1 and D_2; and on the right we have $\tilde{\mathcal{D}}_1$ and $\tilde{\mathcal{D}}_2$

If we estimate a fit $\hat{f}_1$ for $\mathcal{D}_1$ and similarly $\hat{f}_2$ for $\mathcal{D}_2$, then we can compute residuals for each sets, say $\hat{e}_1$ for $\mathcal{D}_1$ and $\hat{e}_2$ for $\mathcal{D}_2$. Since the noise is – not only independent from X but also – iid, it follows that $\hat{e}_1$ and $\hat{e}_2$ follow the same distribution – assuming a perfect fit f. We can viualize this by looking at the histograms from the residuals.

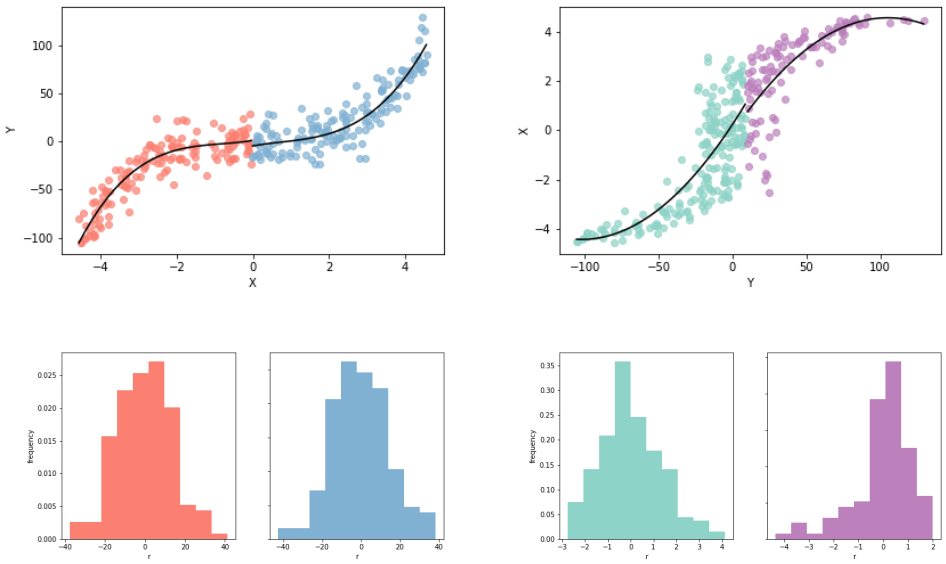


Figure 4.3 – We show the the estimated fits $\hat{f}$ for each parition in black. Bellow each partition we plot the histograms of the residuals – in the same color.

Note that for the reverse model, the noise in $\tilde{\mathcal{D}}_1$ appears to be very different from that of $\tilde{\mathcal{D}}_2$; this is not a coincidence – intuitvely, it seems very unlikely that regressing in the other direction will also result in independence noise. Further, as we briefly mentioned in the early chapters, as Hoyer et al. (2009) show, it is unlikely that for a non-linear we might not have identifiability.

One simple idea is then to quantify these observations; from $\mathcal{D}_1$ and $\mathcal{D}_2$ we compute $\hat{e}_1$, $\hat{e}_1$, and so we can define as a score for these sets:

$$\mathcal{C}(\mathcal{D}_1, \mathcal{D}_2) = \left\| p_1 - p_2 \right\|_1$$

Where p_1 is the empirical distribution of $\hat{e}_1$, and similarly for p_2 and $\hat{e}_2$. We can then apply the score function to $\tilde{\mathcal{D}}_1$ and $\tilde{\mathcal{D}}_2$ and infer causality as follows:

$$\begin{cases} X \to Y & \mathcal{C}(\mathcal{D}_1, \mathcal{D}_2) \leq \mathcal{C}(\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2) \\ Y \to X & \text{otherwise} \end{cases}$$

In the above example, we get that $\mathcal{C}(\mathcal{D}_1, \mathcal{D}_2) = 0.138$ and that $\mathcal{C}(\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2) = 0.480$ where use bins of size 5 for discretization; We are able to predict the causal direction with high confidence.

Assuming the regressions are **faithfull**, then as $n \to \infty$ we know that both p_1 and p_2 will

converge to the same p_Z and so $\mathcal{C}(\mathcal{D}_1, \mathcal{D}_2) \to 0$. On the other hand, it is unlikely that the residuals of $\tilde{\mathcal{D}}_1$ and $\tilde{\mathcal{D}}_2$ follow the same distribution (due to the non-linearlities introduced by f and the additivy of the noise) and so we can be pretty confident that asympotically the procedure will correct. In fact, assuming that **ANM** $X \to Y$ is identifiabile will be enough to show that this procedure is consistent.

In essence the algorithm consists of the parts:

1. Partition the data
2. Estimate regressions and residuals
3. Compute scores

Before giving the general description of the algorithm we will comment on each of these parts in more detail.

4.1.1 Partition

Say that we partition $\mathcal{D}$ into $\mathcal{D}_1, ..., \mathcal{D}_k$; then these partitions need to satisfie three requirements:

1. The partitions need to be **dense**: $|\mathcal{D}_i| \geq \rho|\mathcal{D}|, \quad \forall i \in [k]$, for some $\rho \in (0,1)$.
2. The partitions need to be disjoint
$$\mathcal{D} = \bigcup_i \mathcal{D}_i \quad \text{and} \quad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \quad \forall i \neq j$$
3. We need to be able to order the paritions, say $\mathcal{D}_1, ..., \mathcal{D}_k$, such that if $i < j$ then[1]:
$$\max\{x : (x, *) \in \mathcal{D}_i\} \leq \min\{x : (x, *) \in \mathcal{D}_j\}$$

The first condition – that of dense partitions – is to avoid getting trivial large deviations between residuals in the subsets; the second reason is that if they are dense, then we can give asymptotic guarantees about each subset. The second and third conditions simply ensure that we are not mixing data and that it is coherent to make regression in each subset.

If we use K-means (perhaps the most popular clustering algorithm), then conditions 2. and 3. are met. The only questions is in regards to conditions 1. K-means starts by randomly initialisting two or n centers (depending on the number of clusters that we want), and the updates the centers that they locally minimizes within-cluster variances. If our data is infinite support,

[1] The $*$ is to indicate a dummy variable, as we do not care for the value of y.

and we re-run K-means if there is some cluster i s.t. $|\mathcal{D}_i| < \rho|\mathcal{D}|$; then if we have enough data and for some ρ we can be quite certain that the algorithm will eventually terminate.

In practice this has always been the case (for $\rho = .3$); so we conjecture that one can prove the above statement rigorously.

The last question is, "How many clusters do we want?". Obviously for the small data regime we must be content wuth only two clusers; but what if we have a lot of data? As we will see, we observe experimentally that if we we choose the number of partition as an increasing function w.r.t. sample size, then we can get better accuracies.

One reason that more partitions are desirable is that then the regression problem becomes easier; Indeed, as we zoom into a function it tends to become smoother.

4.1.2 Regression

A benefit of partitioning the is that we are also partitioning the function we are trying to estimate; in particular one would expect that the regression will be easier, e.g. a low order polynomial might be enough.

We do elementary model selection, we take the best model with BIC score for degrees 1 to 6.

Show example of many partitions

4.1.3 Score functions

We have seen in previous chapters various ways to measure the distances between two distribtions say p_1 and p_2, via some score function $\mathcal{D}(p_1, p_2)$; for example $\mathcal{D}$ could be the MMD metric, or the l_1 distance.

Now instead we have a set of distributions, say $P_k = p_1, ..., p_k$, and we wish to see how homogenous P_k is compared to some other set $\tilde{P}_j$ – recall that we wish to see in which of the two, the distribtions are more likely to be the same, i.e. we are testing the iid assumption.

There are several simple ways to go about this:

$$C(P_k) = \max_{i,j} \mathcal{D}(p_i, p_j)$$

Another option is to take an average of the pairwise score:

$$C(P_k) = \frac{1}{\binom{k}{2}} \sum_{i<j} \mathcal{D}(p_i, p_j)$$

or even

$$C(P_k) = \frac{1}{\binom{k}{2}} \sum_{i<j} \mathcal{D}(p_i, p_\mu), \quad p_\mu = \frac{1}{k} \sum_i p_i$$

Thus if $C(P_k) < C(\tilde{P}_j)$, we can say that the distributions in P_k are more homogenous; e.g. they more likely to produce similar looking noise.

We have tested all of the above and find that the first method – using the maximum score between pairs – gives the best performance.

4.1.4 Algorithm

We note that the algorithm is a general framework as we are free to choose the partition, regression method and score function. We would like to point out that we recycle the data, i.e. we use the same data for estimating the regression and subsequent score function; obviously if one wishes one can easily split the data in the paritions to use a different portion of the data for estimation and for evaluating the score function.

Algorithm 2 Twin method: General procedure to decide whether $p_{x,y}$ satisfies and ANM $X \to Y$ or $Y \to X$

Input:

1. I.i.d samples $\mathcal{D} = \{(x_i, y_i)\}_{i\in[N]}$ of X and Y
2. Partition procedure
3. Regression method
4. Score estimator $\hat{C} : R^{***} \to \mathbb{R}$, where E is a set of vectors.

Output: $\hat{C}_{X\to Y}$, $\hat{C}_{Y\to X}$, dir

1. $\tilde{\mathcal{D}} := \{(y_i, x_i)\}_{i\in[N]}$
2. **Partition** the data into subsets[2]:
   - $\{\mathcal{D}_i\}_{i\in[k]}$ s.t. $\mathcal{D}_i \subset \mathcal{D}, \forall\, i \in [k]$
   - $\{\tilde{\mathcal{D}}_i\}_{i\in[j]}$ s.t. $\tilde{\mathcal{D}}_i \subset \tilde{\mathcal{D}}, \forall\, i \in [j]$
   - Where integers $j, k > 1$ are determined by the partition procedure.
3. **Estimate regressions** and residuals for each subset

   for $i \in [k]$:

   - Let $\mathbf{x}$, $\mathbf{y}$ be the vectors formed from $\mathcal{D}_i$
   - $\hat{f}_Y$ of the regression function $x \mapsto \mathbb{E}(Y|X = x)$
   - $\hat{\mathbf{e}}_\mathbf{Y}(i) := \mathbf{y} - \hat{f}_Y(\mathbf{x})$

   end for

   $\mathbf{E}_\mathbf{Y} := \{\hat{\mathbf{e}}_\mathbf{Y}(i)\}_{i\in[k]}$

   for $i \in [j]$:

   - Let $\mathbf{x}$, $\mathbf{y}$ be the vectors formed from $\tilde{\mathcal{D}}_i$
   - $\hat{f}_X$ of the regression function $x \mapsto \mathbb{E}(X|Y = y)$
   - $\hat{\mathbf{e}}_\mathbf{X}(i) := \mathbf{x} - \hat{f}_X(\mathbf{y})$

   end for

   $\mathbf{E}_\mathbf{X} := \{\hat{\mathbf{e}}_\mathbf{X}(i)\}_{i\in[j]}$
4. **Compute scores** to measure the difference between the residuals
   - $\hat{C}_{X\to Y} := \hat{C}(\mathbf{E}_\mathbf{Y})$
   - $\hat{C}_{Y\to X} := \hat{C}(\mathbf{E}_\mathbf{X})$
5. Output $\hat{C}_{X\to Y}$, $\hat{C}_{Y\to X}$, and

$$\text{dir} := \begin{cases} X \to Y & \text{if } \hat{C}_{X\to Y} \le \hat{C}_{Y\to X} \\ Y \to X & \text{otherwise} \end{cases}$$

4.1.5 Consistency

We will show consistency for a simple set up of the twin test – but we note that generalising it to the more general set up should be direct from our proof.

The setup was the linear **ANM**:

$$\begin{cases} Y = f(X) + Z \\ X \perp\!\!\!\perp Z,\ X \sim p_x,\ Z \sim p_Z \end{cases}$$

In practice we are given samples $\mathcal{D} = \{x_i, y_i\}_{i\in[n]}$ from an ANM $X \to Y$; next the algorithm will proceed to split the data into sets $\mathcal{D}_1$ and $\mathcal{D}_2$. It then proceeds to compute residuals and compute some scores between them.

To simplify the proof, we will **skipe the partition procedure** and assume that we are directly given $\mathcal{D}_1$ and $\mathcal{D}_2$, each with n samples (Note that, if p_x is uniform, then having these sets be of equal size would happen exponentially fast; indeed, in general, we will have dense partitions).

Next, we will assume that on each interval, the data is linear with slope a_1 and a_2 resp.

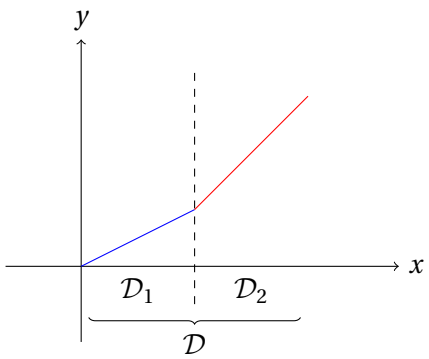


Figure 4.4 – Slopes a_1 and a_2 in blue and red respectively.

Thus our problem can be seen as getting data from two difference **ANM**, both with identical noise, but with a different truncation of X:

$\mathcal{D}_1$ is sampled from

$$\begin{cases} Y_1 = a_1 X_1 + Z \\ X_1 \perp\!\!\!\perp Z,\ X_1 \sim p_{X_1},\ Z \sim p_Z \end{cases}$$

and $\mathcal{D}_2$ is sampled from

$$\begin{cases} Y_2 = a_2 X_2 + Z \\ X_2 \perp\!\!\!\perp Z,\ X_2 \sim p_{X_2},\ Z \sim p_Z \end{cases}$$

Note that without loss of generality we can assume $X_1 \sim X_2 \sim X \sim P_X$.

We will call this scenario the **simplified Twin Test scenario**.

We next describe the steps of the algorithm after partitioning:

We first split $\mathcal{D}_1$ in two sets of equal size $\mathcal{D}_1^{train}$ and $\mathcal{D}_1^{test}$. We first use $\mathcal{D}_1^{train}$ to estimate a_1, say $\hat{a}_1$ via regression. Then, using $\mathcal{D}_1^{test}$, we estimate the residual:

$$\hat{Z}_1 = Y_1 - \hat{a}_1 X$$

We then discretise $\hat{Z}_1$ and form a distribution say $\hat{p}_1$; we can do the same thing for $\mathcal{D}_2$, and by doing the same procedure obtain $\hat{p}_2$.

We discretise both with a fixed step size, say s; as we will see, the only requirement is that we fix the size beforehand.

We use the l_1 distance as our score:

$$\hat{C}_{X\to Y} = \left\| \hat{p}_1 - \hat{p}_2 \right\|_1$$

As we have seen, $\hat{C}_{Y\to X} > 0$ holds in general except in very particular situations. So, to prove that our algorithm is consistent, we need to show that:

$$\hat{C}_{X\to Y} \to 0 \qquad \text{as} \qquad n \to \infty$$

This is precisely what we will show:

Theorem 4 *The **simplified Twin Test scenario**t is consistent, i.e.*

$$\hat{C}_{X\to Y} \to 0 \qquad as \qquad n \to \infty$$

The idea of the proof is to observe the following:

If we have enough data, i.e. when n is large enough then assuming the regression is faithfull, we can chose any α such that

$$|a_1 - \hat{a}_1| \leq \alpha \qquad \text{and} \qquad |a_2 - \hat{a}_2| \leq \alpha$$

This means that

$$\hat{Z}_1 = Z + (a_1 - \hat{a}_1)X \quad \Longrightarrow \quad Z - \alpha X \leq \hat{Z}_1 \leq Z + \alpha X$$

and similarly

$$\hat{Z}_2 = Z + (a_2 - \hat{a}_2)X \quad \Longrightarrow \quad Z - \alpha X \leq \hat{Z}_2 \leq Z + \alpha X$$

Note that $\hat{Z}_1 \sim P_{Z+\Delta_1 X}$ and $\hat{Z}_2 \sim P_{Z+\Delta_2 X}$, where for brevity we denote $\Delta_1 = a_1 - \hat{a}_1$ and $\Delta_2 = a_2 - \hat{a}_2$. We can visualise the distance between these distribtions as follows: (see[3] figure 4.5)

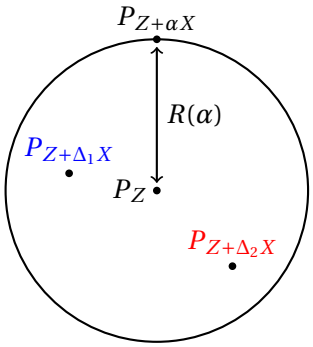


Figure 4.5 – The l_1 "ball" around P_Z, for brevity we denote $\Delta_1 = a_1 - \hat{a}_1$ and $\Delta_2 = a_2 - \hat{a}_2$

The idea is then the following, given some $\epsilon > 0$, we want to show that asympotically

$$\left\| \hat{p}_1 - \hat{p}_2 \right\|_1 > \epsilon$$

cannot happen. The game plan will be to find a bound on the $R(\alpha)$; once we have one, we are done, we will simply pick an α s.t. $R(\alpha) < \epsilon$. Recall that α is the error in of our $\hat{a}$ estimate, which we can get arbitrarily small with enough samples.

We begin by proving lemmas to find bounds for the radius $R(\alpha)$.

Lemma 4 *Given $\alpha, \delta > 0$, random variables Z and X s.t. $\alpha X + Z$, where $X \perp\!\!\!\perp Z$, $X \sim P_X$, $Z \sim P_Z$, with $\int P'_Z(t)\,dt < \infty$. $\alpha X + Z \sim P_{\alpha X + Z}$, there is some $C > 0$ s.t.*

[3] Note that the illustration does not follow the actual gemoetry of the space, we draw it solely to gain intuition about the problem.

$$\|P_Z - P_{\alpha X+Z}\|_1 \leqslant \alpha C + \delta$$

Proof:

First note that $\alpha X \sim \frac{1}{\alpha} P_X\left(\frac{\tau}{\alpha}\right)$ by applying the change of variable rule.

Next, since $X \perp\!\!\!\perp Z$, we may write $P_{\alpha X+Z}$ as a convolution:

$$P_{\alpha X+Z}(t) = \int P_Z(t-x)\frac{1}{\alpha}P_X\left(\frac{x}{\alpha}\right)dx = \int P_Z(t-\alpha x)P_X(x)dx$$

Let $T^{\alpha X}P_Z(t) := P_Z(t-\alpha x)$ (we use the notation introduced by Lagrange for the shift operator).

Hence we may write (with P_X as the underlying measure):

$$P_{\alpha X+Z}(t) = \mathbb{E}\left[P_Z(t-\alpha X)\right]$$

We proceed as follows:

$$\begin{aligned}
\|P_Z - P_{\alpha X+Z}\|_1 &= \int |P_Z(t) - \mathbb{E}\left[P_Z(t-\alpha X)\right]|dt \\
&\leq \int \mathbb{E}|P_Z(t) - P_Z(t-\alpha X)|dt \\
&= \mathbb{E}\int |P_Z(t) - P_Z(t-\alpha X)|dt \\
&= \mathbb{E}\left[\left\|P_Z - T^{\alpha X}P_Z\right\|_1 \,||X| \leq k\right] P\left(|X| \leq k\right) + \mathbb{E}\left[\left\|P_Z - T^{\alpha X}P_Z\right\|_1 \,||X| > k\right] P\left(|X| > k\right) \\
&\leq \mathbb{E}\left[\left\|P_Z - T^{\alpha X}P_Z\right\|_1 \,||X| \leq k\right] + 2P(|X| > k) \\
&\leq \int \left|P_Z(t) - P_Z(t-\alpha C^*)\right|dt + 2P(|X| > k) \\
&\leq \int \left|P_Z(t) - (P_Z(t) + \alpha C^* P_Z'(t))\right|dt + 2P(|X| > k) \\
&\leq \alpha C^* \int P_Z'(t)dt + 2P(|X| > k) \\
&\leq \alpha C + \delta
\end{aligned}$$

The first equality follows by the aforementioned observation. The first inequality follows from the triangle inequality; the equality that comes after is due to Fubini's theorem, we can swap the expectation (which is also an integration) since all measures are measurable. We next use the law of total probability by splitting the expectation w.r.t some k to be chosen later.

The next upper bounds follows by noting that $\|p-q\|_1 \leq 2$ for any distribtions p and q. Next observe that $\mathbb{E}\left[\left\|P_Z - T^{\alpha X} P_Z\right\|_1 \mid |X| \leq k\right]$ is the average l_1 distance between $\|P_Z\|$ and random shifts of itself. Hence if we choose the shift as follows:

$$C^* = \text{argmax}_{s\in[-k,k]} = \int |P_Z(t) - P_Z(t-\alpha s)| dt$$

we can upper bound the average distance by the largest l_1 distance between any shift; note that there is at least one maximiser since we are optimising over a compact set.

TODO – Taylor approx bound –

We next approximate $P_Z(t-\alpha C^*)$ by using it's taylor expansions – this is justified by the fact that we will pick $\alpha \approx 0$.

Using taylor expansions we get that:

$$P_Z(t-\alpha C^*) = P_Z(t) + \sum_{n\geq 1} \left(\alpha C^*\right)^n P_Z(t)^{(n)}$$

Note that in our original equation this means that we need to bound

$$\sum_{n\geq 1} \left(\alpha C^*\right)^n \int P_Z(t)^{(n)} dt$$

Note that we can pick any α, indeed we will later make it smaller than ϵ. But I am not sure how to bound the derivatives of the pdf...

One easy fix is to assume

$$\int P_Z(t)^{(n)} dt < K, \quad \forall n \geq 1$$

for some constant K. This is true for example for exponential / unfirom distribtions.

It would be nice to not have such an assumption though, maybe something like sub-gaussianity? Really not sure about this part.

TODO – Taylor approx bound end –

Finally, since we can pick k, we choose it large enough s.t. $2P(|X| > k) \leq \delta$, we then let $C = C^* \int P'_Z(t) dt$ and we are done.

We note that while C^* will depend on δ, this is not a problem. We will first pick δ to be some small constant, this will then determine C^* to be some other constant; after this we will be

choose α. □

So we have found a bound on the l_1 distance between two continuos distributions; however in our application, these will be quantised versions of these distributions. The following lemma tells us that this is not a problem, the l_1 distance between the quantised version cannot be bigger than that of their continuos counter parts. The only requirement is that we fix the quantisation scheme beforehand and use the same one for both.

Lemma 5 *Let P and Q be two continuos distribtions, then let P^*, Q^* resp. be discretized versions. Then if $\|P-Q\|_1 \leq \delta$ for some δ, then*

$$\left\|P^* - Q^*\right\|_1 \leq \delta$$

Proof: We first quantize $\mathbb{R}$ in bins of length w, say $I_i = [wi, w(i+1))$, note $\bigcup_{i\in\mathbb{Z}} I_i = \mathbb{R}$.

Given continuos distribtions P and Q, we form their quantised counter parts as follows:

$$P^*(k) := \sum_{i\in\mathbb{Z}} \int_{I_i} p(t)dt\, \mathbb{1}_{k\in I_i}, \quad Q^*(k) := \sum_{i\in\mathbb{Z}} \int_{I_i} q(t)dt\, \mathbb{1}_{k\in I_i}$$

We then conclude as follows by a applying the triangle inequality twice:

$$\begin{aligned}
\left\|P^* - Q^*\right\|_1^2 &= \sum_{k\in\mathbb{Z}} \left|P^*(k) - Q^*(k)\right| \\
&\leq \sum_{k\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} \left|\int_{I_i} \big(p(t)-q(t)\big)\,dt\right| \mathbb{1}_{k\in I_i} \\
&\leq \sum_{k\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} \int_{I_i} \left|p(t)-q(t)\right| dt\, \mathbb{1}_{k\in I_i} \\
&= \sum_{i\in\mathbb{Z}} \int_{I_i} \left|p(t)-q(t)\right| dt \sum_{k\in\mathbb{Z}} \mathbb{1}_{k\in I_i} \\
&= \sum_{i\in\mathbb{Z}} \int_{I_i} \left|p(t)-q(t)\right| dt \\
&= \int_{\mathbb{R}} \left|p(t)-q(t)\right| dt \\
&= \|P-Q\|_1^2
\end{aligned}$$

Note that $\sum_{k\in\mathbb{Z}} \mathbb{1}_{k\in I_i} = 1$ for any i and k; since I_i is a disjoint partition k will always belong to one of them, and one alone. □

We can now conclude by proving consistency, recall that we want to show that:

$$P\left(\left\|\hat{p}_1 - \hat{p}_2\right\|_1 > \epsilon\right) \to 0$$

Proof:

$$\begin{aligned} P\left(\left\|\hat{p}_1 - \hat{p}_2\right\|_1 \geq \epsilon\right) &\leq P\left(\left\|\hat{p}_1 - p_Z\right\|_1 + \left\|\hat{p}_2 - p_n\right\|_1 \geq \epsilon\right) \\ &\leq P\left(\left\|\hat{p}_1 - p_Z\right\|_1 + \left\|\hat{p}_2 - p_Z\right\|_1 \geq \epsilon \,|\, |a_1 - \hat{a}_1| \leq \alpha\,, |a_2 - \hat{a}_2| \leq \alpha\right) \\ &\quad + P\left(|a_1 - \hat{a}_1| > \alpha \text{ or } |a_2 - \hat{a}_2| > \alpha\right) \end{aligned}$$

The first inequality follows by the triangle inequality, and the second one by using the law of total probability.

Observe that

$$P\left(|a_1 - \hat{a}_1| > \alpha \text{ or } |a_2 - \hat{a}_2| > \alpha\right) = P\left(|a_1 - \hat{a}_1| > \alpha\right) + P\left(|a_2 - \hat{a}_2| > \alpha\right)$$

Both of which go to zero for any $\alpha > 0$ assuming that we have a suitable regression.

It remains to bound

$$P\left(\left\|\hat{p}_1 - p_Z\right\|_1 + \left\|\hat{p}_2 - p_Z\right\|_1 \geq \epsilon \,|\, |a_1 - \hat{a}_1| \leq \alpha\,, |a_2 - \hat{a}_2| \leq \alpha\right) \tag{4.1}$$

Note that $\hat{p}_1 \to P^*_{Z+\Delta_1 X}$ as $n \to \infty$; where $P^*_{Z+\Delta_1 X}$ is the discretized distribution of $P_{Z+\Delta_1 X}$ – recall that we form $\hat{p}_1$ by creating a discretized histogram from the residuals.

Thus if we are given some $\epsilon > 0$, we first pick $\delta = \frac{\epsilon}{4}$, and $\alpha > \frac{\epsilon}{4C}$. We do the same for $\hat{p}_2$, which implies by lemma 4 and 5 that:

$$\left\|\hat{p}_1 - p_Z\right\|_1 + \left\|\hat{p}_2 - p_Z\right\|_1 < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

And so as $n \to \infty$

$$P\left(\left\|\hat{p}_1 - p_Z\right\|_1 + \left\|\hat{p}_2 - p_Z\right\|_1 \geq \epsilon \,|\, |a_1 - \hat{a}_1| \leq \alpha\,, |a_2 - \hat{a}_2| \leq \alpha\right) \to 0$$

□

4.2 The residual method

4.2.1 Introduction

The residuals method is linked to the twin test. Here we assume we know the noise, etc.

4.2.2 Proof of consistency: A tale of two bounds

The setup was the linear **ANM**:

$$\begin{cases} Y = aX + E_Y \\ X \perp\!\!\!\perp E_Y,\ X \sim p_x,\ E_Y \sim p_{E_Y} \end{cases}$$

From n samples (X_i, Y_i) we estimate $\hat{f}_Y$ by regressing X on Y and $\hat{f}_X$ for the reverse model. We then comput the residuals

$$\hat{e}_Y = Y - \hat{f}_Y(X) \tag{4.2}$$
$$\hat{e}_X = X - \hat{f}_X(Y) \tag{4.3}$$

We note that for the ease of analysis, it would first be wise to use some fraction of the data to first estimate the regression, and then use the remanining for the test.

For n large enough we have that

$$\hat{e}_Y \approx E_Y \sim P_{E_Y}$$

The idea is then to first discretise[4] P_{E_Y} into m bins, call this discrete distribution Q. We apply the same discretization to obtain $B = (b_1, ..., b_m)$ from $\hat{e}_Y$ and $\tilde{B} = (\tilde{b}_1, ..., \tilde{b}_m)$ from $\hat{e}_X$.

We then decide the causal direction as follows

$$\begin{cases} X \to Y & \text{if} \quad C \leq W \\ Y \to X & \text{if} \quad C > W \end{cases}$$

Where

[4] We do so in a naive manner we split it uniformly into m bins.

$$C = \|B - U\|_{L_1}$$
$$W = \left\|\tilde{B} - U\right\|_{L_1}$$

s.t. $U = (\frac{1}{m}, ..., \frac{1}{m})$.

Given our assumption about the **ANM**, the probability to outpout the correct causal direction is:

$$P_{\text{correct}} = \mathbb{P}[C \leq W]$$

We next upper bound thise quantity in order to show consistency

$$\mathbb{P}[C \leq W] \geq \mathbb{P}\left[\bigcup_{\tau \in \mathbb{Q}} C \leq \tau \cap W > \tau\right] \tag{4.4}$$
$$\geq \mathbb{P}[C \leq \tau \cap W > \tau] \tag{4.5}$$
$$\geq \mathbb{P}[C \leq \tau] - \mathbb{P}[W \leq \tau] \tag{4.6}$$

The first inequality is due to the fact that we are only taking the union in the rationals[5]. The second inequality is done by looking at the probability of a fixed τ; and the final one follows by:

$$1 \geq \mathbb{P}[C \leq \tau \cup W > \tau] = \mathbb{P}[C \leq \tau] + \mathbb{P}[W > \tau] - \mathbb{P}[C \leq \tau \cap W > \tau]$$

We will next find appropriate bounds for $\mathbb{P}[C \leq \tau]$ and $\mathbb{P}[W \leq \tau]$.

4.2.3 Bounding the false false postive

We will first lower bound $\mathbb{P}[C \leq \tau]$ by upper bounding the complement event.

[5]We note that we can only take unions over countable sets; recall also that the rationals are dense in the irrationals, so the inequality is very close to equality (and in practice and among friends it would be).

$$\mathbb{P}[C \geq \tau] = \mathbb{P}\left[\sum_{i=1}^{m}\left|b_i - \frac{1}{m}\right| \geq \tau\right] \tag{4.7}$$

$$\leq \mathbb{P}\left[m \max_i \left|b_i - \frac{1}{m}\right| \geq \tau\right] \tag{4.8}$$

$$= \mathbb{P}\left[\bigcup_i \left|b_i - \frac{1}{m}\right| \geq \frac{\tau}{m}\right] \tag{4.9}$$

$$\leq m\,\mathbb{P}\left[\left|b_0 - \frac{1}{m}\right| \geq \frac{\tau}{m}\right] \tag{4.10}$$

$$\leq m2\exp\left(-2n\frac{\tau^2}{m^2}\right) \tag{4.11}$$

The second to last inequality follows by the union bound and by noting that all b_is are the same since they are discretized empirical distribution coming from a uniform source. For the final inequality we use Hoeffding's inequality.

4.2.4 Bounding the false negatives

Recall that what is left to bound is the following quanity, $\mathbb{P}[W \leq \tau]$; for this we first define the following set of probability distributions:

$$\Gamma_\tau = \{\pi \in \Delta_m : \|\pi - U\|_{L_1} \leq \tau\}$$

Where the Δ_m is the m dimensional simple and U the uniform vector as before.

Observe that:

$$\{W \leq \tau\} = \{\tilde{B} \in \Gamma_\tau\}$$

In essence, we are asking: "what is the chance that the realisation of $\tilde{B}$ – which is the empricial distribution of some distribution Q – lies inside some set of distribtions Γ_τ.

We note that bounding this kind of event is exactly what Sanov's theorem[6] gives us, an important result from large deviation theory that also exploits concentration of measure.

Let $\mathbf{x} = (x_1, ..., x_n)$ be a sequence of n each drawn independently from a finite universe U with $|U| = m$. Denote by $P_\mathbf{x}$ the empirical distribution – or type – for a given sequence $\mathbf{x}$. Let Q^n be the product distribution n independent samples of Q.

[6] See the section on Information Theory and statistics in Cover (1999)

Theorem 5 (Sanov's theorem) *Let* Π *be a set of distributions on* U*, and* $m = |U|$*. Let*

$$P^* = \operatorname{argmin}_{P \in \Pi} D(P \| Q)$$

Then

$$\mathop{\mathbb{P}}_{Q^n}[P_{\mathbf{x}} \in \Pi] \leq (n+1)^m 2^{-nD(P^* \| Q)}$$

Applying the above theorem, and noting that Γ_τ takes the place of Π, $\tilde{B}$ that of $P_{\mathbf{x}}$ and the discretized distribution $\hat{e}_X = X - \hat{f}_X(Y)$ that of Q we get:

$$\mathbb{P}[W \leq \tau] = \mathbb{P}\left[\tilde{B} \in \Gamma_\tau\right] \leq (n+1)^m 2^{-nD(\tau)} \tag{4.12}$$

Where $D(\tau) := D(P^* \| Q)$, we make the τ relation excplicit to keep in mind that the minimisation is contrainred to the set Γ_τ which depends on τ.

We remark that the only place of concern is if $D(P^* \| Q) = 0$; assuming however that $Q \neq U$, then there will be some τ s.t. $Q \notin \Gamma_\tau$ and thus $D(P^* \| Q) \neq 0$.

We can now conclude by putting everything together; recall that we had shown that we could bound the success probability as follows:

$$\mathbb{P}[C \leq W] \geq \mathbb{P}[C \leq \tau] - \mathbb{P}[W \leq \tau] \tag{4.13}$$

$$\geq 1 - 2m \exp\left(-2n\frac{\tau^2}{m^2}\right) - (n+1)^m 2^{-nD(\tau)} \tag{4.14}$$

This, if we fix m, and if there exists some τ s.t. $D(\tau) > 0$ then we get consistency by letting $n \to \infty$.

We note that to get the best bound we may maximisise the r.h.s. w.r.t. τ.

5 Experiments

5.1 Benchmark

We benchmark on five bivariate cause-effect datasets[1] , covering a wide range of associations:

1. **Cha** (300 cause-effect pairs) pairs from the challenge of Guyon (2013)
2. **Net** (300 cause-effect pairs) artificial cause-effect pairs generated using random distributions as causes, and neural networks as causal mechanisms
3. **Gauss** (300 cause-effect pairs) generated by Mooij et al. (2016), using random mixtures of Gaussians as causes, and Gaussian process priors as causal mechanisms.
4. **Multi** (300 cause-effect pairs) built with random linear and polynomial causal mechanisms (by Goudet et al. (2017)). In this dataset, additive or multiplicative noise, is applied before or after the causal mechanism.
5. **TCEP** (108 cause-effect pairs)[2] is the Tübingen Cause Effect Pair data set which consists of various domains such as climatology, finance, and medicine (Mooij et al. (2016)).

Cite competitors. (as done in GCNN)

The first is via the Area Under the Precision Recall curve, and the second only checks at accruacy.

[1] The TCEP dataset can be found here and all the other datasets can be found here
[2] Note that 6 of these pairs are not bivariate.

method	Cha	Net	Gauss	Multi	TCEP
Best fit	56.4	77.6	36.3	55.4	58.4 (44.9)
LiNGAM	54.3	43.7	66.5	59.3	39.7 (44.3)
CDS	55.4	89.5	84.3	37.2	59.8 (65.5)
IGCI	54.4	54.7	33.2	80.7	60.7 (62.6)
ANM	66.3	85.1	88.9	35.5	53.7 (59.5)
PNL	73.1	75.5	83.0	49.0	68.1 (66.2)
Jarfo	79.5	92.7	85.3	94.6	54.5 (59.5)
GPI	67.4	88.4	89.1	65.8	66.4 (62.6)
CGNN $\left(\widehat{\text{MMD}}_k\right)$	73.6	89.6	82.9	96.6	79.8 (74.4)
CGNN $\left(\widehat{\text{MMD}}_k^m\right)$	76.5	87.0	88.3	94.2	76.9 (72.7)
TwinTest	66.3	81.9	85.1	39.8	77.0 (82.0)

Table 5.1 – Cause-effect relations: Area Under the Precision Recall curve on 5 benchmarks for the cause-effect experiments (weighted accuracy in parenthesis for TCEP)

Table 5.1 is taken from Goudet et al. (2017)

Model	TCEP	TCEP with 75 samples
BCI	0.64	0.60
ANM-HSIC	0.63	0.54
ANM-MML	0.58	0.56
IGCI	0.66	0.62
CGNN	0.70	0.69
TwinTest	62.4	TODO

Table 5.2 – Accuracy for TCEP Benchmark

Table 5.2 is taken from Kurthen and Enßlin (2018)

6 Conclusion

Bibliography

Bartlett, P. (2008). Reproducing kernel hilbert spaces, cs281b/stat241b (spring 2008) statistical learning theory.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. (2017). Causal generative neural networks. *arXiv preprint arXiv:1711.08936*.

Gretton, A. (2019). Introduction to rkhs, and some simple kernel algorithms.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Guyon, I. (2013). Chalearn cause effect pairs challenge.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Jordan, M. (2004). Gaussian processes and reproducing kernels.

Kpotufe, S., Sgouritsa, E., Janzing, D., and Schölkopf, B. (2014). Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, pages 478–486.

Kurthen, M. and Enßlin, T. A. (2018). Bayesian causal inference. *arXiv preprint arXiv:1812.09895*.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443.

Nowzohour, C. and Bühlmann, P. (2016). Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485.

Oreskes, N. (2011). *Merchants Of Doubt : How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, New York, USA.

Pearl, J. (2000). Causality: Models, reasoning, and inference.

Peters, J. (2008). Asymmetries of time series under inverting their direction. Diploma Thesis, University of Heidelberg. http://stat.ethz.ch/people/jopeters.

Schölkopf, B. (2019). Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009). On integral probability metrics, phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. (2008). Injective hilbert space embeddings of probability measures. In *21st Annual Conference on Learning Theory (COLT 2008)*, pages 111–122. Omnipress.