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 12 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 12, 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

Acknowledgements **i**

Abstract **iii**

Introduction **1**
0.1 Problem and Motivation . . . 1
0.1.1 TODO ideas . . . 2
0.2 SNR and causality . . . 2

1 Tables and Figures **5**
1.1 Tables . . . 5
1.2 Figures . . . 6

2 Mathematics **11**
2.1 Very important formulas . . . 11
2.2 algo . . . 12

3 Causal Inference **13**
3.1 The residual method . . . 13
3.1.1 Introduction . . . 13
3.1.2 Proof of consistency: A tale of two bounds . . . 13
3.1.3 Bounding the false false postive . . . 15
3.1.4 Bounding the false negatives . . . 15
3.2 The twin test . . . 16
3.2.1 Intuition . . . 16
3.2.2 Algorithm . . . 17
3.2.3 Theory . . . 17

A An appendix **19**

Bibliography **21**

Bibliography **21**

Introduction

0.1 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...

0.1.1 TODO ideas

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

0.2 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:

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

[1] quote comm book
[2] Shannon had a very deep...

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.

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:

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

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:

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

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

[3]If it is not injective, then the function is not invertible, and thus only one causal direction is possible.

$$\tilde{a} = \frac{x^\mathsf{T} y}{\|y\|_2^2} = \frac{a\|x\|_2^2 + z^\mathsf{T} x}{a^2\|x\|_2^2 + 2ax^\mathsf{T} z + \|z\|_2^2}$$

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:

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

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

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 it is trivial to see which is the causal direction.

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.

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

1 Tables and Figures

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

1.1 Tables

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

The table 1.1 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
    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,

Table 1.1 – A floating table.

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

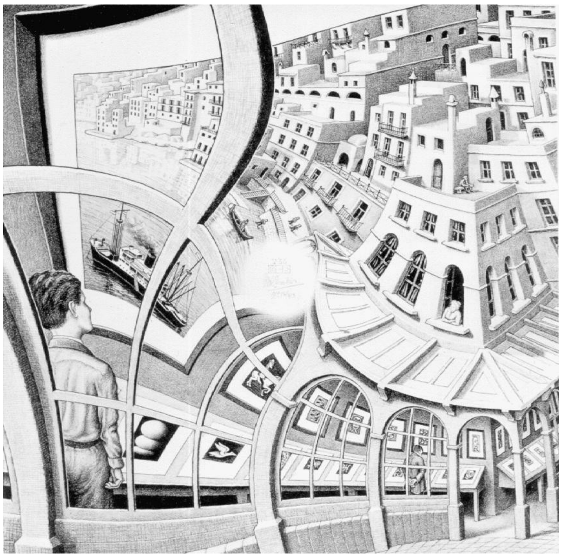


Figure 1.1 – A floating figure (the lithograph *Galleria di stampe*, of M. Escher, got from http://www.mcescher.com/).

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.

1.2 Figures

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

The figure 1.1 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}
```

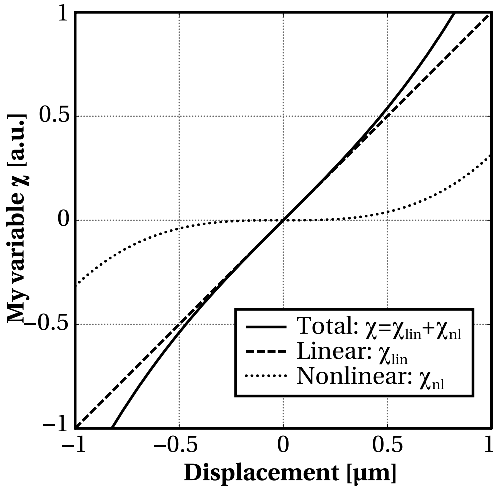


Figure 1.2 – A floating figure with text typeset in "Utopia Latex", a font provided in the template-folder for typesetting figures with greek characters. The text has been "outlined" for best compatibility with the repro during the printing.

```
\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 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.

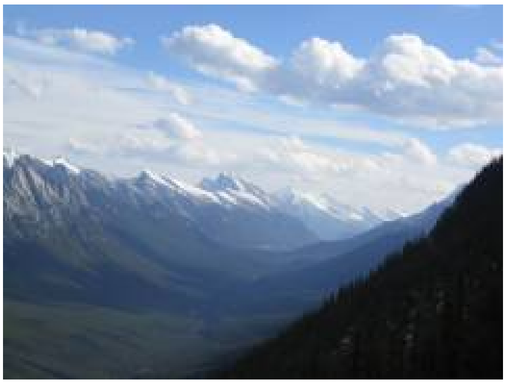
Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit 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 1.3 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}} \\
```

(a) Asia personas duo.

(b) Pan ma signo.



(c) Methodicamente o uno.

(d) Titulo debitas.

Figure 1.3 – Tu duo titulo debitas latente.

```
\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, 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. 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.

2 Mathematics

In this chapter we will see some examples of mathematics.

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.

2.1 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.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} P_0 \\ P_1 \\ P_T \end{bmatrix} = \begin{bmatrix} \frac{P_1}{\tau_{10}} + \frac{P_T}{\tau_T} - \frac{P_0}{\tau_{ex}} \\ -\frac{P_1}{\tau_{10}} - \frac{P_1}{\tau_{isc}} + \frac{P_0}{\tau_{ex}} \\ \frac{P_1}{\tau_{isc}} - \frac{P_T}{\tau_T} \end{bmatrix} \tag{2.1}$$

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.

$$\bar{I}_f(\vec{r}) = \gamma(\vec{r})\left(1 - \frac{\tau_T P_T^{eq}\left(1-\exp\left(-\frac{(T_p-t_p)}{\tau_T}\right)\right)}{1-\exp\left(-\frac{(T_p-t_p)}{\tau_T}+k_2 t_p\right)} \times \frac{\left(\exp\left(k_2 t_p\right)-1\right)}{t_p}\right) \tag{2.2}$$

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.

2.2 algo

Algorithm 1 Counting mismatches between two packed strings

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

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

3 Causal Inference

3.1 The residual method

3.1.1 Introduction

The residual method is very simple...

3.1.2 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

$$\hat{e}_Y = Y - \hat{f}_Y(X) \tag{3.1}$$

$$\hat{e}_X = X - \hat{f}_X(Y) \tag{3.2}$$

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:

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

We next upper bound thise quantity in order to show consistency

$$\mathbb{P}\left[C \leq W\right] \geq \mathbb{P}\left[\bigcup_{\tau \in \mathbb{Q}} C \leq \tau \cap W > \tau\right] \tag{3.3}$$
$$\geq \mathbb{P}\left[C \leq \tau \cap W > \tau\right] \tag{3.4}$$
$$\geq \mathbb{P}\left[C \leq \tau\right] - \mathbb{P}\left[W \leq \tau\right] \tag{3.5}$$

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:

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

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

[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).

3.1.3 Bounding the false false postive

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

$$\mathbb{P}[C \geq \tau] = \mathbb{P}\left[\sum_{i=1}^{m} |b_i - \frac{1}{m}| \geq \tau\right] \tag{3.6}$$

$$\leq \mathbb{P}\left[m \max_i |b_i - \frac{1}{m}| \geq \tau\right] \tag{3.7}$$

$$= \mathbb{P}\left[\bigcup_i |b_i - \frac{1}{m}| \geq \frac{\tau}{m}\right] \tag{3.8}$$

$$\leq m\,\mathbb{P}\left[|b_0 - \frac{1}{m}| \geq \frac{\tau}{m}\right] \tag{3.9}$$

$$\leq m 2 \exp\left(-2n\frac{\tau^2}{m^2}\right) \tag{3.10}$$

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.

3.1.4 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 1 (Sanov's theorem) *Let* Π *be a set of distributions on* U*, and* $m = |U|$*. Let*

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

Then

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

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

$$\mathbb{P}[W \leq \tau] = \mathbb{P}\left[\tilde{B} \in \Gamma_\tau\right] \leq (n+1)^m 2^{-nD(\tau)} \tag{3.11}$$

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:

$$\begin{align} \mathbb{P}[C \leq W] &\geq \mathbb{P}[C \leq \tau] - \mathbb{P}[W \leq \tau] \tag{3.12} \\ &\geq 1 - 2m \exp\left(-2n\frac{\tau^2}{m^2}\right) - (n+1)^m 2^{-nD(\tau)} \tag{3.13} \end{align}$$

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.

3.2 The twin test

3.2.1 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:

3.2.2 Algorithm

Algorithm 2 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 ← partition(x, y)
    for i ← 1 to k do
        f̂_i ← fit(X_i, Y_i)
        e_i ← Y_i − f̂_i(X_i)
    Ĉ ← max_{i,j} ‖p̂_{e_i} − p̂_{e_j}‖_{L_1}
    return Ĉ
```

Algorithm 3 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)
    Ĉ_{X→Y} ← twinscore(x, y)
    Ĉ_{Y→X} ← twinscore(y, x)
    return Ĉ_{X→Y} > Ĉ_{Y→X}
```

3.2.3 Theory

proooofff

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

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

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