ONLINE LEARNING AND CAUSALITY

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

Thèse n. 1234 2020
présentée le 14 juillet 2020
à la Faculté des sciences de base
laboratoire SuperScience
programme doctoral en SuperScience
École polytechnique fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Paolino Paperino

acceptée sur proposition du jury :

Prof Name Surname, président du jury
Prof Name Surname, directeur de thèse
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur

Lausanne, EPFL, 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 14, 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

The endevour of science is in some sense the uncovering of causal structures: does the mass of an object influence its acceleration in free fall? What genomes influence height (etc)? The ability to do experiments in Physics is what has allowed to confirm or uncover relations among objects; indeed, this is also how we learn best, by tweaking a system and having a direct feedback which allows us to evaluate our mental models. In the realm of causality, such a setting is what is known as the "Causal intervention framework" (need to check). Give quickly example about smoking, and illustrate how it would be harder without interventions.

An excellent question about a multivariable causal system is to ask, "what is the minimum number of interventions one has to do to achieve some alpha-confidence about the causal effect"

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 non-numbered chapter...

1.2 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 theorist, 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.

[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.

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:

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

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:

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

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

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

Thus for large n we have that:

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

and

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

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.

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

1.4 Outline

2 Methods for Causal Inference

yes

2.1 Introduction

hello

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

3 Statistical Thoery

In this chapter we will see some examples of mathematics.

Change title To

Measures, bounds and Hilbert spaces.

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.

3.1 Bounds

3.1.1 Hoeffding

3.1.2 Sanov

3.2 Regression

3.2.1 thoery

3.3 K-means

3.4 Probability metrics

3.4.1 Introduction

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

We will next see two

3.4.2 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

3.4.3 Integral Probability Metric

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.

This motivates the following

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

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

See for example Sriperumbudur et al. (2009) for a detailed analaysis

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 end up with different popular distance measures, such as the Wasserstein distance or Total variation distance to name a few.

Note that the variational nature of the measure makes it a hard optimisation problem to tackle. If we could somehow decompose g linearly into simple components (think Fourier transforms), then we could use the linearity of expectation to simplify the problem: this is exactly what the following restriction achieves.

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.

First note that if ϕ is the associated feature map to the kernel k from the associated RKHS $\mathcal{H}$ then we have $g(x) = \langle g, \phi(x)\rangle$.

We can thus do the following simplification by linearity of expectation

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

Where we have defined[1]

$$\mu_P = \mathop{\mathbb{E}}_{X\sim P} \phi(X)$$

With this simplification and Cauchy Schwartz we have the following chain of equalities

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

[1] Note that this is well defined so long as ϕ is is continuous and bounded; one can also show that this object is continuous (See Peters (2008)). Note that $\mu_P \in \mathcal{H}$.

We can therefore see the MMD as the feature mean difference of the distributions.

We can further simplify the MMD

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

which we can straightforwardly estimate with samples; all the we require is to specify a kernel: so how do we choose a kernel?

We need to ensure that MMD$P,Q)=0$ iff $P=Q$, in other words, μ_P needs to be injective as a function of P. There are

We note that HSIC is to MMD, what the Mutual Information is to the Kullback–Leibler divergence.

Benefits -> no bins!

but this is a lie! HSIC needs to choose a kernel, and the hyperparameters...

In their study (Sriperumbudur et al. (2009)) – it is shown that 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

3.5 Reproducing Kernel Hilbert Space

We will begin by defining the kernel,

Definition 1 *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*[2].

It is easy construct new kernels since they are preserved under addition, multiplication and

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

other operations. (See for example...)

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

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

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.

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

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

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

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

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

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

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

It is possible to equip it 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).

We will now show case a brief example to illustrate the power of the RHKS. 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.

We can estimate f as follows

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

An amazing result is that an optimisation of the above form will always admit a representation of the form – assuming Ω is increasing?:

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

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

This is known as the Representer Theorem. If we wish to approximate a prediction for some x, we can do so as follows:

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

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, we k is the Gaussian Kernel, then the solution will be a linear combination of scaled gaussians centered at the data points [4].

As a final remark we will explain the role of the penalty $\Omega \left\| f \right\|_{\mathcal{H}}^2$; from the ML world, we now that this kind of term is known as regularisation and is supposed to help the model choose a "simpler" model. This is exactly what is happening, as this term will penalise non-smooth functions.

This can be seen by using Mercer's Theorem – a Generalisation of the spectral theorem for

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

[4] In fact this will always be the case when we can write $k(x_i, x) = k(x_i - x)$

positive-semidefinite matrices[5].

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:

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

We can thus rewrite the solution as follows

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

Note that due to the $\Omega \left\| f \right\|_{\mathcal{H}}^2$ penalty, f_i^* must decay for higher values of i. Note that for example in the Fourier Transform[6], in the basis $\{\psi_i\}$, higher values if i correspond to higher frequency functions; similarly, for the Gaussian Kernel, higher indicies basis functions correspond to higher frequecies. Thus, a higher Ω will force a faster decay on f_i and thus result in smoother functions – in principle, less overfitting.

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

[6] Heloo

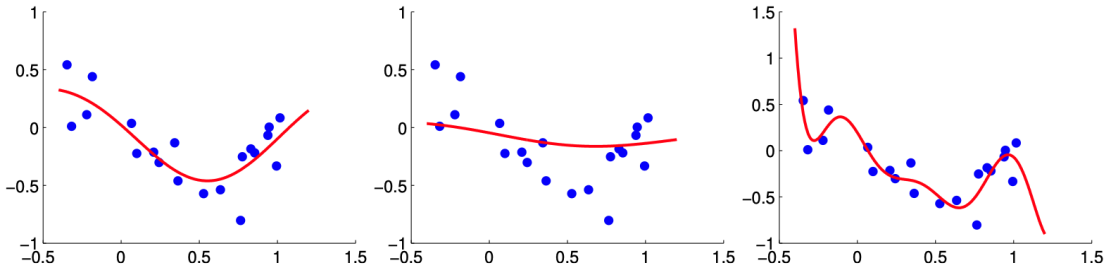


Figure 3.1 – Small RKHS norm results in smooth functions. From left to right $\Omega = .1$, $\Omega = 10$, $\gamma = 1e{-}7$, we fix the Gaussian kernel with $\gamma = 0.6$

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

prooooff

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 Figures

Let's see now how to put one or several images in your text.

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

```
\begin{figure}[tb]
\centering
\includegraphics[width=0.5\columnwidth]{galleria_stampe}
\caption[A floating figure]{A floating figure ... }
\label{fig:galleria}
\end{figure}
```

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

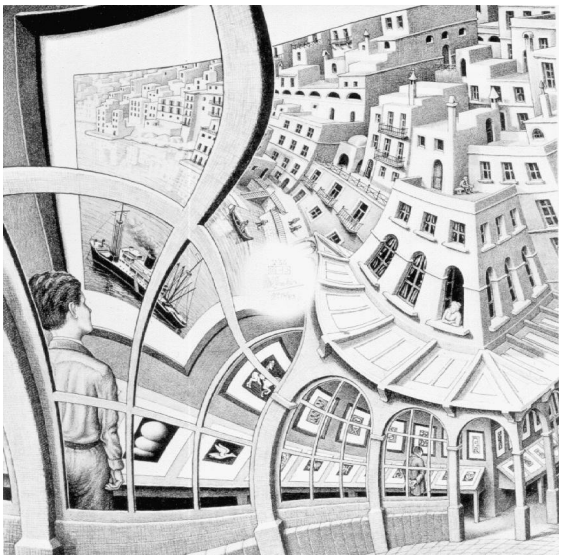


Figure 5.1 – A floating figure (the lithograph *Galleria di stampe*, of M. Escher, got from http://www.mcescher.com/).

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.

The figure 5.3 is a floating figure and was obtained with the following code:

```
\begin{figure}[tb]
\centering
\subfloat[Asia personas duo.]
{\includegraphics[width=.45\columnwidth]{lorem}} \quad
\subfloat[Pan ma signo.]
{\label{fig:ipsum}%
\includegraphics[width=.45\columnwidth]{ipsum}} \\
\subfloat[Methodicamente o uno.]
{\includegraphics[width=.45\columnwidth]{dolor}} \quad
\subfloat[Titulo debitas.]
{\includegraphics[width=.45\columnwidth]{sit}}
\caption[Tu duo titulo debitas latente]{Tu duo titulo debitas latente.}
\label{fig:esempio}
\end{figure}
```

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique,

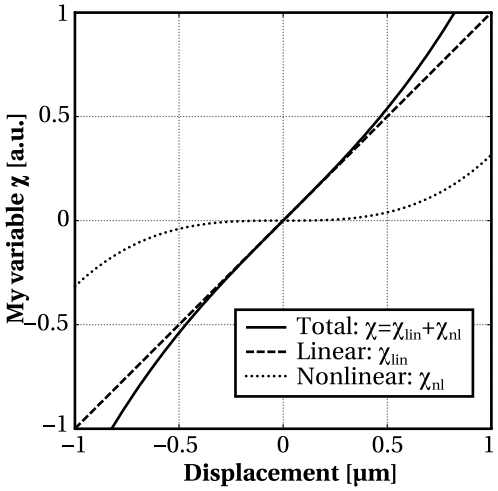


Figure 5.2 – A floating figure with text typeset in "Utopia Latex", a font provided in the template-folder for typesetting figures with greek characters. The text has been "outlined" for best compatibility with the repro during the printing.

libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

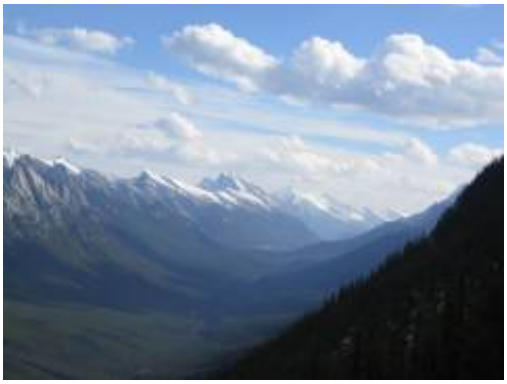
Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl.

(a) Asia personas duo.

(b) Pan ma signo.



(c) Methodicamente o uno.

(d) Titulo debitas.

Figure 5.3 – Tu duo titulo debitas latente.

Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit

amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

5.5 Very important formulas

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.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} P_0 \\ P_1 \\ P_T \end{bmatrix} = \begin{bmatrix} \frac{P_1}{\tau_{10}} + \frac{P_T}{\tau_T} - \frac{P_0}{\tau_{ex}} \\ -\frac{P_1}{\tau_{10}} - \frac{P_1}{\tau_{isc}} + \frac{P_0}{\tau_{ex}} \\ \frac{P_1}{\tau_{isc}} - \frac{P_T}{\tau_T} \end{bmatrix} \tag{5.1}$$

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

$$\bar{I_f}(\vec{r}) = \gamma(\vec{r})\left(1 - \frac{\tau_T P_T^{eq}\left(1 - \exp\left(-\frac{(T_p - t_p)}{\tau_T}\right)\right)}{1 - \exp\left(-\frac{(T_p - t_p)}{\tau_T} + k_2 t_p\right)} \times \frac{\left(\exp\left(k_2 t_p\right) - 1\right)}{t_p}\right) \tag{5.2}$$

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

5.6 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 ← x ⊕ y                          ▷ ⊕: bitwise exclusive-or
3:     δ ← 0
4:     for i ← 1 to n do
5:         if z_i ≠ 0 then
6:             δ ← δ + 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

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.

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

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

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*.