First Principle Methods For Causal Discovery

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the registrar's office.

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

Wings are a constraint that makes
it possible to fly.
— Robert Bringhurst

To my parents. . .

Acknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Lausanne, July 17, 2020 D. K.

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

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 – but he has no idea of anything she might be up to.

Bob then observes the following trend:

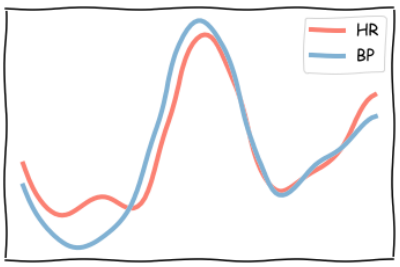


Figure 1.1 – Heart rate (HR) and Blood pressure (BP) of Alice.

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 – having neglected biology as being beneath him – and 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:

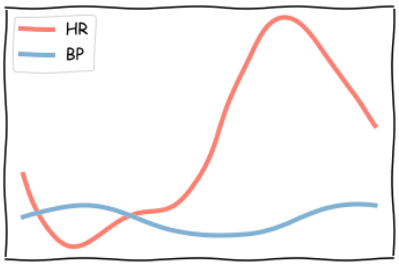


Figure 1.2 – Heart rate (HR) and Blood pressure (BP) of Alice.

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 very 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 google images.

If Bob was able to incorporate these notions into his machine learning models, then he 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 Scholkopf argues (Schölkopf (2019)).

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? In fact[5] X causes Y, and we will show algorithms that can accuretly predict causality in such settings with as few as 75 samples.

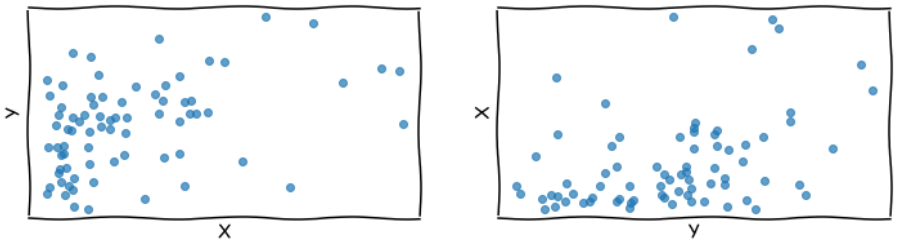


Figure 1.3 – 75 samples of data X, Y.

1.2 Causality

We have seen in the introduction that determining and understanding causality may indeed be useful; but what exactly is causality? When we say X causes Y i.e. $X \to Y$, we have an intuitive idea of what it means; but how should we formalise it?

[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)).

[5] We model $y = f(x) + n$, where x is exponential distribution, and n normaly, both independent from each other; f is a non-linear function. This kind of set up is known as an additive model and it will be a central object of study.

Suppose that we are given two random variable X, Y with joint distribution $\mathbb{P}(x, y)$. Intuitively we would say that $X \to Y$, if we intervene in X and then see an effect Y. In particular we will denote do(x) – short for 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 $\mathbb{P}_{Y|do(x)}$.

In supervised learning the goal is often to estimate $p_{y|x}$; while it is tempting to say that this is the same as $p_{y|do(x)}$, they may in fact be very different. Suppose x is 1 if a person smokes and 0 otherwise, similarly let y be 1 if a person has cancer and 0 otherwise. If we then ran immoral **randmoized trials**[6] and forced a subset of the population at random to smoke, then we could estimate $p_{y|do(x)}$. If it turned out that $p_{y|do(x)} \approxeq p_{y|x}$ then this would be very strong evidence that indeed smoking caused cancer.

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'*

Borrowing from probability, it might be tempting to say something like

$$p_{y|x}$$

From our own lives, we know that it is by intervening in systems that we infere causal directions.

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 pleasing analysis. However both are based on a first principle approach with known asymptotics in mind.

We note that in the analysis / procedure we split the data in 2, first to train the model, and then to perform the score computation - Need to expand on this

Will be easier to do once they are finished.

1.4 Outline

The thesis is divided in two parts,

background etc blup blue die blah blu

[6] The power of randmoized trials is that they average out potential confounder – e.g. a gene that causes both an inclination to smoke and cancer. This is also knowm as A/B testing and is essential to determine if a new medicine has the desired effect and that a new UI might maximise user addicton to an app.

Do once finished, but main idea

Theory and background material

Methods

Experiments

Part I Preliminaries

2 Causal Inference

2.1 Bivariate causal model

2.1.1 Introduction

2.1.2 ANM

2.1.3 Identifiability

2.2 Methods

2.2.1 LinGam

blah

NOTE talk about the regression -> residual decompo (with indep test at the end)

2.2.2 HSIC

Talk about IPM vs $f-$divergence.

To test whether two random variables are independent X, Y

2.2.3 MAX entropy

related to what we did, note that it would work if everything was discrete!

2.3 SNR and causality

In virtually most of the predictive fields, noise is the enemey; Indeed, in the absence of noise, finding the best linear fit to a linear model is trivial. Emre Telatar, a powerful information theo-

rist, liked to jest in his digital communication course that "without noise, we communication engineers would be without a job".

Indeed, for most of the early 20th century[1] noise was keeping engineers busy as they devised clever schemes to fight noise. At the time the whole business was very experimental as no one had come close to understanding noise in the context of transmission; questions such as "Is it possible to send a message with arbitrary reliability?" and "What is the theoretical maximum amount of information that can be reliably sent?", were questions that no one had come close to solving.

Then Shannon came along, in his Magnum Opus, Shannon (1948), he not only formalised the foundations of information Theory, but he also proved[2] most of the main results in it. In particular, he showed that for the AWGN (Additive white nois gaussian channel), it is possible to realiably transmit at most C bits per time unit:

$$C \propto log(1 + SNR)$$

Where SNR is the celebrated signal to noise ratio – $SNR = \frac{\mathbb{E}(X^2)}{\mathbb{E}(N^2)}$. As we would expect, if $SNR \to \infty$ then we can send an arbitrary amount of information per time unit (the only limit is the physical one, i.e. the speed of light). Conversely if $SNR = 0$ then we find ourseleve at a rave, not matter how much we yell, our friends will not be able to understand us.

If we return to the question of causality; a somehwhat trivial observation is that if the mechanism is injective[3], in the absence of noise, it is not possible to say anything about the causal direction of the mechanism. Here too noise is the benelovent giver of jobs, albeit not for the same reasons. Interestingly, we can use noise to help us deduce the causal nature of a process.

We will now see what perhaps could be considered the most simple causal set up, and describe a method for causal inference. We will then see that SNR also plays an important role.

Consider the linear additive noise model – our first causal model!

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

Suppose we are given n samples of the above process:

$$y_i = ax_i + z_i, i \in [n]$$

[1] quote comm book
[2] Shannon had a very deep...
[3] If it is not injective, then the function is not invertible, and thus only one causal direction is possible.

We collect these into vectors say y, x and z; note that we do not have access to the latter, but it will come in handy for the derivation that follows.

One common idea is to first compute the residuals for both possible regression models, and the to check which residual is less dependent on x and y respectively – we are testing for the noise / data independence hypothesis.

We first regress y on x. i.e.

$$\hat{a} = \text{argmax}_\alpha \left\| y - \alpha x \right\|_2^2$$

We differentiate w.r.t to α:

$$-2y^\intercal x + 2\alpha \left\| x \right\|_2^2 = 0 \qquad \Rightarrow \qquad \alpha = \frac{y^\intercal x}{\left\| x \right\|_2^2} = \frac{a \left\| x \right\|_2^2 + z^\intercal x}{\left\| x \right\|_2^2}$$

Note that by symmetry, we find that if we regress x on y we get

$$\tilde{a} = \frac{x^\intercal y}{\left\| y \right\|_2^2} = \frac{a \left\| x \right\|_2^2 + z^\intercal x}{a^2 \left\| x \right\|_2^2 + 2ax^\intercal z + \left\| z \right\|_2^2}$$

As $n \to \infty$ we can invoke the Law of Large Numbers[4] and we thus – given that $E(z) = \mathbb{E}(N) = 0$ – find:

$$\mathbb{E}(\hat{a}) = \frac{a\,\mathbb{E}(\left\| x \right\|_2^2) + 0}{\mathbb{E}(\left\| x \right\|_2^2)} \xrightarrow{p} a$$

$$\mathbb{E}(\tilde{a}) = \frac{a\,\mathbb{E}(\left\| x \right\|_2^2) + 0}{a^2\,\mathbb{E}(\left\| x \right\|_2^2) + 0 + \mathbb{E}(\left\| z \right\|_2^2)} \xrightarrow{p} \frac{aSNR}{a^2 SNR + 1}$$

Thus for large n we have that:

$$r_{x \to y} \approx y - ax = z$$

and

$$r_{y \to x} = x - \tilde{a} y \approx x - \frac{aSNR}{a^2 SNR + 1}(ax + z)$$

[4] The samples are iid, and we note that convergance in probability is preserved when taking products and continuous mappings.

Observe that if $SNR = 0$, then $r_{y\to x} \approx x$, in which case whaat??

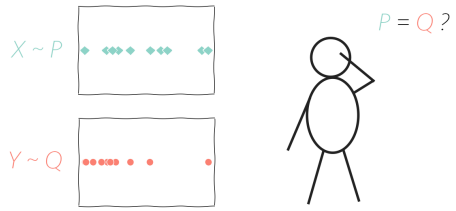
If however $SNR \to \infty$, then $r_{y\to x} \approx -\frac{1}{a}z \approx \frac{1}{a}z$, thus the residuals carry no information about causality.

We note that this somewhat formalises the intuition that we had about the role of noise in causality; it also shows that indeed, SNR plays an inverted role vis-a-vis that of communication theory.

If we are interested in finite sample results if we demand a certain accuracy for a given sample size, then we conjecture that SNR will play a key role in determining this.

3 Statistical distance

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 2 *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 3 *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. $k(x, .) \in \mathcal{H} \quad \forall x \in \mathcal{X}$
2. $\left\langle f, k(x, .) \right\rangle = f(x) \quad \forall f \in \mathcal{H}$

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

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

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

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

[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

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

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 kernes:

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

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 1 (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.

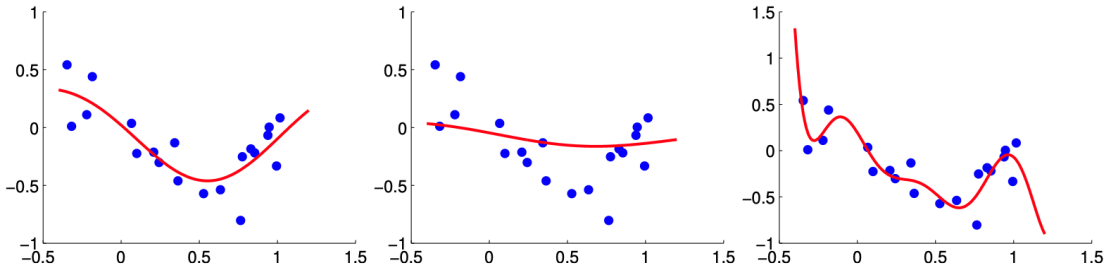


Figure 3.1 – Small RKHS norm results in smooth functions. From left to right $\Omega = .1$, $\Omega = 10$, $\gamma = 1e{-}7$, we fix the Gaussian kernel with $\gamma = 0.6$

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))

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

This motivates the following construction

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

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 1 *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 1 – 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.

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

Generalisation of the usual divergence, exploit jensen.

Talk about f-divergence, and give proof that D(p, q) >=0 and eq iff p == q

Talk about IPM vs $f-$divergence.

To test whether two random variables are independent X, Y

talk about L1 being f-divergence

3.4 Large deviations

Consider making this into a chapter

3.4.1 Hoeffding

3.4.2 Sanov

3.4.3 K-means

Proposed methods Part II

4 Causal Inference

4.1 The residual method

4.1.1 Introduction

The residual method is very simple...

4.1.2 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

$$\hat{e}_Y = Y - \hat{f}_Y(X) \tag{4.1}$$
$$\hat{e}_X = X - \hat{f}_X(Y) \tag{4.2}$$

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[1] 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

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

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

$$\mathbb{P}[C \leq W] \geq \mathbb{P}\left[\bigcup_{\tau \in \mathbb{Q}} C \leq \tau \cap W > \tau\right] \tag{4.3}$$
$$\geq \mathbb{P}[C \leq \tau \cap W > \tau] \tag{4.4}$$
$$\geq \mathbb{P}[C \leq \tau] - \mathbb{P}[W \leq \tau] \tag{4.5}$$

The first inequality is due to the fact that we are only taking the union in the rationals[2]. 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 $\mathbb{P}[W \leq \tau]$.

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

[2] 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).

4.1.3 Bounding the false false postive

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

$$\mathbb{P}[C \geq \tau] = \mathbb{P}\left[\sum_{i=1}^{m} |b_i - \frac{1}{m}| \geq \tau\right] \tag{4.6}$$

$$\leq \mathbb{P}\left[m \max_i |b_i - \frac{1}{m}| \geq \tau\right] \tag{4.7}$$

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

$$\leq m\,\mathbb{P}\left[|b_0 - \frac{1}{m}| \geq \frac{\tau}{m}\right] \tag{4.9}$$

$$\leq m2\exp\left(-2n\frac{\tau^2}{m^2}\right) \tag{4.10}$$

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.

4.1.4 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[3] 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

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

$|U| = m$. 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.

Theorem 2 (Sanov's theorem) *Let* Π *be a set of distributions on* U, *and* $m = |U|$. *Let*

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

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:

$$\mathbb{P}[W \leq \tau] = \mathbb{P}\left[\tilde{B} \in \Gamma_\tau\right] \leq (n+1)^m 2^{-nD(\tau)} \tag{4.11}$$

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:

$$\mathbb{P}[C \leq W] \geq \mathbb{P}[C \leq \tau] - \mathbb{P}[W \leq \tau] \tag{4.12}$$

$$\geq 1 - 2m \exp\left(-2n \frac{\tau^2}{m^2}\right) - (n+1)^m 2^{-nD(\tau)} \tag{4.13}$$

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.

4.2 The twin test

4.2.1 Intuition

Suppose that we have our typical **ANM**

$$Y = f(X) + N$$

The key observation is that if we partition the data in some intervals (e.g. uniform intervals), then if we look at two of these intervals we note that, while the distribution of y will differ – due to either X not being uniform and or the non-linearities due to f – the residuals will in fact be the same for both intervals due to the i.i.d. assumption.

In fact, if we a large enough number of samples, then – assuming that we find good models – we can be source that the difference between the empirical distribution of the residuals between these subets of X goes to 0. By the LLN the empriical CDFs will in fact converge a.s. to the CDF of N.

If on the other hand, we wrongly assume that Y -> X, we can be nearly certain that the additive noise that we find when fitting the reverse model will depend on Y. These observations motivate the following algorithm:

4.2.2 Algorithm

Algorithm 1 Given data x, y, the algorithm returns the predicted causal direction.

Precondition: x and y are vectors of the same length

function TWINSCORE(x, y)
  $X, Y, k \leftarrow partition(x, y)$
  for $i \leftarrow 1$ to k **do**
    $\hat{f}_i \leftarrow fit(X_i, Y_i)$
    $e_i \leftarrow Y_i - \hat{f}_i(X_i)$
  $\hat{C} \leftarrow \max_{i,j} \left\| \hat{p}_{e_i} - \hat{p}_{e_j} \right\|_{L_1}$
  return $\hat{C}$

Algorithm 2 Given data x, y, the algorithm returns the predicted causal direction.

Precondition: x and y are vectors of the same length

function TWINTEST(x, y)
  $\hat{C}_{X\to Y} \leftarrow twinscore(x, y)$
  $\hat{C}_{Y\to X} \leftarrow twinscore(y, x)$
  return $\hat{C}_{X\to Y} > \hat{C}_{Y\to X}$

4.2.3 Theory

proooooff

5 Experiments

Blah blah

5.1 Benchmark

blup di blue

5.2 Examples

In this chapter we will see some examples of tables and figures.

5.3 Tables

Let's see how to make a well designed table.

The table 5.1 is a floating table and was obtained with the following code:

```
\begin{table}[tb]
\caption[A floating table]{A floating table.}
\label{tab:example}
\centering
\begin{tabular}{ccc}
\toprule
```

Table 5.1 – A floating table.

name	weight	food
mouse	10 g	cheese
cat	1 kg	mice
dog	10 kg	cats
t-rex	10 Mg	dogs

```
    name    & weight & food   \\
\midrule
    mouse   & 10  g  & cheese \\
    cat     &  1 kg  & mice   \\
    dog     & 10 kg  & cats   \\
    t-rex   & 10 Mg  & dogs   \\
\bottomrule
\end{tabular}
\end{table}
```

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

5.4 algo

Algorithm 3 Counting mismatches between two packed strings

Precondition: x and y are packed strings of equal length n

1: **function** DISTANCE(x, y)
2: $z \leftarrow x \oplus y$ ▷ $\oplus$: bitwise exclusive-or
3: $\delta \leftarrow 0$
4: **for** $i \leftarrow 1$ to n **do**
5: **if** $z_i \neq 0$ **then**
6: $\delta \leftarrow \delta + 1$
7: **return** δ

6 Conclusion

Good job

6.1 TODO

Talk about SNR, role with shannon, and how it affects prediction in a reverse way here! Cite shanon!

Note on how SNR makes also the Kmeans based algo hard; i.e. the noise that is different is in the edges and becomes negligble.

Note on how the X indep N -> Xtilde indep N tilde only true for gaussian; for others, there will be dependence which the algo we propose can exploit (new one)

Briefly discuss AIC / model selection intution about using poly reg since it's local aprox https://stats.stackexchange.com/questions/9171/ aic-or-p-value-which-one-to-choose-for-model-selection

Note that problem is similar to change detection but it should be easier? -> we don't need to known when it changes

6.1.1 Ideas

-> no free lunch theorem <-> covariate shift?

How can we tackle causality?

In the absence of noise, and the process is bijective, then it is impossible to distinghuis, if however, ...

Shannon answered the question: given the most simple communication system: "How reliably can we communicate given a certain noise level"

In some sense what we would like to answer is, given a certain noise level, how reliably can we predict the causal relation.

Some points:

1. In causality we use noise, whereas in virtually all other domains such as communication theory the aim is combat noise.

Interesingly yet again, the Gaussian case ends up being a difficuly case. For instance, the motivation to look at the AGN additive gaussian noice channel is that the gaussian is the most difficult distribution in the entropic sense; but so it is as well in the bianry case setting due to:

thm.

A An appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

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.

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.

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

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

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.

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.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

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.