Data Analysis in the Earth & Environmental Sciences

Julien Emile-Geay
Department of Earth Sciences
julieneg@usc.edu

GEOL425L

USC University of Southern California

Typeset in LaTeX with help from MacTeX and a tweaked version of the Tufte LaTeX book class. Many figures generated via TikZ. Python highlighting by O. Verdier. All software open source.

Please cite as: *Emile-Geay, J., 2020: Data Analysis in the Earth & Environmental Sciences, 265pp, Fourth edition,* `http://dx.doi.org/10.6084/m9.figshare.1014336`.

Recompiled January 2020

CONTENTS

Chapter 1

PREAMBLE

I What is data analysis?

We all have a basic concept of what DATA means, so let us first define the word ANALYSIS. Etymologically, it is the union of the Greek *ana* (up, throughout) + *lysis* (a loosening, from *lyein* "to unfasten"). Dictionary.com gives four definitions of it:

1. the separating of any material or abstract entity into its constituent elements (opposed to synthesis).
2. this process as a method of studying the nature of something or of determining its essential features and their relations: *the grammatical analysis of a sentence*.
3. a presentation, usually in writing, of the results of this process: *The paper published an analysis of the political situation*.
4. a philosophical method of exhibiting complex concepts or propositions as compounds or functions of more basic ones.

Science is fundamentally a reductionist enterprise, and Earth Science is no exception. Faced with complex environmental systems, we wish to break them into simpler parts that we can better understand (definition 1); we must learn this process (definition 2), which is why you are reading these notes. We must also communicate the result of our analysis, in verbal or written form (as definition 3 alludes to), but also graphically. Indeed, a scientific analysis is only as good as it can be understood by a third party, so it behooves the analyst to be educated (know the methods), rigorous (apply the methods correctly, in conformity with their intended uses and underlying assumptions), honest (present candidly without withholding inconvenient information), lucid about the results (interpret them but not over interpret them), and transparent (communicate them clearly).

Now that we have defined the task that will occupy us for the rest of this book, let us take a step back. What do we mean by DATA? Data is a word about which most scientists think they agree, but turn out to disagree a fair bit. Data represent numerical representation of quantities (variables)[1] which, we would like to believe, tell us about a system under investigation. In today's world, data refer both to measurements of physical, chemical or biological quantities, or to their representation in a numerical model, though most experimentalists will scoff at the idea that models produce "data", by which they mean "observations". Still, volumetrically the amount of model-generated data far outweighs any observational data, so this is something with which today's data analyst must contend. On the other hand, most experimentalists believe their data to be hard evidence about a phenomenon of interest. Is that so? Evidence – always. Hard, not always, depending on its associated **uncertainties**. Perhaps the most important task incumbent to the analyst is that of assessing and communicating these uncertainties, and deriving the scientific conclusions that are warranted by them.

[1] a single data point, for those who like Latin, is called a *datum*

Hence, for the purposes of this book, data analysis is the task of decomposing data into simpler components, describing the key features of these components and their associated **uncertainties**, and communicating these key features. We have our work cut out for us. Let us begin.

II What should you expect from this book?

This book was written as notes to a class aptly named "Data Analysis in the Earth and Environmental Sciences" and, as such, introduces students to key methods used in those fields. Most of the examples are drawn from the Earth sciences, attempting to span a broad a spectrum thereof. Many of the concepts, however, are quite general and apply to virtually any dataset. Similarly, we chose to illustrate the numerical methods in Python[2], but most programming languages have equivalent solutions available.

[2] Using the so-called "Pylab stack" using NumPy/SciPy/Matplotlib

Data analysis draws on many mathematical concepts, some of which can be rather advanced. Our purpose is not to write an exercise in mathematical pedantry, but to borrow from mathematics what we need to better understand the Earth. This is a balancing act: in these notes we attempt to introduce the fundamental notions that are needed to understand the essence of each analysis technique, but there is only so much that one can teach in a semester. If you did not study mathematics very extensively, there will come a point where you will find the mathematics abstruse, overwhelming, even obnoxious; conversely, if you were trained in the mathematical sciences, you may at times be saddened, even appalled, by the absence of proofs or the lack of rigor in these

pages. Our intent is to give each student the rudiments to understand and use these techniques –perhaps even teach them, one day. As such, a modicum of mathematical education is required, amounting roughly to Calculus III, introductory linear algebra and trigonometry (Appendices, A, B and C). Probability theory is our main foundation, and none of it can be explained without the language of calculus and linear algebra. If you never studied either of those, find a good MOOC and immerse yourself. If you did study them, but never excelled at them, we hope that the desire to understand the Earth will drive you to learn some new math! Ultimately, however, if you want to use those techniques correctly, you will have to understand where they come from, and to do that you will have to delve into some mathematics – a black box approach will not suffice. How much you want to do this is your choice, but it is hard to know too much about this topic.

Each Earth Science department worth its salt has a class like this one (at USC, GEOL425L); they each make different choices about what topics to emphasize. In the process, they all achieve different compromises between exposing students to the most common techniques used in their field and giving them enough background to understand these techniques at a deep level. It is our belief that knowledge is most efficiently acquired through practice; as such, a key component of this class is a set of weekly laboratory practicums that put into use the notions presented in the lectures. Equally important, and perhaps more relevant for those of you who are becoming researchers (whether graduate or undergraduate), a final project will *apply the concepts taught here on your own dataset*. If you play your cards right, this will form the foundation of a thesis chapter or paper, which will make you – and your advisor – very happy indeed. Do we have your attention now?

III Structure

The class is articulated around three main themes:

1. Living in an uncertain world
2. Living in the temporal world
3. Living in multiple dimensions

Because **uncertainties** will turn out to matter at every step of the way, we need to define a common language to deal with them. Such is the object of *statistics*, rooted in *probability theory*, which are tackled in the first third of these notes. Much of those fields of knowledge rely heavily on results from calculus and linear algebra, which are briefly reviewed in Appendices A & B. Next we focus on datasets that follow a sequential order; we call those timeseries, though the independent variable need

not be time (it could as well be space). Here a fundamental mathematical tool is Fourier analysis, in which we delve at some length, with the ultimate goal of performing spectral analysis and signal processing. For this we need some basic notions of trigonometry and complex algebra, reviewed in Appendix C. Of course, we shall not forget that all our spectral estimates are just that – estimates – so the probabilistic language developed in Part 1 will still come to good use.

We finish with a variety of topics involving the interaction of several variables: this may include predicting one physical variable from measurements of another (linear models, least squares), analysis of spatiotemporal variability (principal component anaysis), or the estimation of hidden parameters from sparse measurements (geophysical inverse theory). Much of this requires simplifying matrices to a diagonal form, which we review in Appendix D.

Without further ado, we now begin our odyssey into data analysis.

IV Acknowledgements

Thorsten W. Becker, who put this class on the books at USC, was instrumental for much of the labs and least squares/inverse theory chapters. Appendix A is a shameless (but authorized) rip-off of a similar appendix in his and B. Kaus' excellent e-book Numerical Modeling of Earth Systems.

Dominique Guillot is to be commended for the excellent review of linear algebra (Appendix B), and much of chapters 3 and 5.

I thank Elisa Ferreira, Laura Gelsomino and Antonio Mariano for their help with LaTeX typesetting, and all the ERTH425L students who have pointed out typographical errors over the years[3]. Despite my best efforts, the probability that some such errors remain in the following pages is close to unity, so I am grateful for any comment that can help us fix that.

[3] non-exhaustive list: Jianghao Wang, Kirstin Washington, Alexander Lusk, Kevin Milner, Billy Eymold, Joseph Ko, Judith Gauriau

Part I

Living in an uncertain world

Chapter 2

PROBABILITY THEORY

> "Probability is the only satisfactory way to reason in an uncertain world."
>
> *Dennis Lindley*

One seldom has all the information that one wishes for: often the measurements are too few, too imprecise (or both) to allow us to conclude much with certainty. Yet this does not mean that we know nothing – we may have a lot of information about the Earth system, and what we need is a way to reasoning quantitatively about it, within these uncertainties. This is the domain of probability theory, which encodes those rules of reasoning in mathematical form.

I Probability Theory as Extended Logic

Probability theory provides an automatic means to conduct plausible reasoning given scientific evidence (observations of the system under consideration).

Deductive vs Plausible reasoning

Suppose some dark night a policeman walks down a street, apparently deserted[1]. Suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a man wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman has little hesitation in concluding that the man is a burglar, and promptly arrests him. But by what reasoning does he arrive at that conclusion?

It should be clear that our policeman's conclusion was not a logical deduction from the evidence; for there may have been a perfectly innocent explanation for all this. It might be, for instance, that the man was the owner of the jewelry store, that he was coming home from a masquerade party, and didn't have the key with him. However, just as

[1] The following two sections borrow heavily from the excellent, if somewhat peevish, work by *Jaynes* (2004)

he walked passed the store, a passing truck threw a stone through the window, and he was only protecting his own property.

Now, while the policeman's reasoning process was not a logical deduction from *perfect* knowledge of the situation, we will grant that it had a certain degree of validity. The evidence above does not make the man's dishonesty *certain*, but it does make it extremely plausible. This is an example of a kind of reasoning in which all of us are proficient, without necessarily having learned the mathematical theory behind it. Every day we make decisions based on imperfect information (e.g. it might rain, should I take my umbrella?), and the fact that there is some uncertainty in no way paralyzes our actions.

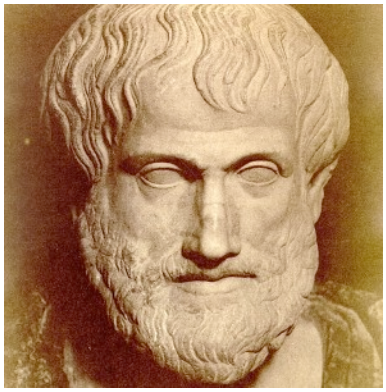


Figure 2.1: A bust of Aristotle, whose real face is quite uncertain.

We therefore contrast this *plausible reasoning* with the *deductive reasoning* whose formalism is generally attributed to Aristotle (Fig. 2.1). The latter is usually expressed as the repeated applications of the *strong syllogism*:

(S1)
If A is true, then B is true;
but A is true.

Therefore B is true. $\therefore$

and its inverse:

(S2)
If A is true, then B is true;
but B is false.

Therefore A is false. $\therefore$

Figure 2.2: The Hungarian mathematician George Pólya, December 13, 1887 – September 7, 1985)

This is the reasoning we would conduct in an ideal world. In the real world, there is almost never sufficient information to determine the absolute veracity of any proposition. However, George Pólya argued that we can still reason via *weak syllogisms*:

(S3)
If A is true, then B is true;
but B is true.

Therefore A becomes more plausible. $\therefore$

Example
$A \equiv$ *"it will start to rain by 10 am at the latest" ;*
$B \equiv$ *"clouds will appear before 10 am".*

Observing clouds at 9:45am does not give us logical certainty that it will rain by 10, only a strong inkling that it might be the case (if the clouds are sufficiently dark). Note that the connection is purely logical, here – not physical. We know that clouds are a necessary connection for rain (Rain $\Rightarrow$ Clouds), while the causal relationship goes the other

way (Clouds ⇒ Rain). Probability theory is not charged with inferring physical causation, only logical connections. Another weak syllogism is:

(S4)
If A is true, then B is true;
but A is false.

Therefore B becomes less plausible. $\therefore$

In this case, the evidence does not *prove* that B is false, only that one of the possible reasons why it might be true has been eliminated, so we feel less confident about it. Scientific reasoning almost always looks like S3 and S4.

Now, our policeman's reasoning was not even of the above types. It is a still weaker form of syllogism:

If A is true, then B becomes more plausible;
but B is true.

Therefore A becomes more plausible. $\therefore$

Despite this apparent weakness, the burglary case should convince you that such form of reasoning may approach the power of deductive reasoning if the evidence is sufficiently strong. One subtlety is that we are not only referring to present evidence, but also to past experience. In this case, our policeman's *prior information* is that people smuggling jewelry out of a broken window tend to be thieves, not jewelers. His judgement would be very different if he lived in a city where things were usually otherwise, and he would soon learn to dismiss the evidence as something perfectly ordinary.

Thus, our reasoning about the world involves several things: linking statements together by way of degrees of plausibility, and prior information about some of these statements. How our brain achieves this is in fact immensely complex, and we conceal this complexity by calling it *common sense*.

Designing a thinking robot

Of course, even with the most common sense in the world, anyone can make mistakes. Objective and scientific though we'd like to be, we inevitably let other considerations enter our judgement about the plausibility of a hypothesis: how elegantly it is stated, who stated it, whether we had coffee that morning, etc. To eliminate the risk of inconsistent judgement, and to be able to process a lot of information at once, we would like to obtain a set of automatic rules for this kind of scientific reasoning, so we can teach a computer to do it faster than we can, and

so our results can be reproducible by anyone given the same information. What do we need to tell our robot?

We first need to operate on *logical propositions*. Given two propositions E and F, and their negation (or complement) $\bar{E}$ (also written NOT E, or E^c) and $\bar{F}$ (NOT F, F^c), we define:

$$E \cap F = \text{Both E and F are true} = \text{AND} \tag{2.1}$$
$$E \cup F = \text{Either E or F are true} = \text{OR} \tag{2.2}$$

This is sometimes represented by a Venn diagram (Fig. 2.3):

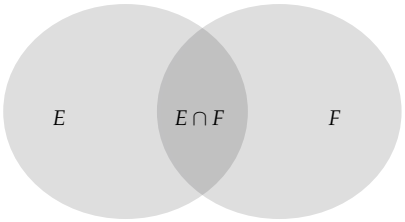


Figure 2.3: Venn diagram of two events, E and F, that are not mutually exclusive and overlap (the intersection $E \cap F$, is non-empty). The total area in gray is the union of E and F ($E \cup F$)

These operations are duals of each other:

$$\overline{E \cap F} = \bar{E} \cup \bar{F} \tag{2.3a}$$
$$\overline{E \cup F} = \bar{E} \cap \bar{F} \tag{2.3b}$$

(e.g. the opposite of "rich and beautiful" is "poor OR ugly"). Eq. (2.3a) are known as DeMorgan's laws. One can show that this basic set of rules allow to decompose any complex proposition into simpler ones, so we need is a set of rules to assign degrees of plausibility to these propositions.

Finally, we will denote by $(E|F)$ the event "E given F", meaning "E given that F is true".

Basic desiderata for plausible reasoning

Measurability Degrees of plausibility are represented by real numbers, which we call a probability, denoted $\mathbb{P}$.

Common sense If we had old information G which gets updated to G' in such a way that E becomes more plausible ($\mathbb{P}(E|G') > \mathbb{P}(E|G)$) but the plausibility of F is unchanged, then this can only increase (not decrease) the plausibility that both E and F are true: $\mathbb{P}(E \cap F|G') > \mathbb{P}(E \cap F|G)$. It also reduces the plausibility that $\bar{E}$ is false: $\mathbb{P}(\bar{E}|G') < \mathbb{P}(\bar{E}|G)$. That is, the rules of probability calculus must conform to intuitive human reasoning (common sense).

Consistency If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same probability. Also, if in two problems our state of knowledge is identical, then we must assign identical probabilities in both. Finally, we must always consider ALL the information relevant to a question, without arbitrarily deciding to ignore some of it. Our robot is therefore completely non-ideological.

Figure 2.4: American physicist Richard T. Cox, ca 1939.

Cox's theorem states essentially that these postulates are jointly sufficient to design a robot that can reason scientifically about the world. This is not, however, how probability theory was initially built, so a historical digression is warranted.

A brief history of probability

Figure 2.5: Blaise Pascal (Source:Wikimedia Commons).

- **1654**: A gambler's dispute led to the creation of a mathematical theory of probability by two famous French mathematicians, Blaise Pascal and Pierre de Fermat.

  Antoine Gombaud, Chevalier de Méré, a French nobleman with an interest in gaming and gambling questions, called Pascal's attention to an apparent contradiction concerning a popular dice game. This problem and others posed by de Méré led to an exchange of letters between Pascal and Fermat in which the fundamental principles of probability theory were formulated for the first time. This was the first occurrence of "chance" and randomness" in scientific parlance.

- **1657**: Christian Huygens published the first book on probability; entitled *De Ratiociniis in Ludo Aleae*. The major contributors during this period were Jakob Bernoulli (1654-1705) and Abraham de Moivre (1667-1754).

Figure 2.6: Pierre de Fermat (Source:Wikimedia Commons).

- **1812**: Pierre Simon Laplace (1749-1827) introduced a host of new ideas and mathematical techniques in his book, *Théorie Analytique des Probabilités*. Before Laplace, probability theory was solely concerned with developing a mathematical analysis of games of chance. Laplace applied probabilistic ideas to many scientific and practical problems. The theory of errors, actuarial mathematics, and statistical mechanics are examples of some of the important applications of probability theory developed in the 19^{th} century.

- **1933** : End of a long struggle to find a definition until Kolmogorov's axioms (1933). The difficulty had to do with finding a mathematical definition that would conform to other (e.g. logical, intuitive) considerations.

Figure 2.7: Pierre Simon Laplace (23 March 1749 – 5 March 1827).

II Notion of probability

It turns out that the notion of probability is difficult enough a concept that statisticians, probability theorists, and philosophers still argue about it. Here we present three views of the topic: the Frequentist view, the Bayesian view, and the axiomatic view.

Frequentist Interpretation

Let E be an event in a random experiment. An experiment might be "Rolling a die"; an event (outcome) of this experiment might be "Obtaining the number 6". Let us repeat this experiment N times, each time updating a counter x_i:

$$x_i = \begin{cases} 1\,, & \text{if } E \text{ occurred}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{2.4}$$

The frequentist interpretation holds that , as N gets large, then the frequency of occurrence of E, $f_N(E)$, converges to a constant, which we call its probability $\mathbb{P}(E)$:

$$\text{(Relative Frequency)} \qquad \lim_{N\to\infty} \frac{x_1 + x_2 + \ldots + x_N}{N} = \mathbb{P}(E) \tag{2.5}$$

This is the so-called *Law of large numbers*.

The probability can also be thought of as:

$$\mathbb{P}\,(E) = \frac{\text{Number of ways } E \text{ can occur}}{\text{Total number of possible outcomes}}\,. \tag{2.6}$$

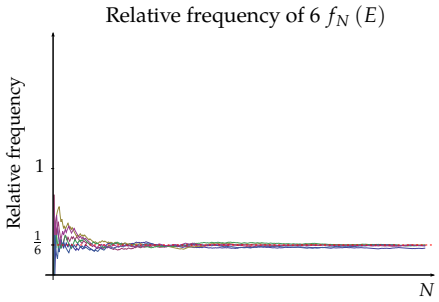


Figure 2.8: Plot of $f_N(E)$ of rolling one fair die in function of the number of experiments, N. We can see that for large N, the relative frequency tends to $\frac{1}{6}$, which we define as the probability of occurrence of the event.

Example (Rolling two dice)

We represent the two dice as (i, j), where i = result of throwing the 1^{st} die and j = result of throwing the 2^{nd} die. Possible outcomes S are:

$$S = \{(1,1)\,,(1,2)\,,\ldots,(1,6)\,,(2,1)\,,\ldots,(6,1)\}\,,$$

where,

$$|S| = 36\,.\ \textit{where}\ |\ |\ \textit{means "number of elements"}$$

What is the probability of the event $E = sum\ > 8$?

Possibilities:

$$E = \{(3,6)\,,(4,6)\,,(5,6)\,,(6,6)\,,(4,5)\,,(5,4)\,,(5,5)\,,(6,3)\,,(6,4)\,,(6,5)\}\,,$$
$$\Rightarrow 10\ \textit{Possibilities}\,.$$

$$\mathbb{P}\,(E) = \frac{|E|}{|S|} = \frac{10}{36} = \frac{5}{18} \approx 0.278\,.$$

The Bayesian Viewpoint

In the frequentist interpretation, probabilities are the frequencies of occurrence of random events as proportions of a whole. But what if we cannot repeat these events to measure a frequency? In the words of

D.G. Martinson, *"one cannot re-run the experiment called the Earth"*. Does this mean that we can form no judgement about the plausibility of an event, like A = "a meteorite killed the dinosaurs at the end of the Cretaceous" (which, by definition, only happened once)? If so, what would be the use of probabilities in Earth Science?

In the Bayesian interpretation, probabilities are *rationally coherent degrees of belief, or a degree of belief in a logical proposition given a body of well-specified information*. Put it another way: it is *a measure of a state of knowledge about a problem*.

Figure 2.9: Rev. Thomas Bayes, 1701-1761, Presbyterian minister.

General Idea:

→ Start from prior knowledge.

→ Collect observations about the system (perform experiments, go out in the field)

→ Update prior knowledge in light of new observations. This is encoded in a *posterior probability*

Example

We roll a dice a small number of times, with the result:

$$1,3,2,1,1,5,4,4,3,5,6,6,3,1,5$$

$$p_i = \mathbb{P}(X = i) = ?$$

Let's assume we have no reason to think that the die is loaded. A priori considerations of symmetry would lead us to set $p_i = 1/6, i \in [1:6]$, *but we should perhaps not be so definite (after all, if we decide from the outset that the die is fair, how could we possibly revise our judgement if the evidence suggests otherwise?). So instead we set a* prior probability distribution *for* p_i, *illustrated in Fig. 3.*

Now, after rolling the die 15 times, we observe the sequence

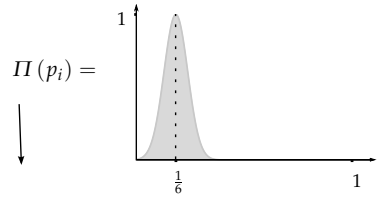
$$X = \{1,3,2,1,1,5,4,4,3,5,6,6,3,1,5\} \tag{2.7}$$

, after which we update our estimate of p_i*:*

$$\mathbb{P}(p_i|X) = \frac{\overbrace{\mathbb{P}(X|p_i)}^{\text{likelihood of obtaining } X \text{ given } p_i} \cdot \overbrace{\mathbb{P}(p_i)}^{\text{prior}}}{\sum_i \mathbb{P}(X|p_i) \cdot \mathbb{P}(p_i)}$$

$\downarrow$

Updated posterior probability.



Prior distribution of p_i

Figure 2.10: Prior distribution of p_i of rolling a dice with mean $\frac{1}{6}$.

And we conclude that this limited dataset contains no damning evidence that the die is loaded.

Advantages / Disadvantages of Bayesian framework:

Advantage One can now compute the probability of any event, regardless of whether it can be repeated *ad nauseam*.

Advantage The framework incorporates all information, which is relevant to the problem, including information known *a priori* (*e.g.* before collecting the data). This may lead to better estimation, especially when the sample size is too small for frequentists theorems to apply.

Disadvantage We now rely on a *subjective* prior, and different priors may lead to different results. On the other hand, the prior can be justified by theory (e.g. considerations of symmetry, physical laws) or other experiments. The role of prior information becomes less important as the number of observations increases, so in the limit $n \to \infty$, the frequentist and Bayesian interpretations often lead to the same results.

Disadvantage In many cases Bayesian analysis is more cumbersome and more computationally demanding.

Laplace's ideas, in hindsight, would now be considered Bayesian. Ironically, it is not clear whether Bayes himself was a Bayesian.[2]

[2] **Reading**: *"Who Discovered Bayes' Theorem"*, Stephen M. Stigler, The American Statistician, Nov. 1983, vol. 37, n° 4.

Axiomatic Definition of Probability

In 1933, Andrei Kolmogorov devised an axiomatic definition of probabilities, which borrows heavily from set theory. This is the definition one finds in most textbooks. It turns out that the very same rules govern operations on logical propositions (Boolean algebra), given a judicious choice of notation, so what we say for sets will apply equally well to propositions. We start be defining some basic terms:

Elementary set theory

Sample space aka "Universe" or "Compound event". Denoted by Ω, this defines all the possible outcomes of the experiment. For example, if the experiment is tossing a coin, $\Omega = \{\text{head, tail}\}$. If tossing a single six-sided die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. For some kinds of experiments, there may be two or more plausible sample spaces available. For example, Temperature = any number from -60°C to 60°C, Precip from 0 to 10 m/day.

Events Any subset of the Universe. e.g. die = 4 or Temperature= 0°C. *Mutually exclusive* events are events that cannot occur simultaneously, *e.g.* temperature < -20 °C and liquid rain. In a Venn diagram, these would correspond to non-overlapping disks.

Figure 2.11: Andrei Kolmogorov, Russian mathematician (1903 – 1987).

For any event E,

$$E \cup \Omega = \Omega, \tag{2.8a}$$
$$E \cap \Omega = E. \tag{2.8b}$$

Operations We have seen the basic operations of complement, union and intersection in Sect. I. To illustrate, let us cast a die and denote the events $A = \{$"Outcome is an even number "$\}$ and $B = \{$" Outcome > 3 "$\}$.

Notation	English translation	Numeric Outcome
$A \cap B$	both A & B	$\{4,6\}$
$A \cup B$	either A or B	$\{2,4,5,6\}$
$\overline{A \cap B}$		$\bar{A} \cup \bar{B} = \{1,2,3,5\}$
$\overline{A \cup B}$	neither A nor B	$\bar{A} \cap \bar{B} = \{13\}$

Note: both operators are commutative (*e.g.* $A \cap B = B \cap A$) and distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Partition A **partition** is a subdivision of Ω made of events $\{\omega_i, i = 1 \cdots n\}$ that verify:

- $\forall i, \omega_i \neq \emptyset$ (non-emptiness)
- $\forall (i,j)/i \neq j, \quad \omega_i \cap \omega_j = \emptyset$ (mutual exclusivity)
- $\bigcup_{i=1}^{n} \omega_i = \Omega$ (complete coverage)

One such partition is represented in Fig. 2.12.

Figure 2.12: **Partition of a Set**. Credit : wikipedia.

Probability Axioms

Given this lingo borrowed from set theory, Kolmogorov defined a probability $\mathbb{P}$ using the following axioms:

Positivity $\forall E, \mathbb{P}(E) \in \mathbb{R}$ and $\mathbb{P}(E) \geq 0$

Unitarity $\mathbb{P}(\Omega) = 1$

Additivity if $E_1, E_2, \cdots$ are mutually exclusive propositions, then $\mathbb{P}(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$. [3]

It turns out that these axioms produce a probability that has exactly the same properties as the logical desiderata of Pólya and Cox. So why the fuss? First, you should feel an immense relief: in designing a thinking robot out of logical principles, we wind up with exactly the same

[3] As a special case, we have the much simpler finite additivity $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$. We could have started from that point, but it turns out that some strange things may happen when trying to generalize this finite additivity to countable additivity (Borel-Kolmogorov paradox), so this complicated definition is warranted. *Jaynes* (2004) contends, however, that if we abide by Cox's principles and apply them carefully, we never run into those kinds of embarrassing paradoxes.

operating principles that mathematicians have imposed on probability for a long time. It is not completely a coincidence, of course, since Kolmogorov intended his axioms to provide a rigorous foundation for the use of probabilities in theory and applications. What is remarkable, however, is that his understanding of probability was rooted in set and measure theory, while Pólya and Cox's viewpoint is rooted in logic and physics. That the two give the same answer is quite remarkable, and means that we can apply the usual rules of probability calculus with the reified understanding that they describe *a way to reason about logical propositions*, and are not limited to games of chance or repeated events. Further, in the limit of certainty (probabilities approaching unity), one can show that the principles of deductive logic dear to Aristotle are recovered as special cases of the more general **probability theory as extended logic**. Satisfied that we are about this mind-blowing realization, let us quit philosophizing and start computing.

III The Calculus of Probabilities

Basic Properties

A direct consequence of these axioms are following properties:

$$A \subseteq B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B) \tag{2.9a}$$

$$\forall A \quad \mathbb{P}(A) \in [0,1] \tag{2.9b}$$

$$\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A) \tag{2.9c}$$

$$\mathbb{P}(\varnothing) = 0 \tag{2.9d}$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \text{ (sum rule)} \tag{2.9e}$$

With three sets (Fig. 2.13): $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$.

More generally,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \left((-1)^{k-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\right) \right) \tag{2.10}$$

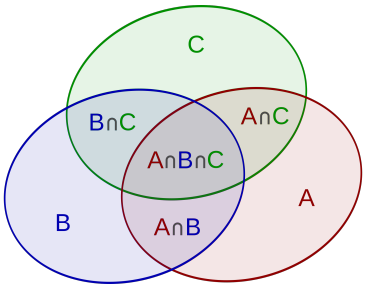


Figure 2.13: Principle of inclusion-exclusion illustrated on three overlapping sets. Credit: Wikipedia

Conditional Probabilities

$$\begin{aligned} \mathbb{P}(A|B) &= \text{Probability of } A \text{ given that } B \text{ occurred} \\ &\equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} . \end{aligned} \tag{2.11}$$

This is equivalent to:

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) \quad \text{(Product rule)}} \tag{2.12}$$

Which is more intuitive; it says that the probability of A and B both being true is the probability of A assuming B is true, times the probability that B is true. In fact, because $\cap$ is a commutative operation, the reverse applies, so we also have:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \tag{2.13}$$

This symmetry is the basis of Bayes' rule[4].

[4] Which, once again, may or may not have been enunciated by Bayes, but we call it that way regardless

Independence

Two events A and B are said to be independent if and only if:

$$\mathbb{P}(A|B) = \mathbb{P}(A) \qquad \& \qquad \mathbb{P}(B|A) = \mathbb{P}(B) \tag{2.14}$$

Equivalently,

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)} \tag{2.15}$$

This means that knowing B makes no difference about our state of knowledge about A, and *vice versa*. A more advanced concept is that of *conditional independence*. Two events A and B are said to be conditionally independent given C if and only if:

$$\mathbb{P}((A \cap B)|C) = \mathbb{P}(A|C) \cdot \mathbb{P}(B|C) \tag{2.16}$$

Example
We roll two dice:

$$X_1 = \textit{result of } 1^{\text{st}} \textit{ die}. \qquad X_2 = \textit{result of } 2^{\text{nd}} \textit{ die}.$$

1. *Given the events:*

$$A : \textit{sum } > 8, \qquad B : X_1 = 6.$$

We want to compute:

$$\mathbb{P}(\textit{sum} > 8 \,|\, X_1 = 6) = \mathbb{P}(X_1 + X_2 > 8 \,|\, X_1 = 6)$$

We can compute the intersection:

$$A \cap B = X_1 + X_2 > 8 \textit{ and } X_1 = 6 = \{(6,3), (6,4), (6,5), (6,6)\},$$

and by definition Eq. (2.11):

$$\mathbb{P}(A \cap B) = \frac{4}{36}, \qquad \mathbb{P}(B) = \frac{6}{36},$$
$$\Rightarrow \qquad \mathbb{P}(A \mid B) = \frac{4/36}{6/36} = \frac{4}{6} = \frac{2}{3}.$$

Equivalently, we could recognize that once $X_1 = 6$, there are only 4 ways to make the sum greater than 8 ($X_2 = (3,4,5,6)$), so the probability is 4/6, hence 2/3. We find the second way faster and more intuitive, but they are tantamount.

2. $\mathbb{P}(X_1 = 2 \cap X_2 = 3)$ *The events are independent:*

$$\begin{aligned} \Longrightarrow \qquad & \mathbb{P}(X_1 = 2 \cap X_2 = 3) = \\ & = \mathbb{P}(X_1 = 2) \cdot \mathbb{P}(X_2 = 3) \\ & = \frac{1}{6} \times \frac{1}{6} \\ & = \frac{1}{36}. \end{aligned}$$

Law of Total Probabilities

Assume $\{\omega_1, \ldots, \omega_n\}$ is a partition of Ω. If E is some event, then via Eq. (2.8b):

$$\mathbb{P}(E) = \mathbb{P}(E \cap \Omega)$$

But $\Omega = \bigcup_{i=1}^{n} \omega_i$ so:

$$\mathbb{P}(E) = \mathbb{P}\left(E \cap (\bigcup_{i=1}^{n} \omega_i)\right)$$

And by definition of conditional probabilities:

$$\mathbb{P}(E \cap \omega_i) = \mathbb{P}(E|\omega_i) \cdot \mathbb{P}(\omega_i)$$

But,

$$(E \cap \omega_i) \cap (E \cap \omega_j) = \emptyset, \qquad (i \neq j) \qquad \text{(Incompatible events)} \quad (2.17)$$

$$\begin{aligned} \Longrightarrow \qquad \mathbb{P}(E) &= \mathbb{P}\left(E \cap \left(\bigcup_i \omega_i\right)\right) \\ &= \mathbb{P}((E \cap \omega_1) \cup (E \cap \omega_2) \cup \ldots \cup (E \cap \omega_n)) \\ &= \mathbb{P}(E \cap \omega_1) + \mathbb{P}(E \cap \omega_2) + \ldots + \mathbb{P}(E \cap \omega_n) \\ &= \mathbb{P}(E|\omega_1) \cdot \mathbb{P}(\omega_1) + \ldots + \mathbb{P}(E|\omega_n) \cdot \mathbb{P}(\omega_n) \\ &= \sum_i \mathbb{P}(E|\omega_i) \cdot \mathbb{P}(\omega_i) \end{aligned}$$

Hence:

$$\boxed{\mathbb{P}(E) = \sum_i \mathbb{P}(E|\omega_i) \cdot \mathbb{P}(\omega_i)} \qquad \text{(Law of Total Probabilities)} \quad (2.18)$$

This is very useful for computing probabilities by considering different events covering all possibilities. What's almost magical about it is that we may choose any partition we'd like, so we can pick one that makes the conditional probabilities easy to evaluate. Then it is just a matter of adding and multiplying.

Note that the above formula bears more than a passing resemblance to Eq. (B.15). This is not a complete coincidence: you may think of a

partition as a sort of "basis" in a probability space. Thus, if you know the probabilities of each partition and can express the probabilities of each event in terms of the partition, you are done.

Example
5 *jars containing white (w) and black (b) balls.*

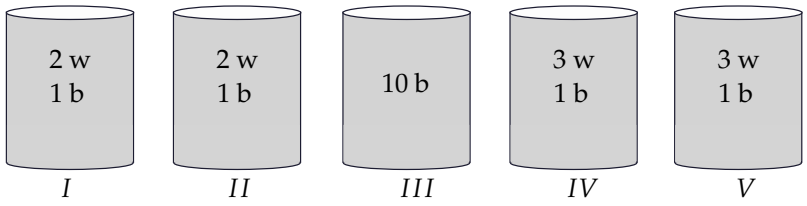


Figure 2.14: Illustrating the experiment of picking a random ball from 5 jars containing white and black balls.

1. *Pick a jar at random* ($P = \frac{1}{5}$).
2. *We draw a ball at random in the chosen jar.*
3. *Drawing a black ball.*

What is the probability of getting a black ball? To answer this, we use the law of total probabilities.

Let $E_i = \{$ *Choosing the* i^{th} *jar* $\}$ $(i = 1, 2, \ldots, 5)$*, where* $\{E_1, \ldots, E_5\}$ *is a partition.*

Now, $\mathbb{P}(E_i) = \frac{1}{5}$ *by hypothesis, and:*

$$\mathbb{P}(B|E_i) = \textit{Probability of drawing a black ball from } E_i$$
$$= \begin{cases} 1/3\,, & I, II\,, \\ 1\,, & III\,, \\ 1/4\,, & IV, V\,. \end{cases}$$

$$\overset{\textit{law of total prob.}}{\Longrightarrow} \qquad \mathbb{P}(B) = \sum_{i=1}^{5} \mathbb{P}(B|E_i) \cdot \mathbb{P}(E_i) = \frac{13}{30}\,.$$

Bayes' Rule

Elementary Case

This is a somewhat obvious application of the definition of conditional probabilities:

$$\begin{aligned} &\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \ , \\ \implies \qquad &\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \end{aligned} \tag{2.19}$$

(Expresses $\mathbb{P}(A|B)$ as a function of $\mathbb{P}(B|A)$)

We obtain the simplest form of Bayes' rule:

$$\mathbb{P}(A|B) = \frac{\overbrace{\mathbb{P}(B|A)}^{\text{likelihood}} \cdot \overbrace{\mathbb{P}(A)}^{\text{prior}}}{\underbrace{\mathbb{P}(B)}_{\text{normalizing constant}}}$$

As we will see later, sometimes one of the two conditional probabilities is easier to compute than the other, so this allows to "switch the conditionals".

More generally, if $\{E_i\}$ is a partition, and $\mathbb{P}(E_i) > 0$. For any B:

$$\begin{aligned} &\mathbb{P}(E_i|B) = \frac{\mathbb{P}(E_i \cap B)}{\mathbb{P}(B)} \ , \\ &\mathbb{P}(E_i \cap B) = \mathbb{P}(B|E_i) \cdot \mathbb{P}(E_i) \ , \\ &\mathbb{P}(B) \overset{\text{law of tot. prob.}}{=} \sum_j \mathbb{P}\left(B|E_j\right) \cdot \mathbb{P}\left(E_j\right) \ , \\ \implies &\boxed{\mathbb{P}(E_i|B) = \frac{\mathbb{P}(E_i) \cdot \mathbb{P}(B|E_i)}{\sum_j \mathbb{P}\left(B|E_j\right) \cdot \mathbb{P}\left(E_j\right)}} \qquad \text{Bayes' Rule} \end{aligned} \tag{2.20}$$

Application: Medical Diagnostic

We want to know if a person chosen at random is affected by a disease using the result of a medical test. Define the events:

S : The person is sick

$\bar{S}$: The person is healthy

T : The test is positive.

$\bar{T}$: The test is negative.

Now, the test is not perfect: it could come back negative when the person is in fact sick ("false negative"), and come back positive while

the person is in fact healthy ("false positive"). We define:

$$\mathbb{P}(\bar{T}|S) \stackrel{\text{def}}{=} f_n \text{ (probability of a false negative)}$$
$$\mathbb{P}(T|\bar{S}) \stackrel{\text{def}}{=} f_p \text{ (probability of a false positive)}$$

We would like to know the probabilities:

$\mathbb{P}(S|T)$ Being sick when the test is positive.

$\mathbb{P}(\bar{S}|\bar{T})$ Not being sick when the test is negative.

To solve this, we can apply Bayes' theorem using the partition $\Omega = \{S, \bar{S}\}$ and we set $\mathbb{P}(S) = f$.

$$\mathbb{P}(\bar{S}|\bar{T}) = \frac{\mathbb{P}(\bar{T}|\bar{S}) \cdot \mathbb{P}(\bar{S})}{\mathbb{P}(\bar{T}|\bar{S}) \cdot \mathbb{P}(\bar{S}) + \mathbb{P}(\bar{T}|S) \cdot \mathbb{P}(S)} = \frac{(1-f_p)f}{(1-f_p)f + f_n(1-f)}.$$

and

$$\mathbb{P}(S|T) = \frac{(1-f_n)f}{(1-f_n)f + f_p(1-f)}.$$

We know from epidemiological studies that the frequency of occurrence of the disease f is one in 200. As we vary f_p for constant f_n:

- $\begin{cases} f_n = 1\% \\ f_p = 1\% \end{cases} \rightarrow P(S|T) = 0.33$
- $\begin{cases} f_n = 1\% \\ f_p = 0.5\% \end{cases} \rightarrow P(S|T) = 0.498$
- $\begin{cases} f_n = 1\% \\ f_p = 0.1\% \end{cases} \rightarrow P(S|T) = 0.83$

That is, as the rate of false positives drops, we get more and more confident about the diagnosis. Worryingly enough, study after study has found that only about 15% of medical doctors (the people whom we pay to interpret such test results) get this right[5]. Most have no clue how to conduct such a simple calculation, and end up "transposing the conditional" (giving $\mathbb{P}(T|S)$ instead of $\mathbb{P}(S|T)$). As the example shows, those will in general be very different, except in special cases (homework: what values of f_n and f_p are needed for the equality $\mathbb{P}(S|T) = \mathbb{P}(T|S)$ to hold? The answer should depend on their ratio).

A remarkable feature is how counterintuitive the results may be for what appear to be relatively small differences between f_p and f_n. This is a reminder that the rules of probability calculus must be applied carefully and thoroughly, lest one commit some grave mistakes.

[5] For this account and for a simple introduction to Bayes' theorem, see http://yudkowsky.net/rational/bayes

Bayes' rule is central to Bayesian inference, which we will touch on briefly in Chapter 5. In this day and age, Bayesians are slowly taking over the world, so having even a vague idea about this theorem can be life-saving. For more than a vague idea, read *Gelman et al.* (2013).

Finally, note that even though we applied Bayes' rule, we were quite happy to equate a proportion of samples in a population (that is, a frequency of occurrence) with a probability. The use of Bayes' rule does not have to make you a strict Bayesian, and in such a case the frequentist interpretation of probability is most convenient and defensible.

Exercise (Cloud Seeding)
The effect of cloud seeding on the production of damaging hail is investigated by seeding and not seeding an equal number of candidate storms. Suppose the probability of damaging hail from a seeded storm is 0.1, and the probability of damaging hail from an unseeded storm is 0.4. If a candidate storm has just produced damaging hail, what is the probability that it was seeded? [6] .

Exercise (Hearing the sea in a sea shell)
Listening to the sea in a seashell, what is the probability that you may be hearing the sea itself? (Fig. 2.15) .

Further reading

Google's Peter Norvig has a great computational essay on probability, all in Python. If you know Python, you will learn a lot about probability. If you know about probability, you will learn some Python. And if you know neither, you should learn a bit about both!

[6] Taken from *Wilks* (2011)

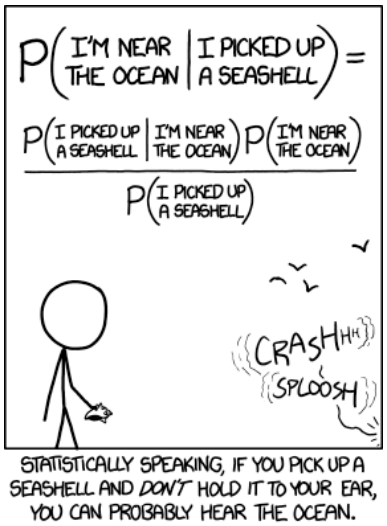


Figure 2.15: Bayes and sea-shells `http://xkcd.com/1236/`

Chapter 3

PROBABILITY DISTRIBUTIONS

> "If we look at the way the universe behaves, quantum mechanics gives us fundamental, unavoidable indeterminacy, so that alternative histories of the universe can be assigned probability.."
>
> *Murray Gell-Mann*

One message from quantum mechanics is that some quantities cannot be known with certainty (at least simultaneously), so the Universe must be described in the language of probabilities. In this chapter we introduce several commonly used probability distributions with broad applicability in the Earth Sciences. In each case we (briefly) introduce the necessary mathematical apparatus to derive their essential properties, and cite some examples of their use in Earth Sciences.

I Random Variables. Probability Laws

Random Variables

Definition 1 *A random variable (RV) is a function* $X(\Omega) \longrightarrow \mathbb{R}$*, which associates a real number to every event.*

$X(\Omega)$ is the image of Ω, the space of all possible values taken by X. In many cases it is a restricted subset of $\mathbb{R}$.

Example
Result of rolling a dice, amount of rainfall in a day, magnitude of an Earthquake, age of a rock, etc.

Random variables provide the link between the oh-so-esoteric "sample space" and the every day world of matter-of-factly measurements. They are the basis for applying probability theory to laboratory or field data, whether or not they were generated by a truly random process. [1] However, there is some order to this randomness: RVs can only take values within a certain interval, and do so according to a *probability law*, which we call a distribution.

[1] This does not necessarily mean that "God is playing with dice", as Einstein famously said, but that we have *uncertain information* concerning the process. Even if you are a staunch determinist, it is hard to argue about the fact that measurements always carry some degree of uncertainty

Since we are dealing with real-world data (which only come in discrete form *e.g.* because of digitization), most observations are discrete, though continuous RV provide an incredibly useful lens through which to view some Earth processes. Note that an RV could be complex; it can also be a vector or scalar.

Distribution Functions

Any discrete random variable admits a probability mass function (PMF) f defined by, $\forall x_i \in X(\Omega)$:

$$f(x_i) = P(X = x_i) \tag{3.1}$$

which describes the "weight" of each x_i in the total outcome. The PMF is often presented as a histogram, binning outcomes together in relatively coarse chunks.

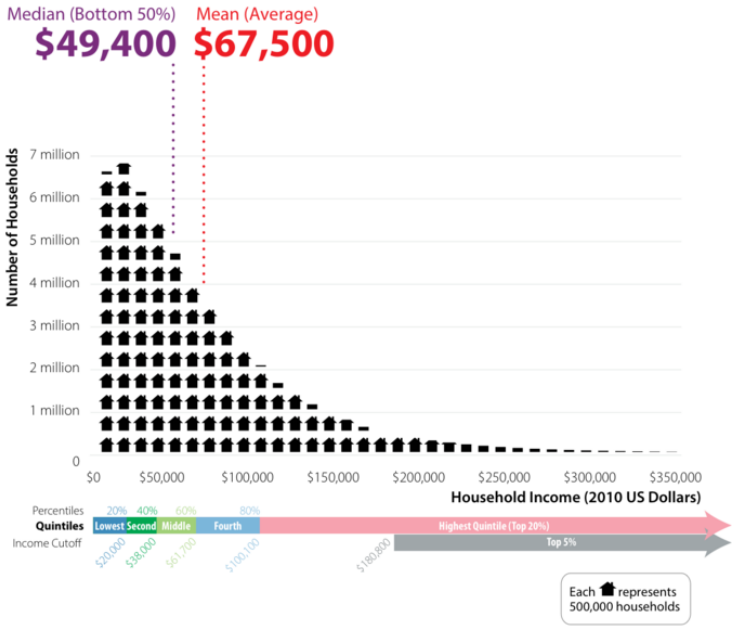


Figure 3.1: US income distribution (2010 data), as represented on http://visualizingeconomics.com/. Several points are apparent: the distribution peaks around $20K, but the median income is $49,400, and the mean is $67,500: this is typical of a *skewed distribution*. Skewness is the mathematical way of describing what economists call inequality: the top 1% earners famously control about 40% of the wealth (see http://youtu.be/QPKKQnijnsM). This is also a *heavy-tailed* distribution.

An example is the US income distribution of Fig. 3.1. How such a distribution may be estimated from observations will be discussed next; how it can be described by numbers or summarized by graphs, will be discussed in Sect. II. An associated definition is the Cumulative Distribution Function (CDF):

$$\forall x \in \mathbb{R}, F(x) = P(X \leq x) = \sum_{i=1}^{N} P(X = x_i) = \sum_{i/x_i \leq x} f(x_i). \tag{3.2}$$

so it is the sum of the PMF up until the largest x_i such that $x_i \leq x$.

A related definition is the so-called *survivor function*

$$P(X \geq x) = 1 - F(x) \tag{3.3}$$

(thus named by biomedical statisticians looking at survival times of patients under a pharmaceutical treatment). An example of survivor function that is foundational to seismology is the Gutenberg-Richter law (Fig. 3.2).

Properties:

- F is always increasing, and piecewise continuous (not differentiable at each x_i).
- $F(x) = 0$ for $x < x_{\min}$, $F(x) = 1$ for $x \geq x_{\max}$
- The CDF and PDF can be obtained from each other easily (by sum or differentiation).

Because of the first property, F is a bijection[2], so it can be inverted. This means that for every F one can find an inverse F^{-1} such that, for any (x, y) such that $y = F(x)$:

$$F^{-1}\left(F(x)\right) = x \quad \text{and } F\left(F^{-1}(y)\right) = y \tag{3.4}$$

F^{-1} is called the *quantile function*, often denoted by the letter Q. It is how things like confidence intervals and percentiles are obtained.

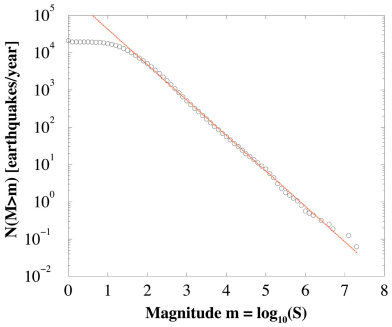


Figure 3.2: Earthquake magnitude distribution showing a power-law behavior over 6 decades. The Y axis represents the probability of observing an Earthquake magnitude greater than m (the X axis), in log scale. The graph follows $\log_{10} N(M > m) \propto -bm$, where b is the Gutenberg – Richter exponent $b = 1$ (dashed red line). The roll-off for $m < 2$ is due to difficulties with detecting very small earthquakes. From http://www.pnas.org/content/99/suppl_1/2509.short

[2] A one-to-one mapping: *i.e.* for each y there is only one x such that $y = F(x)$. In other words, F allows to travel from x to y and *vice versa*, uniquely.

Probability Density Function

So far we have focused on RVs taking discrete values (*i.e.* values that can be counted). An important class of variables are however continuous. In a continuous medium (like air, water, the Earth's mantle, etc) all variables take by definition a continuum of values over some interval $[a; b] \subset \mathbb{R}$, so their CDF $F(x)$ is generally a smooth function of x: it can be differentiated, yielding the *probability density function* (PDF) $f(x) = F'(x)$, which is analogous to the PMF. A PDF verifies 3 properties:

1. f is continuous
2. $\forall x \in \mathbb{R}$, $f(x) \geq 0$
3. f has unit mass: $\int_{-\infty}^{+\infty} f(x)dx = 1$. (Appendix A, Sect. III).

Indeed, one can write $F(x) = \int_{-\infty}^{x} f(u)du$, from which it follows that:

- F is continuously differentiable (being the integral of a continuous function)
- F is monotonically increasing (since its derivative $f(x)$ is positive)

- F goes from 0 to 1:

$$\lim_{x\to-\infty} F(x) = 0; \quad \lim_{x\to+\infty} F(x) = 1;$$

F is also called the cumulative distribution function, and is exactly equivalent to the CDF of a discrete random variable. In fact, one definition of a discrete RV is having a staircase-like CDF (with jumps at each x_i), while continuous RVs have smooth, continuous CDFs.

While the CDFs are equivalent for discrete and continuous RVs, there is, however a crucial difference between the PMF and the PDF: by definition, the probability that X takes values between x_0 and $x_0 + \delta x$ is $F(x_0) - F(x_0 + \delta x)$. But we have just seen that F is continuously differentiable, and its derivative is f, so, applying an order one Taylor expansion[3]:

[3] Appendix A, Sect. III

$$\mathbb{P}(x_0 \leq X \leq x_0 + \delta x) = f(x_0)\delta x + \mathrm{O}\big((\delta x)^2\big) \tag{3.5}$$

For δx small enough, this probability is close to $f(x_0)\delta x$: this quantity is the probability of X being in a neighborhood of x_0 of length δx. Notice, however, that as δx approaches 0, this probability approaches zero as well, since f is well-behaved.

Thus we have the mind-altering result that for a continuous random variable, $\mathbb{P}(X = x_0) = 0$ (that is, its PMF is zero at every point)! Does this mean that the variable can never reach this value? No: what it means is that we can never know this value with absolute certainty, so the probability of *exactly* observing the value x_0 is zero; the probability of observing something close to x_0 is finite, however. In other words, there will always be some uncertainty about our knowledge of X, but thankfully we can get close enough to it for practical purposes[4].

[4] JOKE: A mathematician and a physicist agree to a psychological experiment. The mathematician is put in a chair in a large empty room and a very tasty cake is placed on a table at the other end of the room. The psychologist explains, "You are to remain in your chair. Every five minutes, I will move your chair to a position halfway between its current location and the cake." The mathematician looks at the psychologist in disgust. "What? I'm not going to go through this. You know I'll never reach the cake!" And he gets up and storms out. The psychologist makes a note on his clipboard and ushers the physicist in. He explains the situation, and the physicist's eyes light up and he starts drooling. The psychologist is a bit confused. "Don't you realize that you'll never reach it exactly?" The physicist smiles and replied, "Of course! But I'll get close enough for all practical purposes!"

Empirical Determination of a Distribution

Imagine $X = \{x_1, x_2, \cdots, x_n\}$ is a collection of measurements. How do we find its distribution?

Histograms

The simplest strategy is to bin in different (usually regularly spaced) classes of x: how many between -5 and -4? how many between -3 and -2? And so on. We plot this frequency of occurrence as a function of the midpoint of each interval, say, and kaboom: this is the famed histogram[5].

[5] `np.histogram()`

For instance, we plot in Fig. 3.3 the empirical PMF of precipitation in Boulder, CO. 50 bins were chosen, which seems reasonable given the large number of observations (1192). However, perhaps the little ups and downs are artifacts of choosing such a large number of bins, and

would disappear if we chose a coarser bin size. This is tremendously important, because one might be tempted to interpret those peaks, whereas in fact they could arise from chance alone, simply because of low sample numbers in a few bins. Remember that the choice of the number of bins is subjective and one always has to try several[6]. Lab #3 will have you investigate this granularity issue in detail.

[6] By default, `np.histogram()` uses 10 bins, regardless of sample size

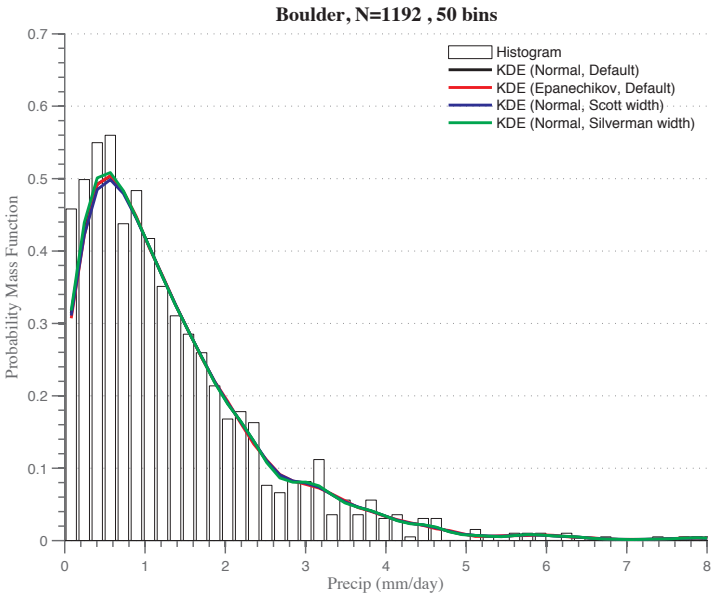


Figure 3.3: Empirical determination of the probability mass function of precipitation in Boulder, CO. Data: GHCN.

Kernel Density Estimation

In the case of precipitation, $X(\Omega)$ is the space of positive real numbers ($\mathbb{R}^+$), so we might want to estimate a continuous distribution (that is, a *probability density function*), rather than a PMF. How do we do this from discrete observations? Kernel Density Estimation achieves this by using smoothing functions called *kernels*:

$$\mathrm{KDE}(X) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.6}$$

where K is a smooth function of unit mass with good localization, and h is the *bandwidth*, which controls how many neighboring points are being used to estimate f around x_i. The two main choices of kernel are:

$$K(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{(Epanechikov),} \\ \frac{1}{\sqrt{2\pi}} e^{-t^2/2} & \text{(Gaussian).} \end{cases} \tag{3.7}$$

(though there are plenty of other reasonable choices).

In the end, this choice doesn't matter nearly as much as the choice of the bandwidth h. A large h includes many points, so will lead to a very smooth estimate, which might gloss over important details. On the other hand, too small an h could lead a jagged PDF full of spurious spikes. As with everything in life, there is a happy Middle Path (Fig. 3.4).

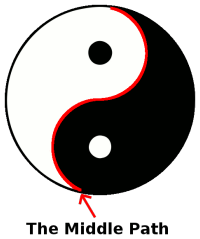


Figure 3.4: One representation of the Middle Path

Choice of initial smoothing parameter: Two main methods on the market, used by most codes out there:

Silverman (1986) $h = \min(0.9 * \sigma, 0.666 * \mathrm{IQR})/n^{1/5}$

Scott (1992) $h = c * \mathrm{IQR}/n^{1/3}$, where $c \in [2; 2.6]$ (2.6 for Gaussian kernels, smaller otherwise)

Let us emphasize that these are **initial** estimates and that a proper determination will always involve trial and error to see whether this choice mattered. Luckily for you, all of this is beautifully coded up in the amazing seaborn package, which we'll abbreviate as `sns`.

Quantiles and Percentiles

Often we are interested in the values below which lie a certain fraction of the mass of a certain distribution. If you've ever spoken of median income or the percentile of your GRE score, then you are already familiar with the notion. The formal way of obtaining this is through the inverse CDF, F^{-1}, aka the *quantile function*[7]. More precisely, we define the α-quantile q_α the number such that:

$$\mathbb{P}(X \leq q_\alpha) = \alpha \tag{3.8}$$

That is, $q_\alpha = F^{-1}(\alpha)$. Unless the inverse CDF can be obtained analytically (which can prove impossible even for usual distributions), it is approximated using numerical root finding methods (*cf.* Appendix A, Sect. III). Going back to Fig. 3.1, how many people earn less than $225,000 a year? About 98%. Below what income do 98% of Americans fall? About $225,000. The first answer used the CDF, the second used the quantile function. The values q_α are called, not coincidentally, **quantiles**. Remarkable quantiles are:

the median the 50% quantile ($q_{0.50}$), such that 50% of the mass of the distribution is to the left of it, 50% of it on the right of it.

terciles split the distribution into 3 regions of equal mass: $(0, 1/3, 2/3, 1)$

quartiles split the distribution into 4 regions of equal mass (0, 25%, 50%, 75%,100%)

deciles idem with 10 regions: $10\%, 20\%, \cdots, 100\%$

percentiles idem with 100 regions.

[7] Many choices to do this in Python, from the simple `np.percentile()` to the comprehensive `scipy.stats.mstats.mquantiles()`, modeled after the R `quantile` function, to internal implementations in `pandas` and `seaborn`, to name just a few.

Distribution Fitting

Often it is of interest to fit a known ("parametric") distribution to an empirical PMF. This can only be done after one has surveyed the data using the above techniques, and having noticed that the data follow a particular pattern. Alternatively, one may have *a priori* ideas of why the data should follow a certain distribution (*e.g.* a Gutenberg-Richter law), so it is interesting to plot that theoretical distribution on the same graph. In general, however, this is a topic of statistical estimation, which we will tackle in Chapter 5. The fitter Python package allows to scan up to 80 distributional candidates to see which one fits best.

Moments of a distribution

Expectance : the gambler returns

The expectance operator gives the expected value of an RV X:

$$E(X) = \mu_1 = \begin{cases} \sum_{i=1}^{N} x_i P(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x f(x) dx & \text{if } X \text{ is continuous} \end{cases} \tag{3.9}$$

This also called the moment of order 1, more commonly known as average, or mean. Indeed, if all outcomes are equally likely $P(X = x_i) = 1/N$, then this is just the arithmetic average of all observations. If the outcomes are not equally likely, they must be weighted by their probability before summation – it is therefore a *weighted mean*.
Key Properties:

Linearity $E(aX + bY) = aE(X) + bE(Y)$

Transfer theorem

$$E(g(X)) = \begin{cases} \sum_{i=1}^{N} g(x_i) P(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} g(x) f(x) dx & \text{if } X \text{ is continuous} \end{cases} \tag{3.10}$$

This last formula is incredibly useful to study error propagation: if the uncertainties in X are known, so are the uncertainties in any function of X. See Chapter 4 and Lab 4.

Higher order moments

By the previous theorem, one may define the moment of order n as:

$$m(X, n) = E(X^n) = \sum_{i=1}^{N} {x_i}^n P(X = x_i) \tag{3.11}$$

The moment of order 2 is closely related to the *variance*, obtained by setting $g(x) = (x - \mu)^2$. Hence:

$$E((X-\mu)^2) = \sum_{i=1}^{N} (x_i - \mu)^2 P(X = x_i) = E(X^2) - \mu^2 \tag{3.12}$$

(the last two expression are exactly analogous in the continuous case). The variance measures the *spread* of a distribution about its central tendency. One often speaks of the standard deviation, $\sigma = \sqrt{V(X)}$, which is the average deviation around the mean, and bears the same units as X[8].

[8] "Standardized data" refer to data having been centered to the mean and divided by their standard deviation

Properties:

$V(X + b) = V(X)$ "Shifting the mean does not shift the variance"

$V(aX) = a^2 V(X)$

$\sigma(aX) = a\sigma(X)$ "Rescaling X by a simply magnifies its excursions by the same amount"

The moment of order 3 measures departures from symmetry about the y axis ($x = 0$). It is closely associated with the *skewness*, defined as the third standardized moment:

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{3/2}} \tag{3.13}$$

Finally, the *kurtosis* is derived from the fourth moment, and quantifies the "peakedness" of a distribution. The list goes on, but no one ever seems to worry about moments beyond $n = 4$.

Note *Knowledge of all moments of X is equivalent to knowledge of the distribution itself (via the characteristic function).*

II Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data for the purpose of formulating hypotheses, complementing the tools of conventional statistics for testing hypotheses[9]. It was so named by *Tukey* (1977) to contrast with Confirmatory Data Analysis, the term used for the set of ideas about hypothesis testing, p-values, confidence intervals, etc. (which formed the key tools in the arsenal of practicing statisticians at the time, and which we will study in Chapter 6).

[9] For additional references, check out Andrew Zieffer's notes and Chapter 3 of *Wilks* (2011)

Often in EDA, one looks for methods that are both *robust and resistant*. Robustness is defined as the insensitivity to assumptions about the data (e.g. 'the data are normally distributed'). Resistance is the insensitivity to large excursions, i.e. outliers. In many cases, EDA is a "first pass" to take a gross look at the data, formulate hypotheses and then go on to apply more specific techniques.

Assume that you get a data vector $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, which we can view as a realization of a random variable X (i.e. we have uncertain knowledge about the process that generated the data). Where do we start and where do we go from there? The first step would be to plot the data themselves in a sensible way (i.e. as a function of a related variable, like the depth of collection, time, distance along a transect and so on). The second would be to obtain numerical summaries of its distribution, then estimate the distribution itself, and finally plot it in fancier ways.

Numerical Summaries

Range

The *range* is simply the difference between the largest and smallest value, and bears the same units as X. The *dynamic range* is usually defined as:

$$\text{DR} = 20 \times \log_{10}\left(\frac{\max(\boldsymbol{x})}{\min(\boldsymbol{x})}\right) \tag{3.14}$$

expressed in decibels (dB) regardless of the units of $\boldsymbol{x}$. For instance, the dynamic range of human hearing is roughly 140 dB.

The range immediately gives us a crude idea of the size of variations to expect, which can be very precious already. For instance, are they of the expected order of magnitude? Does the range encompass 0?

Location

We wish to locate the centroid of the distribution, its *central tendency*, *i.e.* its center of mass on the x axis. The sample mean allows to estimate this simply:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.15}$$

which is simply (3.9) with $P(X = x_i) = 1/N$. This is called a uniform distribution and means that each outcome is equally likely to occur. The formula turns out to be valid even for more general distributions. We shall see later why the sample mean is the optimal estimator of the true mean ("Maximum Likelihood") for Gaussian random variables.

However the sample mean is neither robust nor resistant. Robust measures of location include:

the median: $MED(X) = q_{50\%}$. A different mean and median is a sure sign of skewness.

the trimean: $TM = \frac{q_{25\%}+2\times q_{50\%}+q_{75\%}}{4}$, a weighted mean of the quartiles.

the trimmed mean: a version of the sample mean ignoring the most extreme values.

For more information, read *Wilks* (2011), Chapter 3.

Spread

The most common measure of spread, as discussed above, is the sample standard deviation s:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.16}$$

However, it is neither robust nor resistant (due to the presence of squares, or put another way, the ℓ_2 norm), and may be easily fooled by outliers. Better measures include the:

interquartile range: $IQR = q_{0.75} - q_{0.25}$, which complements the trimean.

mean absolute deviation: $MAD = E\left(|X - q_{0.50}|\right)$, which uses the ℓ_1 norm.

Symmetry

The sample skewness is defined as:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}} \tag{3.17}$$

But a robust and resistant measure is the Yule-Kendall skewness:

$$\text{YKS} = \frac{q_{0.75} + 2q_{0.50} - q_{0.25}}{\text{IQR}} \tag{3.18}$$

To illustrate how these three properties may evolve in warming climate aspects of changing mean, variance and skewness of temperature distributions are presented in Fig. 3.5.

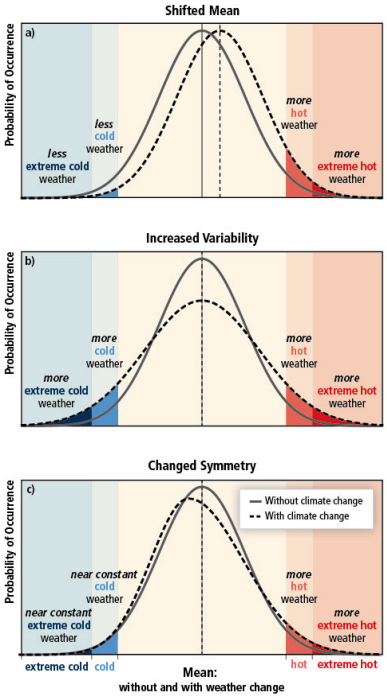


Figure 3.5: Location, scale and symmetry of temperature distributions in a changing climate, from http://ipcc-wg2.gov/SREX/report/

Graphical Summaries

"Numerical quantities focus on expected values, graphical summaries on unexpected values". John Tukey

A picture is worth a thousand words. We've already seen in Fig. 3.3 how one might represent a PMF, and if appropriate, a PDF. Here are a few strategies to display such distributions. One way to do this is to plot distribution quantiles as shaded bands, as in Fig. 3.7. This is a great way to chart evolution over time and get a general impression, though it can become difficult to see exactly what happens to the distribution. Boxplots (Fig. 3.8) and violin plots (Fig. 3.10) are more useful in this regard.

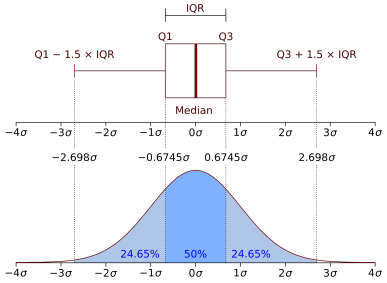


Figure 3.6: Boxplot representation of a normal distribution, illustrating how economically the device can convey the location of various quantiles of interest.

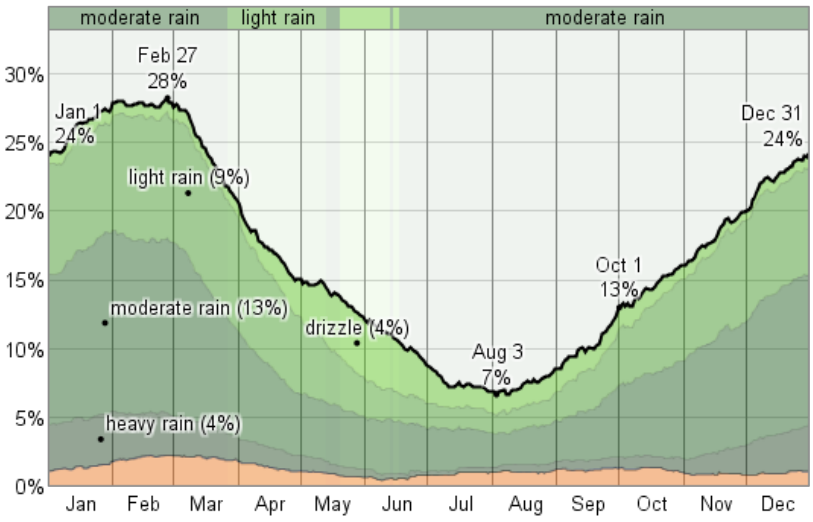


Figure 3.7: Historical probability of precipitation over Los Angeles at some point in the day, expressed in %. Orange represents thunderstorm-related precipitation, which is extremely rare in L.A. Credit: weatherspark

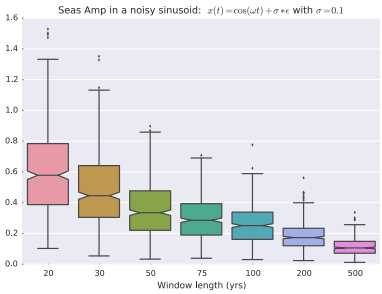


Figure 3.8: Boxplots for grouped observations (here, illustrating convergence as a function of sample size). Note the notch to denote the median's location.

Boxplots

Boxplots are compact ways to represent a distribution, focusing on its quartiles (body), 95% mass (whiskers) and occasionally outliers. An example is shown in Fig. II, showing how the normal distribution may be summarized very succinctly in this fashion. Boxplots are sometimes the only practical way of showing how several distributions compare (Fig. 3.8). Whenever possible, thou shalt notch your boxplot.

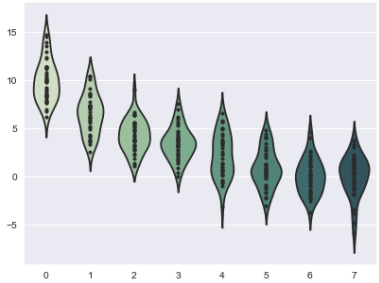


Figure 3.9: A simple violin plot, using `sns.violinplot()`. Credit: seaborn example gallery

Violinplots

Let's face it: boxplots are pretty boxy. A more refined way of plotting distributions are *violin plots*, which elegantly summarize their shape as quantified via kernel density estimates, along with the data as points (Fig. 3.10) or bars ("bean plots"). Good implementations come with a number of customizable options, including the ability to perform side-by-side comparisons as split violins, which must sound awful, but look rather snazzy.

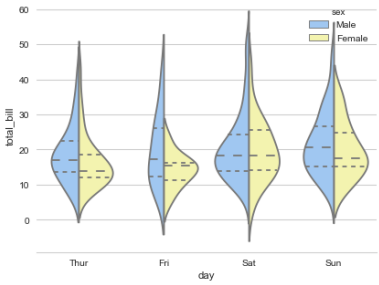


Figure 3.10: A split violin plot. Notice the median and quartiles which allow to quickly compare these quantiles among paired datasets. Credit: seaborn example gallery

Scatterplots

Jumping ahead to part III, we sometimes want to explore the relationship between two variables. One expeditious way to do so is to draw a *scatterplot* of one variable versus the other; even better is to plot their histograms along using seaborn's `sns.jointplot()`, as done in Fig. 3.11.

This is not a bad time to introduce the idea of correlation.

Correlation and covariance

How shall we quantify the relationship between two variables X and Y, such as those of Fig. 3.11? Covariance and correlation are popular measures of association for this purpose. Covariance is defined as:

$$\mathrm{Cov}(X,Y) = E\left[(X - E(X))\,(Y - E(Y))\right] \tag{3.19a}$$

$$= E(XY) - E(X)E(Y) \tag{3.19b}$$

In particular, $\mathrm{Cov}(X,X) \equiv V(X)$ (the variance of X is its covariance with itself). If the two variables are independent, then $\mathrm{Cov}(X,Y) = 0$ (as a consequence of $E(XY) = E(X)E(Y)$).
In order to remove the dependence on the units of X and Y we can normalize by the standard deviations σ_X and σ_Y. We define the *correlation coefficient* as:

$$\rho_{XY} = \mathrm{Cov}\left(\frac{X - E(X)}{\sigma_X}, \frac{Y - E(Y)}{\sigma_Y}\right). \tag{3.20}$$

It has the following properties:

- When $|\rho| < 1$ and $\rho > 0$, we say that the two variables are correlated. When $\rho < 0$ we say they are anti-correlated.
- If $|\rho| = 1$, there is a linear relationship between X and Y i.e. $X = aY + b$ or $Y = cX + d$., where a, b, c, d are constants.
- ρ only measures *linear* association; a non-linear relation may not get picked up). A famous example is $U = \cos X$, $V = \sin X$. Even though $U = \sqrt{1 - V^2}$, you can verify that $\rho(U,V) = 0$!
- ρ^2 represents the fraction of variance shared between the variables X and Y. It is common to speak of "variance explained" by one variable, but we discourage use of the term: if X and Y are correlated because they are both caused by a third variable Z, then what does X possibly *explain* about Y (or *vice versa*)? You should prefer the terminology "X accounts for $\rho^2 \times 100\%$ of variance in Y" (or *vice versa*). Causality aside, if $|\rho|$ is high, one may use Y as a proxy for X and vice versa.

The correlation coefficient for a sample can be estimated as

$$\hat{\rho} \equiv r_{XY} = \frac{\frac{1}{n-1}\sum_i^n (x_i y_i - \bar{x}\bar{y})}{S_X S_Y} \tag{3.21}$$

$$S_X^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2 \qquad S_Y^2 = \frac{1}{n-1}\sum_i^n (y_i - \bar{y})^2. \tag{3.22}$$

This estimate is also known as Pearson's product-moment correlation or Pearson's correlation. It is the most common measure of correlation but is neither robust nor resistant. For instance, this situation:

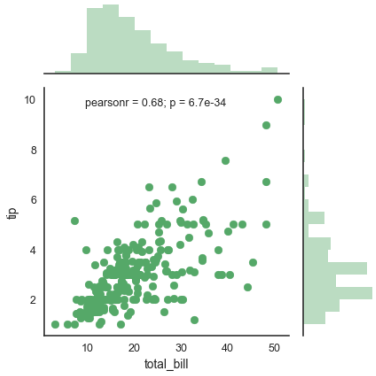


Figure 3.11: A scatterplot between two variables, along with histogram of marginal distributions. Generated using `sns.jointplot()`. Credit: seaborn gallery

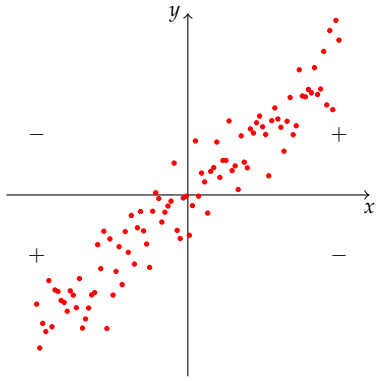
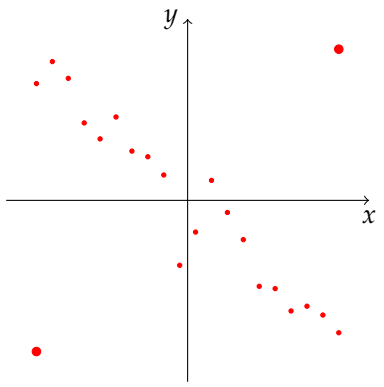


Figure 3.12: Correlation in a scatter plot. Eq. (3.19a) is equivalent to assigning a sign to each quadrant, and multiplying the distance from the origin $(E(X), E(Y))$ by this sign.



would give $\rho > 0$ even if the majority of points shows $\rho < 0$. As an alternative we can use Spearman's correlation[10] or Kendall's τ[11], which are based on rank order statistics, and are therefore more robust. Kendall's τ lends itself to better uncertainty quantification. The wider problem with correlations is how to interpret them. We will come back to this in Chapter 6, Sect. VII.

Fig. 3.13 displays three types of cases: linear relation with various amounts of scatter (top), perfect correlation (middle) and various ways of obtaining a zero linear correlation (bottom), even when nonlinear associations are extremely strong.

[10] `scipy.stats.spearmanr()`
[11] `scipy.stats.kendalltau()`

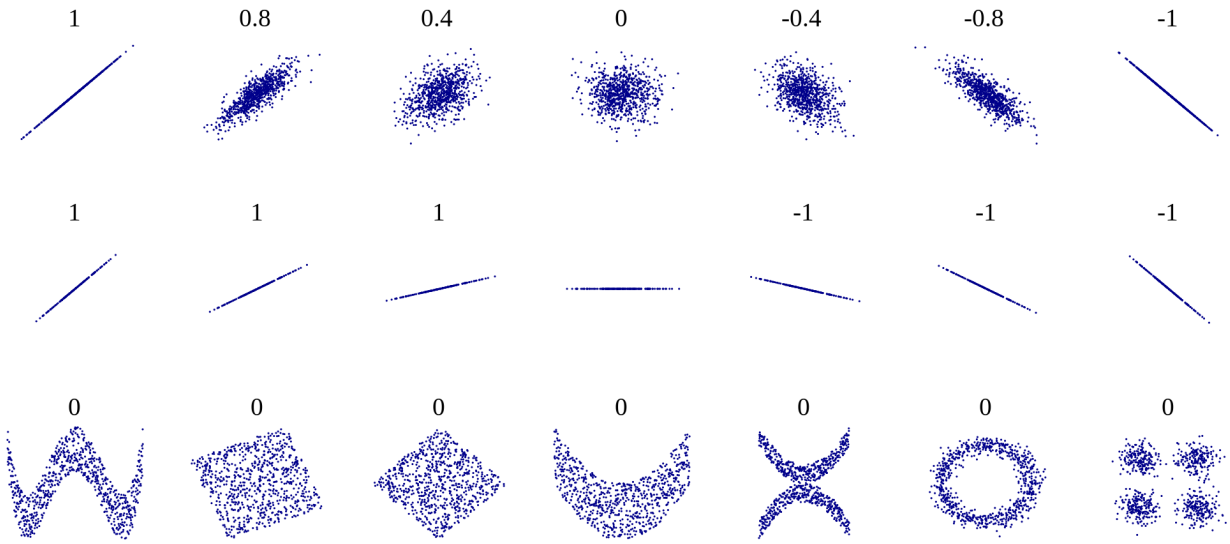


Figure 3.13: A taxonomy of correlations. (top) typical situations of a linear relationship of varying strength; (middle) perfect correlation: one is measuring the two variables twice, under different names; (bottom) examples of nonlinear associations yielding zero linear correlation. See Lab 6 for some practical flavors. Credit: Wikipedia

Bivariate densities

More generally, we are often interested in the joint probability distribution of several variables. As in 1 dimension, kernel density estimation can help estimate a continuous density from discrete observations, turning a cloud of points into a field of color (Fig. 3.14). Alternatively, the 2-dimensional analog of the histogram visualizes a bivariate PMF as hexagonal tiles (Fig. 3.15). Both figures were generated via `sns.jointplot()`.

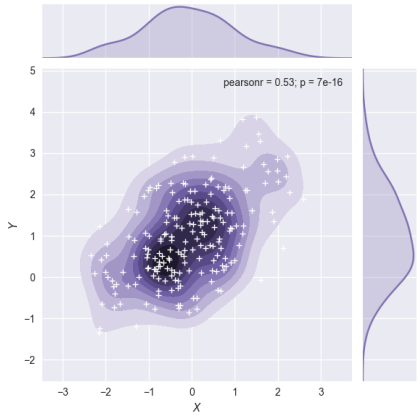


Figure 3.14: Joint plot of two variables, illustrating joint and marginal kernel density estimates, as well as individual observations as white crosses. Credit: Seaborn manual

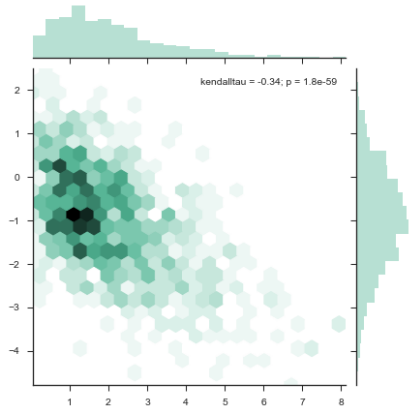


Figure 3.15: The bivariate analogue of a histogram is known as a "hexbin" plot, because it shows the counts of observations that fall within hexagonal bins. Credit: Seaborn manual

III Common Discrete Distributions

We now introduce some classic theoretical distributions, which are very useful in modeling natural phenomena. The underlying idea is that, in cases where little to no data are available, we might anticipate measurements to follow a certain theoretical distribution; if so, this knowledge will greatly help the assessment of uncertainties and the estimation of parameters of interest. These distributions form the basis of so-called "parametric statistics", because they are characterized by one or more *parameters*. Here we focus on the distributions themselves – we will reserve the difficult topic of statistical estimation (of said parameters) for Chapter 5.

Bernoulli Distribution

A Bernoulli random variable is the easiest way to encode a binary outcome: heads or tails? Rain or shine? Big earthquake or small earthquake? We say that a r.v. $X \sim \mathrm{B}(p)$ if and only if X takes the value 1 with probability p ("success") and 0 with probability $q = 1 - p$ ("failure"). Verify that $E(X) = p$, $V(X) = pq$.

The Bernoulli distribution is not very interesting in its own right, but serves as a building block for a very important distribution:

Binomial Distribution

A random variable $X \sim \mathcal{B}(n, p)$ is said *binomial* if $\forall k \in [1; n]$:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.23}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. The latter have the useful property that:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \quad \text{(Pascal's rule)} \tag{3.24}$$

The binomial distribution measures the probability of k successes amongst n trials. It is easy to show that $X \sim \mathcal{B}(n, p)$ can be obtained

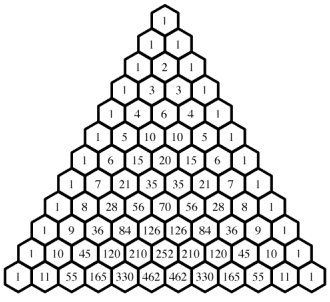


Figure 3.16: Pascal's Triangle, an illustration of Pascal's rule. Credit: Math Forum

as the sum of n *independent* Bernoulli trials with identical probability p. If follows that: $E(X) = np$, $V(X) = npq$[12]. Its PMF and CDF are shown in Fig. 3.17 (left column).

[12] For independent variables, the variance of the sum is simply the sum of the variances

Application *Many processes can be modeled as a sum of independent,* ***binary*** *events. For instance, if the probability of a winter frost day is 0.1, and there are 90 days, what is the probability of getting at least 10 frost days per winter? (assume a winter is 90 days long)*

$$\mathbb{P}(X=0) = \binom{90}{0} 0.1^0 (1-0.1)^{90} \simeq 7.6e^{-5} \tag{3.25a}$$

$$\mathbb{P}(X=1) = \binom{90}{1} 0.1^1 (1-0.1)^{89} \simeq 7.6e^{-4} \tag{3.25b}$$

$$\vdots = \vdots \tag{3.25c}$$

$$\mathbb{P}(X=9) = \binom{90}{9} 0.1^9 (1-0.1)^{81} \simeq 0.139 \tag{3.25d}$$

$$\mathbb{P}(X \geqslant 10) = 1 - \mathbb{P}(X < 10) = \sum_{k=0}^{k=9} \mathbb{P}(X=k) \simeq 0.412 \approx 41\% \text{ chance}$$

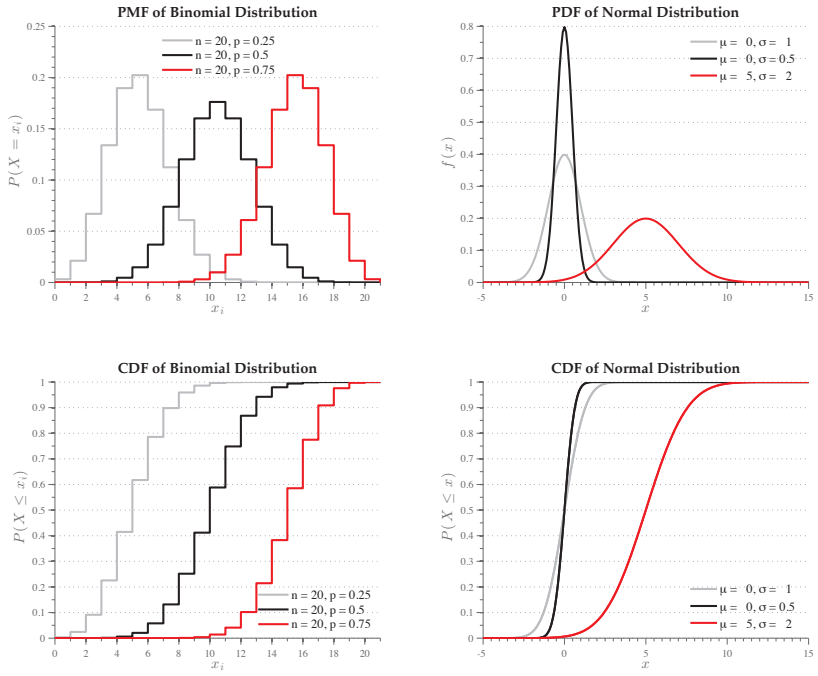


Figure 3.17: Binomial vs Normal distribution. Top: PMF or PDF. Bottom: CDF. Left: Binomial distribution with $n = 20$ and three values of p. Right: Normal distribution for 3 values of μ and σ.

Poisson Distribution

A random variable is said to be Poisson with rate $\lambda \in \mathbb{R}^+$ (usually written $X \sim \mathcal{P}(\lambda)$) if and only if, $\forall k \in \mathbb{N}$:

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \tag{3.26}$$

Proof of the fact that all probabilities sum to unity is provided via the exponential series: $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ (Appendix A, Sect. III). From this and a simple rearrangement of dummy indices, one can easily show that $E(X) = \lambda$ and $V(X) = \lambda$. For $\lambda \leq 1$, the PMF decreases strongly with k, hence the nickname: 'law of rare phenomena'.

A *Poisson process* is a collection of Poisson random variables. Examples that are well-modeled as Poisson processes include the radioactive decay of atoms, telephone calls arriving at a switchboard, page view requests to a website, or the number of US hurricane landfalls. If N_t is the number of such occurrences over a time interval t, then:

$$\mathbb{P}(N_t = k) = f(k; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \tag{3.27}$$

In particular, the number of events over two intervals Δt_1 and Δt_2 are independent, and the probability that the time between events exceeds Δt is $e^{-\lambda \Delta t}$, which decays exponentially fast to zero.

The Poisson distribution has all sorts of wonderful properties. In particular, it is easy to show that if $X_1 \sim \mathcal{P}(\lambda_1)$ and $X_2 \sim \mathcal{P}(\lambda_2)$, then the sum $X = X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$ (additivity of the two variables). The Poisson process is intimately tied to the **exponential** distribution, which describes *memoryless* processes.

Interesting property: the Binomial distribution can be approximated by a Poisson distribution when $n \to \infty$ if $\lambda = np$ is kept constant. In turn, it converges to a normal distribution for λ larger than ~ 5 (see Fig. 3.18, cyan curve, and Chapter 4, section II).

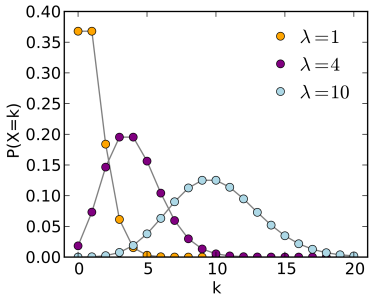


Figure 3.18: PMF of the Poisson distribution for 3 values of λ. Credit:Wikipedia

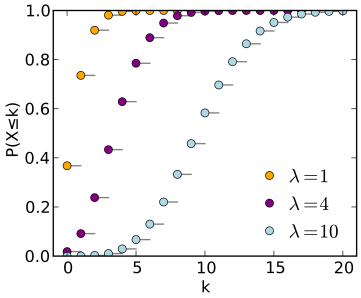


Figure 3.19: CDF of the Poisson distribution for 3 values of λ. Credit:Wikipedia

Geometric Distribution

The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, ...\}$. Also called "waiting time until the first success" . $X \sim \mathcal{G}(p)$ iff $\forall k > 1$:

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p \tag{3.28}$$

Again, this has no upper bound for large k, so we will need to sum until infinity. This is made possible by the geometric series:

$$\sum_{k=0}^{\infty} a^k = \frac{1}{1 - a} \quad \forall |a| < 1 \tag{3.29}$$

From this it follows that:

$$\mathrm{E}(X) = \frac{1}{p} \quad ; \quad \mathrm{var}(X) = \frac{1 - p}{p^2} \tag{3.30}$$

The fact that the expected time until the first success is inversely proportional to the probability of the event (a frequency of occurrence), should not be surprising. It is the basis for estimating "return periods" from the statistics of past events, a practice as fraught as it is misleading. Most of the public thinks of a "100-year flood" as a flood that recurs every 100 years, whereas it really is a flood that has a 1% chance of happening every year, assuming stationarity (Lab 5).

Application *Waiting time until you get a 6 by rolling a die, waiting time until any event with some estimated frequency (e.g. droughts, floods, earthquakes).*

IV Common Continuous Distributions

Normal Distribution

The most emblematic curve in all of statistics, the Normal (or Gaussian) distribution (Fig. 3.17) will in fact prove to be the *central* distribution towards which all the others gravitate. Its properties are so numerous, and its range of applications so vast, that we will here content ourselves with a very cursory description, but devote an entire Chapter (4) to normality.

Property	Value
support	$x \in \mathbb{R}$
pdf	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
cdf	$\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$
mean	μ
median	μ
mode	μ
variance	σ^2
skewness	0

Table 3.1: Essential Properties of the Gaussian Distribution

Gamma Distribution, $\Gamma(k, \theta)$

Very versatile distribution that allows to fit many different behaviors. Characterized by 2 real numbers k (shape parameter) and θ (scale parameter):

$$f(x; k, \theta) = \left(\frac{x}{\theta}\right)^{k-1} \frac{\exp(-x/\theta)}{\Gamma(k)\,\theta}$$

where $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt$ is the famous Gamma function (a generalization of the factorial function). As is apparent from Fig 3.20 and Fig 3.21, the qualitative behavior is radically different for different values of k . Only for $k \geq 1$ does the distribution start at zero, with a tail that can usefully approximate things like rainfall statistics. It is in fact one of the simplest distributions to represent skewed variables, since the skewness is solely constrained by k. For large k, it starts looking like a Gaussian (but then again, most things do). Two notable special cases are $k = 1$ (Exponential distribution) and $\theta = 2$ (Chi-squared Distribution), which we will see again.

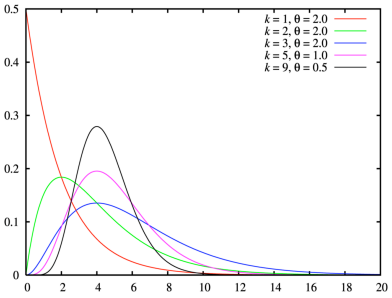


Figure 3.20: Gamma distribution (PDF)

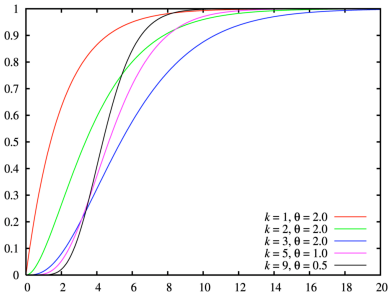


Figure 3.21: Gamma distribution (CDF)

Property	Value for Exponential Distribution
support	$x \in \mathbb{R}^+$
pdf	$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$
cdf	$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$
mean	$\frac{1}{\lambda}$
median	$\frac{\ln(2)}{\lambda}$
mode	0
variance	$\frac{1}{\lambda^2}$
skewness	2

Property	Value for Gamma Distribution
support	$x \in \mathbb{R}^+$
pdf	$x^{k-1}\frac{\exp(-x/\theta)}{\Gamma(k)\,\theta^k}$
cdf	$\frac{\gamma(k,x/\theta)}{\Gamma(k)}$
mean	$k\theta$
median	no simple closed form
mode	$(k-1)\theta$ for $k \geq 1$
variance	$k\theta^2$
skewness	$\frac{2}{\sqrt{k}}$

Table 3.2: Essential Properties of the Exponential and Gamma Distributions

Extreme Values: the Weibull Distribution

The Weibull is a special case in a broader class called *generalized extreme value* distributions. It is given here because of its simplicity and applicability in Earth Sciences, particularly for extreme events like upper ocean velocities, floods, hurricane wind speeds, and all manner of imaginable natural hazards. Non-surprisingly, it is a darling of the insurance and reinsurance industry. The Weibull distribution is also suitable for modeling particle size distributions, or the time before an instrument fails. The density writes:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.31}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution. Its complementary cumulative distribution function is a stretched exponential function. The Weibull distribution is related to a number of other probability distributions; it interpolates between the exponential distribution ($k = 1$) and the Rayleigh distribution ($k = 2$).

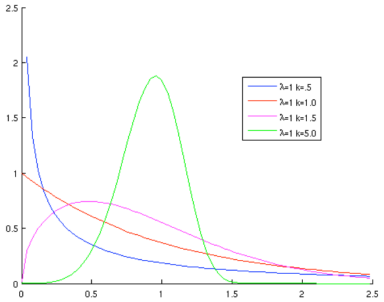


Figure 3.22: PDF of the Weibull distribution

Table 3.3: Weibull Distribution

Property	Value for the Weibull Distribution
support	$x \in [0; +\infty)$
pdf	$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$
cdf	$1 - e^{-(x/\lambda)^k}$
mean	$\lambda\Gamma\left(1+\frac{1}{k}\right)$
median	$\lambda(\ln(2))^{1/k}$
mode	$\lambda\left(\frac{k-1}{k}\right)^{\frac{1}{k}}$ if $k > 1$
variance	$\lambda^2\Gamma\left(1+\frac{2}{k}\right) - \mu^2$
skewness	$\frac{\Gamma(1+\frac{3}{k})\lambda^3 - 3\mu\sigma^2 - \mu^3}{\sigma^3}$

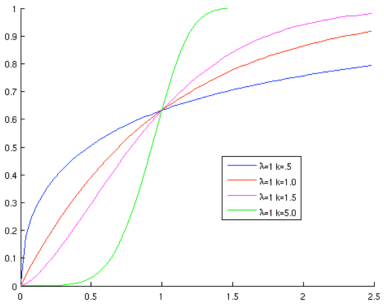


Figure 3.23: CDF of the Weibull distribution

In lab #5, we shall see how these distributions can be fit to geophysical and geochemical data and what it enables us to do.

Chapter 4

NORMALITY. ERROR THEORY

> Why have we for so long managed with the normality assumption?
>
> *George Barnard*

I The Normal Distribution

Distribution Functions

A random variable X is said to be normal, or Gaussian, with parameters μ and σ ($X \sim \mathcal{N}(\mu, \sigma^2)$) if and only if its Probability Density Function (PDF) verifies:

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (\text{“density"}). \tag{4.1}$$

Special Case: (the standard normal distribution): $\mu = 0,\ \sigma = 1$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{4.2}$$

The CDF (Cumulative Distribution Function) of $X \sim \mathcal{N}(0,1)$ is, generally, described by Φ:

$$\Phi(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt. \tag{4.3}$$

Note *One can show that the function e^{-t^2} does not have any antiderivative, i.e., if $g'(t) = e^{-t^2}$ then $g(t)$ cannot be expressed as a finite combination of elementary functions (though it can be approximated that way). $\Longrightarrow$ Do not try to compute $\Phi(x)$ analytically (use a table or a computer).*

In Python, the PDF can be generated at the values stored in array x via `scipy.stats.norm.pdf(`x`)`, the CDF via `scipy.stats.norm.cdf(`x`)`. As in all languages, the default location and scale parameters are 0 and 1, respectively.

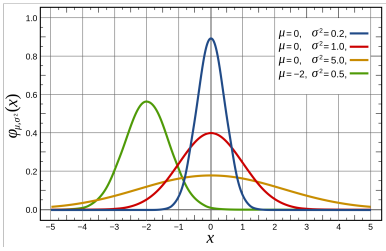


Figure 4.1: PDF of the Normal distribution. Credit:Wikipedia

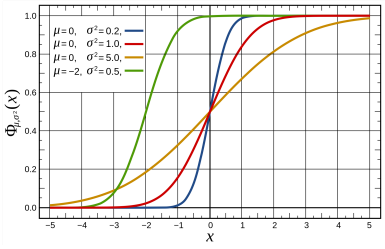


Figure 4.2: CDF of the Normal distribution. Credit:Wikipedia

Standard Normal

The standard normal is the only one we really need to worry about, because of the wonderful affine invariance of the normal distribution: if $X \sim N\left(\mu, \sigma^2\right)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}\left(0,1\right) \Rightarrow X = \sigma Z + \mu$. Hence:

$$\begin{aligned} F_X\left(x\right) &= P\left(X \leq x\right) \\ &= \mathbb{P}\left(\sigma Z + \mu \leq x\right) \\ &= \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

This means that we can express the PDF and CDF of any normal distribution with the standard normal PDF and CDF, simply by centering and rescaling.

Example

Assume that the temperature at some place follows a normal distribution with mean $\mu = 20°C$, and standard deviation $\sigma = 5°C$. What is the probability that the temperature drops below $12°C$?

$$T \sim N\left(\mu = 20, \sigma^2 = 25\right); \quad P\left(T \leq 12\right) = ?$$

$$\begin{aligned} P\left(T \leq 12\right) = & \\ &= \mathbb{P}\left(Z\sigma + \mu \leq 12\right) \\ &= \mathbb{P}\left(Z \leq \frac{12-\mu}{\sigma}\right) \\ &= \mathbb{P}\left(Z \leq -\frac{8}{5}\right) \\ &= \Phi\left(-\frac{8}{5}\right) = \Phi\left(-1.6\right) \\ &\approx 5.5\%. \qquad \textit{(Unlikely)} \end{aligned}$$

Moments

Consider $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then:

$$\begin{aligned} E\left(X\right) &= \mu, \qquad \text{Expected value ("mean").} \\ E\left(\left(X-\mu\right)^2\right) &= \sigma^2, \qquad \text{Variance.} \end{aligned}$$

So:

$$\mathcal{N}(\underbrace{\mu}_{\text{mean}}, \underbrace{\sigma^2}_{\text{variance}})$$

where σ is the standard deviation.

In general, if $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, it looks like Fig. 4.4.

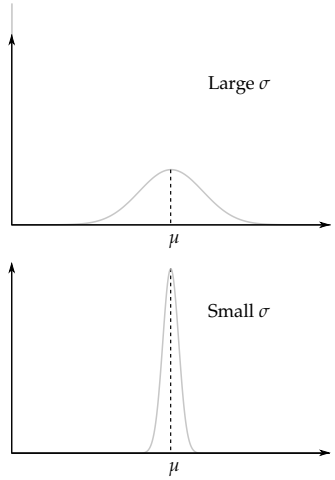


Figure 4.3: Two normal distributions with the same mean, μ, and different scale parameters σ.

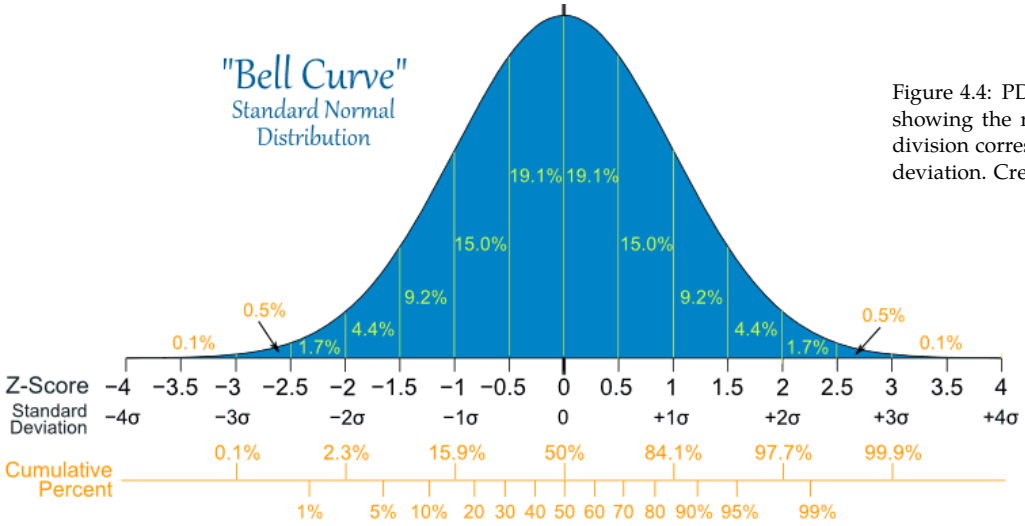


Figure 4.4: PDF of the standard normal, showing the respective percentage subdivision corresponding to each standard deviation. Credit:Math Is Fun

The normal distribution is unimodal and symmetric, so the mean, the mode and the median coincide (this is rare). Hence, it is not skewed, so $E\left((X-\mu)^3\right) = 0$. Its fourth moment is related to the *kurtosis*, which measures the "peakedness" of the distribution:

$$\frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4} \tag{4.4}$$

This quantity is 3 for the normal distribution, and it is common to define $\frac{\mu_4}{\sigma^4} - 3$ as *excess kurtosis*. By definition, the normal distribution has zero excess kurtosis. More peaked distributions are called *leptokurtic*, less peaked distributions are called *platykurtic*.

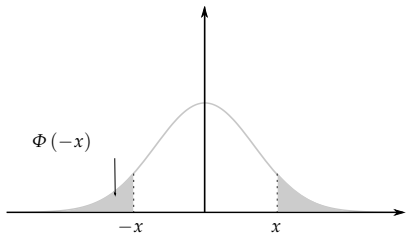


Figure 4.5: Representation of the symmetry of the cumulative distribution function since $\Phi(-x)$, the left gray region, is equal to its complement $1-\Phi(x)$, the right gray region.

Notable properties

1. Symmetry

$$\boxed{\Phi(-x) = 1 - \Phi(x)}$$

Proof *:*

Using the unit measure and the previous result:

$$\begin{aligned} 1 &= \mathbb{P}(X \leq -x) + \mathbb{P}(-x \leq X \leq x) + \mathbb{P}(X \geq x) \\ &= \Phi(-x) + \underbrace{\mathbb{P}(-x \leq X \leq x) + \Phi(-x)}_{\Phi(x)}. \end{aligned}$$

So $\Phi(-x) = 1 - \Phi(x)$. *(cf Fig. 4.5)*

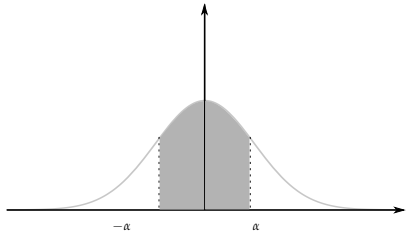


Figure 4.6: Illustration of property 2 of the normal CDF

2.

$$\boxed{\mathbb{P}(|X| \leq x) = 2\Phi(x) - 1}$$

Proof $\mathbb{P}(X \leq -x) = \mathbb{P}(X \geq x)$ *But:*

$$\begin{aligned} 1 &= \mathbb{P}(X \leq -x) + \mathbb{P}(-x \leq X \leq x) + \mathbb{P}(X \geq x) \\ &= 2(1 - \Phi(x)) + \mathbb{P}(-x \leq X \leq x) \,. \\ \implies \quad \mathbb{P}(|X| \leq x) &= 2\Phi(x) - 1 \,. \; \textit{QED} \end{aligned}$$

Stability

The normal distribution is stable under scaling, addition and subtraction. If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}\left(\mu_y, \sigma_y^2\right)$ are **independent** normal random variables, and a and b are constants, then:

Scaling If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}\left(a\mu + b, (a\sigma)^2\right)$. In other words, the mean gets an affine transform, and the standard deviation is scaled by a factor a.

Addition The sum of two normals is another normal whose mean and variances are the sums of individual means and variances.

$$U = X + Y \sim \mathcal{N}\left(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2\right) \,. \tag{4.5a}$$

The converse is also true (Cramer's theorem)[1]

[1] If X, Y independent and $X + Y$ is normal, then X and Y must be normal.

Subtraction The difference of two normals is another normal whose mean is the difference of the individual means, but whose variance is the sum of the individual variances:

$$X - Y \sim \mathcal{N}(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2) \,. \tag{4.5b}$$

$$\downarrow$$

Be careful! Variances are added.

Although we've seen additivity before, you should realize how remarkable it is: most distributions do not display this property. And the remarkable result is that if you subtract two normal RVs, the means are subtracted but the variances (measures of uncertainty) are added.

II Limit Theorems

Central Limit Theorem

Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Sir Francis Galton (1822 – 1911)

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d) random variables each having mean μ and variance σ^2.

Central limit theorem (CLT): As n increases, the sample average:

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n},$$

its distribution converges to: $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Put differently, if $X_1, \ldots, X_n$ be independent with common finite mean μ and variance σ^2 and:

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

where $\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$. Then, $Z_n \to N(0,1)$, hence:

$$\boxed{\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)} \tag{4.6}$$

That is, $\mathcal{N}(0,1)$ is the **asymptotic distribution** of Z_n, regardless of the original distribution of the $X_1, \ldots, X_n$. This is absolutely mind-boggling.

Remarkable facts

→ Works for "any" distribution as long as μ and σ exist.

→ Notice the factor $\sigma/\sqrt{n}$: *uncertainties decrease as the square root of the number of observations*. This is the theoretical basis for well-known laboratory practice of averaging independent measurements together to decrease uncertainty.

→ normal asymptotics apply for any n larger than about 30. Infinity is within reach of anyone who can count to 30!

→ An analog demonstration of the CLT is given by Galton boards, illustrated by the Quincux game

In summary, we have three extremely powerful properties that conspire to make the normal (Gaussian) distribution the **central** distribution of statistics:

1. The average of n independent and identically-distributed variables tends to be Gaussian for n sufficiently large, *regardless of their original distribution*.

2. For certain values of their parameters, many distributions (Binomial, Poisson, Gamma, Chi-squared, student's T, etc) converge to a Gaussian form, for similar reasons (they arise as the sum of many independent increments). The normal distribution is therefore an *attractor* of the distribution space.

3. For analytical reasons, once a Gaussian shape is attained, it is preserved under a variety of operations (scaling, convolution, product, Fourier transform). The consequence is that the sum or difference of normally-distributed variables is Gaussian too, and so it is for all other linear transformations... so one never escapes the Gaussian "black hole" once in its vicinity.

This will make the normal distribution a tool of choice to represent and analyze errors in all manner of measurements.

Application

A certain strain of bacteria occurs in all raw milk.

$$X = \text{bacteria count per milliliter of milk}\,.$$

According to the health department, if the milk is not contaminated, the mean μ of x is ≈ 2500 with a standard deviation σ of about 300.

An inspector measures x on 42 samples of milk that has been held in a cold storage awaiting shipment. The mean bacteria count $\overline{x}$ is 2700. What would you do if you were the inspector?

Solution *By the CLT, if the milk is not contaminated,* $\overline{x} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$,

$$\mu_{\overline{x}} = \mu = 2500\,, \qquad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{300}{\sqrt{42}} \approx 46.3\,,$$

$$\mathbb{P}\left(\overline{x} \leq 2650\right) = 0.9994\,.$$

$\Longrightarrow$ *Observing* $\overline{x} = 2700$ *is very unlikely if the milk is not contaminated.*
$\Longrightarrow$ *The milk must not be sold!*

Other limit theorems

Convergence of Binomial law to Poisson law

The binomial distribution is extremely useful, but the calculation of the $\binom{n}{k}$ factors, made of factorials, is extremely tedious, even for modern computers (try computing 100! if you don't believe it). It turns out that in the case where n gets large but $np = \lambda$ (a constant), the binomial distribution converges in distribution to the Poisson distribution. That is, if $X \sim \mathcal{B}(n, p)$:

$$\mathbb{P}(X = k) \xrightarrow[n\to\infty]{np=\lambda} e^{-\lambda}\frac{\lambda^k}{k!} \tag{4.7}$$

which is more manageable if you can easily compute exponentials. But it does get better: we may dispense with all these pesky factorials altogether, as shown below.

Convergence of Poisson law to Normal law

One can show using Stirling's approximation that, in turn, the distribution of a Poisson RV X with parameter λ converges to a normal distribution as $n \to \infty$. That is, the CDF of $\frac{X-\lambda}{\sqrt{\lambda}}$ converges to Φ.

So if we had $X \sim \mathcal{B}(n, p)$ such that $n \to \infty$ and np stays constant, we could use the normal approximation to evaluate the Poisson PMF. That is just how DeMoivre and Laplace discovered it in the first place (see Sect. III). Put it another way, the standardized sum:

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}; \quad S_n = \sum_{i=1}^{n} X_i \tag{4.8}$$

converges in distribution to a standard normal.

Example

Suppose that a coin is tossed 100 times and lands heads up 60 times. Should we be surprised and doubt that the coin is fair?

$$\mathbb{P}(S \geq 60) = \mathbb{P}\left(\frac{S-50}{5^2} \geq \frac{60-50}{5^2}\right) \approx 1 - \Phi(2) = 0.0228 \tag{4.9}$$

With n this large, there is only a 2% probability of observing this result, so the fairness of the coin is called into question. This logic will be used again in Chapter 6 with statistical hypothesis tests.

III A bit of history

Physicists believe that the Gaussian law has been proved in mathematics while mathematicians think that it was experimentally established in physics.

Henri Poincaré, 1854 –1912

Figure 4.7: The venerable Carl Friedrich Gauss, who contrary to popular belief did not discover the Gaussian law.

Origins

Though usually named after Gauss, the "Gaussian" (normal) distribution was actually discovered in 1733 by DeMoivre, who found it as a limit case of the binomial distribution with $p = 1/2$, but did not recognize the significance of the result. In the general case $0 < p < 1$, Laplace (1781) had derived its main properties and suggested that it should be tabulated due to its importance. Gauss (1809) considered another derivation of this distribution (not as a limiting form of the binomial distribution); it became popularized by his work and thus his name became attached to it. It seems likely that the term "normal" is associated with a linear regression model for which Gauss suggested the least-squares (LS) solution method and called the corresponding system of equations the *normal equations* (Chapter 13). One more name – *central distribution* – originates from the central limit theorem. It was suggested by Pólya and actively backed by *Jaynes* (2004). A well-known historian of statistics, Stigler, formulates an universal law of eponymy that "no discovery is named for its original discoverer."[2] *Jaynes* (2004) remarked that "the history of this terminology excellently confirms this law, since the fundamental nature of this distribution and its main properties were derived by Laplace when Gauss was six years old; and the distribution itself had been found by de Moivre before Laplace was born." (*Kim and Shevlyakov*, 2008)

[2] As already remarked for Bayes' theorem

Herschel, Maxwell, and the Error Distribution

Herschel (1850) arrived at a normal distribution when trying to describe errors in the location of stars obtain from optical measurements in the (x, y) plane. His reasoning was the following:

- Errors in x should be independent of errors in y, so the joint density $f(x, y) = f(x)f(y)$
- The distribution should depend only on the distance to the origin, so $f(x, y) = f(\sqrt{x^2 + y^2})$.

It turns out that the conjunction of these two seemingly simple properties is enough to enforce the bivariate gaussian form:

$$f(x, y) \propto \exp\left(\alpha(x^2 + y^2)\right) \tag{4.10}$$

Clearly, α must be negative for f to be integrable, and after normalization to unit mass, it can be rewritten in the usual form:

$$f(x,y) = \frac{1}{\sigma^2 2\pi} \exp\left(-\frac{x^2+y^2}{\sigma^2}\right) \tag{4.11}$$

which reduces to Eq. (4.1) in the univariate case (with $\mu = 0$).

This form is eerily reminiscent of the Maxwell-Bolztmann distribution of velocities in a gas, which forms the basis for the kinetic theory of gases. There is in fact a deep link between Gaussians and diffusion processes, known as the Fokker-Planck equation in statistical mechanics. Essentially, when one has big ensemble of particles merrily bumping into each other, their velocities wind up with a Gaussian distribution. For a gas at rest, the mean velocity is zero, and the variance is proportional to the temperature (which is therefore a measure of molecular agitation).

Information Theory

Shannon and Nyquist pioneered the concept of *information entropy* associated with a probability distribution. As in thermodynamics, information entropy can only increase under spontaneous transformations. It turns out that, by just specifying the mean and the variance of a measurement error, the maximum entropy principle leads to a Gaussian form for their distribution. *Jaynes* (2004) interprets the result this way:

> *The normal distribution provides the most honest description of our knowledge of the errors given just the sample mean and variance of the measurements.*

This brings a theoretical justification for what countless astronomers and experimentalists have observed empirically since the 19th century: the normal distribution is therefore a legitimate error distribution for most applications, and it should always be used as an error distribution unless one has cogent evidence that errors are not normally distributed.

Maximum Error Cancellation

It can be shown that the sample mean $\overline{x} = \frac{1}{n}\sum x_i$ brings about the maximum chance of error cancellation for symmetrically distributed errors (equal chance of positive and negative excursions means that the average will be closer to the 'true' value).

IV Error Analysis

Error Terminology

Lack of mathematical culture is revealed nowhere so conspicuously as in meaningless precision in numerical computations.

Carl Friedrich Gauss, 1777 – 1855.

Absolute and relative error

Experimental errors are often reported in two ways:

$$\text{Absolute error:} \qquad x = x_{best} + \Delta x \tag{4.12a}$$

$$\text{Relative error:} \qquad x = x_{best}\left(1 \pm \frac{\Delta x}{|x_{best}|}\right) \tag{4.12b}$$

Accuracy vs. precision

Precision measures the spread of repeated measurements, *i.e.* their tendency of cluster close each other. Accuracy, in contrast, measures the distance to the "true" value of the quantity being measured (Fig. 4.8).

By the central limit theorem, precision can always been improved by repeating measurements, but even devilish levels of precision can lead to inaccurate measurements if systematic *biases* exist (*e.g.* a thermometer is always 0.2K too cold, a mass spectrometer has a 0.5‰ offset, etc). One needs external information to determine those (e.g. data run in another lab or another machine, operator log, cross-validation, etc.).

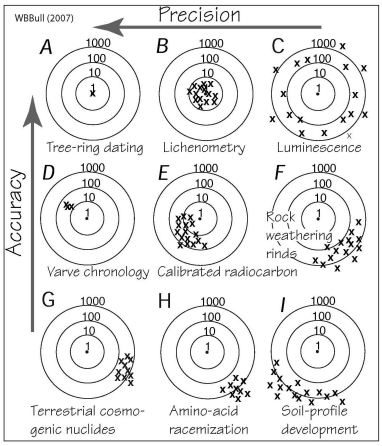


Figure 4.8: Diagram of precision and accuracy for various dating methods. Credit:Tectonic Geomorphology of Mountains (Figure 6.1, p. 211)

Modeling Errors Using the Normal Distribution

Common Model:

$$\text{Measurement} = \text{"true quantity"} + \text{error}\,.$$

$$y = \mu + \varepsilon\,.$$

where $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. Measurement $\rightarrow$ random variable.

Recall that $P\left(-\sigma \leq \varepsilon \leq \sigma\right) \approx 68.1\%$. Hence μ lies in the interval $[\mu - \sigma, \mu + \sigma]$ with a probability of 68.1%. We shall see later that if we estimate μ via a sample average $\bar{x}$, then: $[\bar{x} - 1.96\sigma, \bar{x} + 1.96\sigma]$ is a 95% *confidence interval* for μ.

Covariance, Correlated Errors

X, Y random variables.

Covariance

$$\mathrm{Cov}(X,Y) = E\left((X - E(X))(Y - E(Y))\right) . \tag{4.13}$$

Special Case: $\mathrm{Cov}(X,X) = \mathrm{Var}(X) = E\left((X - E(X))^2\right)$.

X, Y independent $\Longrightarrow$ $\mathrm{Cov}(X,Y) = 0$. The converse is **false** in general; except, of course, when X, Y are both normal, in which case: $\mathrm{Cov}(X,Y) = 0 \Longrightarrow X$, Y independent. Thus, in the Gaussian world, independence and zero covariance are the same thing. One can show that:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(X_i, X_j\right) . \tag{4.14}$$

In particular, $X \sim N\left(0, \sigma_X^2\right)$ and $Y \sim N\left(0, \sigma_Y^2\right)$:

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\mathrm{Cov}(X,Y) . \tag{4.15a}$$

i.e. $$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(x,y) \geq \mathrm{Var}(X) + \mathrm{Var}(Y) . \tag{4.15b}$$

(the equality holds if and only if the variables are not correlated.)

Simple Rules of Error Propagation

Using these properties, one can derive simple rules for various functions of a normal RV.

Addition/Subtraction In general, with $X \pm \Delta X$; $\quad Y \pm \Delta Y$:

The uncertainty on $X + Y$ is at most $\Delta X + \Delta Y$. But assuming X, Y are independent and normal,

$$\begin{aligned} \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \Longrightarrow \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_X^2}\,, \\ \Longrightarrow \Delta(X+Y) = \Delta(X-Y) = \sqrt{(\Delta X)^2 + (\Delta Y)^2} \leq \Delta X + \Delta Y . \end{aligned} \tag{4.16a}$$

Put differently, the most conservative error estimate for sums or differences is the sum of their uncertainties, but for independent errors this reduces considerably to a **quadratic sum** (the norm of the uncertainty vector)[3].

[3] The inequality stems from Pythagoras' theorem

Product/Division

$$\log(xy) = \log(x) + \log(y) \ .$$
$$\log\left(\frac{x}{y}\right) = \log(x) - \log(y) \ .$$
$$\Delta \log x \approx \frac{\Delta x}{x} \ . \qquad \text{(First-order approximation)}$$

If we call $u = xy$ and $v = x/y$:

$$\Delta \log u \approx \frac{\Delta u}{u} \leq \Delta \log x + \Delta \log y \approx \frac{\Delta x}{x} + \frac{\Delta y}{y} \ .$$

$$\boxed{\frac{\Delta(xy)}{xy} = \frac{\Delta(x/y)}{x/y} \approx \frac{\Delta x}{x} + \frac{\Delta y}{y} \ .} \tag{4.16b}$$

which is just the same as Eq. (4.16a), but with relative instead of absolute uncertainties. In words, the relative uncertainty of a product (or ratio) is the sum of the relative uncertainty of its components (unless there is some partial cancellation).

Nonlinear Propagation

For any well-behaved function $\psi(X_1, X_2, ..., X_n)$, with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, where the X_i are independent of each other[4]:

$$\sigma_\psi^2 \approx \left|\frac{\partial \psi}{\partial X_1}\right|_{\mu_1}^2 \sigma_1^2 + \left|\frac{\partial \psi}{\partial X_2}\right|_{\mu_2}^2 \sigma_2^2 + \cdots + \left|\frac{\partial \psi}{\partial X_n}\right|_{\mu_n}^2 \sigma_n^2 \tag{4.16d}$$

Where the subscript means that we evaluate the function around the value $\psi(\mu_1, \mu_2, \cdots, \mu_n)$, and higher order terms are assumed negligible. This result stems from performing a first-order Taylor expansion of ψ about the mean, using the transfer theorem (Eq. (3.12)) to propagate the errors through ψ, and using the linearity of the expectance operator. This allows us to compute the contributions to total uncertainty due to several input sources (*cf.* Lab 4). This may be particularly helpful when considering which aspects of experiment design should be improved to yield to lowest overall uncertainties.

[4] In the more general case, the equation writes

$$\sigma_\psi^2 \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \left|\frac{\partial \psi}{\partial X_i}\right|_{\mu_i} \left|\frac{\partial \psi}{\partial X_j}\right|_{\mu_j} \mathrm{Cov}(X_i, X_j) \tag{4.16c}$$

where μ_i is the mean of X_i

Chapter 5

PRINCIPLES OF STATISTICAL ESTIMATION

"There are three types of lies: lies, damn lies and statistics"

Benjamin Disraeli

Having built an understanding of measurements as realizations of random processes governed by probability laws, and having seen how many natural processes can be modeled by relatively simple, parametric distributions (the normal distribution being the most ubiquitous), we now turn to the more technical topic of *fitting those distributions to observations*. Because those theoretical distributions are entirely characterized by a handful of parameters $\theta = (\theta_1, \cdots, \theta_p)$, this amounts to estimating θ based on n observations[1] $\boldsymbol{x} = (x_1, \cdots, x_n)$, ideally with uncertainty bounds for those parameters.

[1] $n \geq 1$, hopefully.

I Preamble

Methods

We shall see three estimation methods:

1. **The method of moments**, which is simple, but risky.
2. **Maximum likelihood estimation**, which is not much more complicated, but optimal by design.
3. **Bayesian estimation**, which is much richer, but usually more difficult.

In the following, we will denote estimates by a hat ($\hat{\theta}$) to make plain that they are only a proxy for the real thing. The meaning of this notation is worth explaining: we imagine a world where true parameters θ roam free, and we try to ascertain what values they might assume in our material world (possibly, your lab bench). This frequentist view applies to the first two. The Bayesian view is that the parameters are

estimated from uncertain data, so they are themselves uncertain, and therefore must be characterized by a probability distribution. If the estimation is good, that distribution is narrowly focused around a central value, otherwise it will be spread out.

Point vs Interval Estimation

In the frequentist world, it is common to separate *point estimation* from *interval estimation*. Point estimation applied to the age of the Earth would be to say: given 10 Ar-Ar dates from mineral inclusions, what's your best guess for the age of the Earth? Interval estimation would be to say: how well do you know that? In the Bayesian view, these distinctions are somewhat superficial for the reason noted above (give me a distribution, I'll give you both its central tendency and its spread). However, since the frequentist viewpoint has long pervaded the experimental sciences, it is worth knowing this terminology.

II Method of Moments

We've seen in Chapter 3 that the moments of most distributions usually have a simple relationship to their parameters; the normal distribution is an extreme example of this, because its two parameters are essentially its first two moments. Hence the idea, due to Karl Pearson ca. 1890, to use those moments to estimate the parameters. It is quite simple to use, and almost always yield some sort of estimate. Unfortunately, in many cases, yields estimates that leave a lot to be desired (e.g. a probability greater than one, a sample number smaller than zero). However, it is a good place to start when other methods prove intractable, and it is used in Lab 5.

Theorem 1 (Strong Law of large numbers) *The n-sample mean*

$$\bar{X}_n = \sum_{i=1}^{n} x_i / n \tag{5.1}$$

converges in probability to the true mean as the number of trials tends to infinity:

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu \right) = 1. \tag{5.2}$$

This law justifies the intuitive interpretation of the expected value of a random variable when sampled repeatedly as the long-term average. It is a justification for the ergodic assumption frequently used in physics.

According to the law of large numbers, the k^{th} moment of a random variable X may be estimated as:

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \approx E\left(X^k\right) = \mu_k \,. \tag{5.3}$$

Method of Moments:

1. Express the unknown parameter(s) θ as a function of the moments μ_k ($k \leq 1$):
$$\theta = g(\mu_1, \ldots, \mu_k) . \tag{5.4a}$$

2. Approximate the moments μ_k by the sample moments $\hat{\mu_k}$ and solve for θ:
$$\begin{cases} \mu_1 = \mu_1(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} x_i , \\ \mu_2 = \mu_2(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} x_i^2 , \\ \vdots \\ \mu_k = \mu_k(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} x_i^k , \end{cases} \tag{5.4b}$$

Example (Normal method of moments)

$X_1, \ldots, X_n$ *i.i.d.* $N(\mu, \sigma^2)$ ($X \sim \mathcal{N}(\mu, \sigma^2)$).

$\begin{cases} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{cases}$ *parameters to estimate.*

$$\mu_1 = E(X) = \mu .$$
$$Var(X) = \sigma^2 = E\left(X^2\right) - E^2(X) = \mu_2 - \mu^2 .$$
$$\Longrightarrow \mu_2 = \mu^2 + \sigma^2 .$$

Faster to Solve:

$$\begin{cases} \mu_1 = \mu = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i , & \text{(E1)} \\ \mu_2 = \mu^2 + \sigma^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 . & \text{(E2)} \end{cases}$$

We must solve the system simultaneously for μ and σ^2. We know that $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$. Substituting (E1) into (E2):

$$\begin{aligned} \mu^2 + \sigma^2 = \overline{X}^2 + \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 . \\ = \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 . \end{aligned}$$

So, $\begin{cases} \hat{\mu} = \overline{X} . \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 \equiv S^2 . \quad \leftarrow \textit{Sample variance.} \end{cases}$

Example
Given θ balls in a jar numbered $1, 2, \ldots, \theta$. We draw balls at random ($P(x_i) = 1/\theta$, uniform distribution). We want to estimate θ.

$$X_1, \ldots, X_n \text{ i.i.d. sample.}$$

Method of moments:

$$\mu_1 = E(X) = \sum_{i=1}^{\theta} i \cdot P(X = i) \quad = \sum_{i=1}^{\theta} i \cdot \frac{1}{\theta} = \frac{1}{\theta} \sum_{i=1}^{\theta} i \quad = \frac{1}{\theta} \frac{\theta(\theta+1)}{2} = \frac{\theta+1}{2}.$$

System of equations to solve: $\mu_1 = \frac{\theta+1}{2} = \frac{1}{n} \sum_{i=1} n x_i = \overline{X}$.

$$\boxed{\hat{\theta}_{MoM} = 2\overline{X} - 1} \tag{5.5}$$

Example
$X_1 = 3$, $X_2 = 2$, $X_3 = 10$, $X_4 = 4$, $X_5 = 2$.

$$\overline{X} = 4.$$
$$\hat{\theta}_{MoM} = 2\overline{X} - 1 = 7.$$

(A rather poor estimate since we already know that 10 has been drawn, so we know for sure that θ should be greater than 10!)

III Maximum Likelihood Estimation

The Big Idea

The main idea is to treat the conditional probability of the observations $f(X|\theta)$ as a function of the unknown parameters θ (with X fixed), and find the value of the parameters that maximizes this function. That is, we seek the value of the parameters that maximizes the probability of observing the sample. This is much more defensible than the method of moments, and leads to all sorts of nice properties.

The starting point is that the n observations are considered a **random sample** of size n, *i.e.* realizations of n i.i.d. random variables $X_1, \ldots, X_n$.

Discrete Case: estimating a Bernoulli parameter

Example
Sequence of Bernoulli trials (Tossing a coin, for example)

$$X = \begin{cases} 1, & \mathbb{P}(X = 1) = p, \\ 0, & \mathbb{P}(X = 0) = 1 - p. \end{cases}$$

In other words:

$$\mathbb{P}(X = x|p) = p^x (1-p)^{1-x}, \text{ with } x = 0, 1 \tag{5.6}$$

Given the sequence

$\boldsymbol{x} = \{1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0\,.\}$ *(7 ones and 8 zeros)*

We want to estimate p (= θ). How do we proceed?

The silly way would to try different values of p, simulate from that model, and see if the statistics are compatible with this sequence.

$$\begin{aligned} &P\,(\text{observing } x_i \text{ if } p = 0.99) && \rightarrow \text{Small}\,. \\ &P\,(\text{observing } x_i \text{ if } p \approx 0.5) && \rightarrow \text{Large}\,. \\ &P\,(\text{observing } x_i \text{ if } p = 0.001) && \rightarrow \text{Small}\,. \end{aligned}$$

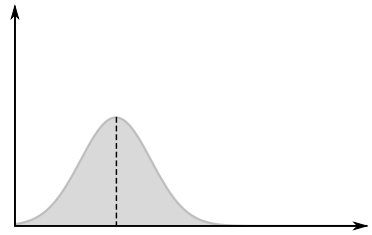


Figure 5.1: Likelihood of the Bernoulli trial with mean $\bar{x} = 0.5$.

We can, of course, be more efficient about it. We can recognize the probability of observing each datum as conditional on the parameters:

$$P\,(X_i = x_i|\,\theta) = f_\theta\,(x_i)$$

and we seek the probability of observing $(x_1, \cdots, x_n)$ simultaneously, that is, we seek the *joint probability distribution* of $(X_1, \cdots, X_n)$

From the independence assumption,

$$\begin{aligned} \mathbb{P}\,(X_1 = x_1, \ldots, X_n = x_n) &= \mathbb{P}\,(X_1 = x_1) \times \cdots \times \mathbb{P}\,(X_n = x_n) \\ &= \prod_{i=1}^{n} \mathbb{P}\,(X_i = x_i) \\ &= \prod_{i=1}^{n} f_\theta\,(x_i) \\ &\equiv L(\theta|\boldsymbol{x}) \end{aligned} \tag{5.7}$$

Which we call the *likelihood function*. For each value of θ, the likelihood $L(\theta|\boldsymbol{x})$ is the probability of observing $\{X_1 = x_1, \ldots, X_n = x_n\}$ for that distribution and that parameter θ. Using Eq. (5.6):

$$L\,(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{5.8}$$

Further, recognizing that $p^a \times p^b = p^{a+b}$, we can transform the product as a sum:

$$L\,(p) = p^{\sum x_i}(1-p)^{n-\sum x_i} = p^y(1-p)^{n-y} \tag{5.9}$$

wherein $y = \sum_{i=1}^{n} x_i$, the number of successful trials.

Now, it is generally more practical to work with the so-called log-likelihood $\ell\,(\theta) = \log L\,(\theta)$ instead of $L\,(\theta)$: the product becomes a sum, and that greatly simplifies calculations (which is exactly why the log was invented in the first place).

Definition 2 *The log-likelihood is given by:*

$$\begin{aligned}\ell(\theta) &\equiv \log L(\theta) \\ &= \log \prod_{i=1}^{n} f(x_i|\theta) \\ &= \sum_{i=1}^{n} \log f(x_i|\theta) \ . \end{aligned} \tag{5.10}$$

Since log is an increasing function, the maximum of L is the same as the maximum of ℓ, hence:

$$\ell\left(\hat{\theta}_{MLE}\right) = \max_{\theta} \ell(\theta) \tag{5.11}$$

Definition 3 *A parameter $\hat{\theta}$ maximizing $L(\theta)$ is said to be a **maximum likelihood estimator** (MLE) of θ.*

$$\boxed{L\left(\hat{\theta}_{\mathrm{MLE}}\right) = \max_{\theta} L(\theta)} \tag{5.12}$$

All we have to do now to find the MLE of p, is to compute $\ell(p)$ for $0 < p < 1$[2], and find its maximum:

$$\ell(p) = \log\left(p^y (1-p)^{n-y}\right) = y\log(p) + (n-y)\log(1-p) \tag{5.13}$$

Now, the extrema $\ell(\theta)$ occur whenever $\frac{\partial \ell}{\partial \theta} = 0$[3].
If we do this, we get :

$$\frac{y}{p} - \frac{n-y}{1-p} = 0 \tag{5.14}$$

After a few rearrangements, we get the solution:

$$\hat{\theta}_{MLE} = \hat{p} = y/n = \bar{x}. \tag{5.15}$$

That is, our best guess for the trial probability p is just the average of the successes. Intuitive though this may seem, it is immensely satisfying to arrive at this result via a rigorous optimality principle.

Remark: $\hat{\theta}_{MLE}$ does not have to be unique (cf Fig. 5.2), although it is often the case in practice.

[2] Strange things may happen at the boundary, so one needs to verify that other maxima are not present there

[3] A necessary, but not sufficient, condition for finding a maximum. For instance, the likelihood may have more than one maximum, or a minimum as in Fig. 5.2. We'll see later how to ensure that a unique maximum has been reached.

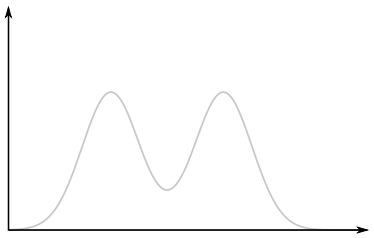


Figure 5.2: A bimodal likelihood.

Continuous Case: normal observations

Similarly in a continuous case with IID observations, the likelihood is the joint density of the observations f_J which factorizes as the product of the density functions:

$$L(\theta) = f_J(x_1, \cdots, x_n \mid \theta) \tag{5.16a}$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad \text{Independence} \tag{5.16b}$$

$$L\left(\hat{\theta}_{MLE}\right) = \max_{\theta} L(\theta) \,. \tag{5.16c}$$

If ℓ (or L) is differentiable with respect to θ, possible candidates for the MLE of θ are solutions of:

$$\frac{\partial L}{\partial \theta} = 0\,, \qquad \left(\text{or } \frac{\partial \ell}{\partial \theta} = 0\right) .$$

For a vector $\theta = (\theta_1, \ldots, \theta_k)$, this implies:

$$\frac{\partial L}{\partial \theta_i} = 0\,, \qquad \left(\text{or } \frac{\partial \ell}{\partial \theta_i} = 0\right) \quad \forall i \in \{1, \cdots, k\}.$$

Estimating the parameters of a normal model

Consider a random normal sample of size n, that is: $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$.

$$\begin{cases} \hat{\mu} = ?\,, & \mu \in \mathbb{R}\,, \\ \hat{\sigma}^2 = ?\,, & \sigma^2 \in (0, \infty)\,. \end{cases}$$

Calling $\begin{cases} \theta_1 = \mu\,, \\ \theta_2 = \sigma^2\,, \end{cases}$ and given that $f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, then:

$$\begin{aligned} L\left(\mu, \sigma^2\right) &= \prod_{i=1}^{n} f\left(x_i|\mu, \sigma^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}\,, \quad \rightarrow \quad \text{Likelihood function.} \end{aligned} \tag{5.17}$$

$\underline{log - \text{likelihood}}$:

$$\ell\left(\mu, \sigma^2\right) = \log L\left(\mu, \sigma^2\right) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\,.$$

We first want to maximize $\ell(\theta)$ with respect to $\theta_1 = \mu$:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0 \iff \sum_{i=1}^{n} x_i = n\mu$$
$$\iff \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}\,.$$

So the MLE of the mean is just the sample mean? Really, all this for that? Well, sure, it's intuitive, but now it's also rigorously established[4]. What about σ? Can it be estimated via the sample standard deviation? It turns out that this is indeed the MLE. Now maximizing $\ell(\theta)$ with respect to $\theta_2 = \sigma^2$:

[4] Note that Gauss famously obtained the normal distribution as one that would have this property: that is, he considered the inverse problem of designing a distribution for which the sample mean would be the best estimate of the true mean (*Kim and Shevlyakov*, 2008)

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^4} = 0\,.$$
$$\iff n\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \mu)^2\,,$$
$$\iff \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2\,, \qquad \text{but we just showed that } \hat{\mu} = \overline{x},$$
$$\iff \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = s^2 \quad \text{sample variance}$$

So,
$$\begin{cases} \hat{\mu} = \overline{x}\,, \\ \hat{\sigma}^2 = s^2\,. \end{cases}$$

This happens to coincide with the result of the method of moments, but it will turn out to be an exception.

Now, is $(\overline{x}, s^2)$ really a maximum of the likelihood?
$$\begin{cases} \text{Max ?} \\ \text{Min ?} \\ \text{Saddle point ?} \end{cases}$$

A sufficient condition if for L or ℓ to have negative curvature around there:

$$1) \qquad \frac{\partial^2 \ell}{\partial \mu^2} < 0\,, \qquad \frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2} < 0\,, \tag{5.18a}$$

$$2) \qquad \frac{\partial^2 \ell}{\partial \mu^2} \frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2} - \frac{\partial^2 l}{\partial \mu \sigma^2} > 0\,, \tag{5.18b}$$

It is easy to verify that:

$$\hat{\theta}_{MLE} = \left(\hat{\mu}, \hat{\sigma}^2\right) = \left(\overline{x}, s^2\right)\,. \tag{5.19}$$

These estimates are unique for the normal distribution (it is unimodal), and show that the sample mean and sample variance are legitimate estimators of the true parameters (they are the most consistent with the observations in this model).

The method of moments revisited, and defeated

Let us compute the MLE for the previous example where

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta. \\ 0, & otherwhise. \end{cases}$$

$X_1, \ldots, X_n$ i.i.d.

$$\hat{\theta}_{MoM} = 2\overline{X} - 1, \qquad \text{(Unacceptable if } \max x_i > 2\overline{X} - 1)$$

$\hat{\theta}_{MLE} = ?$

Likelihood Function:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|0) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 \le x_i \le \theta. \\ 0, & \text{otherwise.} \end{cases}$$

where $\theta \ge \max x_i$. So $\hat{\theta}_{MLE} = \max(x_i)$

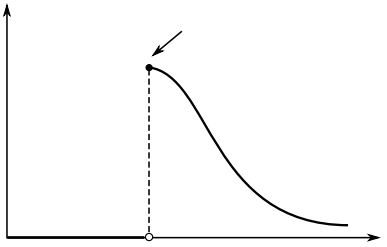


Figure 5.3: Likelihood function of the Example III. Here we can see that $\theta \le \max x_i$.

IV Quality of Estimators

By now you should be familiar with the idea that inference is the task of estimating unknown, sometimes latent parameters from observable data, given some model. We've seen two ways to do this, and we may ask, how good are they?

Recall here that estimators are *statistics*, that is, deterministic functions of observations, which we hypothesize as originating from a random process (one about which we have imperfect knowledge). Therefore estimators like $\hat{\mu}$ and $\hat{\sigma}$ are themselves random variables: they are characterized by a distribution. There are two desirable properties of estimators:

Accuracy (No or low bias):

$$\begin{aligned} \text{We want} \quad & E(\hat{\theta}) = \theta, \\ \text{We define} \quad & \text{bias}(\hat{\theta}, \theta) \equiv E(\hat{\theta}) - \theta. \end{aligned}$$

Precision : small variance of $\hat{\theta}$.

Example (with $\mathcal{N}(\mu, \sigma^2)$)

$$\begin{aligned} \hat{\mu} &= \overline{x}, \\ E(\overline{x}) &= E\left(\frac{x_1 + \ldots + x_n}{n}\right) \\ &= \frac{1}{n}\sum_i E(x_i) = \mu. \end{aligned}$$

No bias!

Mean Square Error (MSE)

Definition 4

$$MSE(\hat{\theta}, \theta) = E\left[(\hat{\theta} - \theta)^2\right] \tag{5.20}$$

The MSE incorporates the two types of errors: variability (precision) and bias (accuracy). An estimator that has good MSE properties has small combined bias and variance. Indeed,

$$\begin{aligned} E\left[(\hat{\theta}-\theta)^2\right] &= E\left(\hat{\theta}-E\left(\hat{\theta}\right)+E\left(\hat{\theta}\right)-\theta\right)^2 \qquad \rightarrow \text{(bias-variance decomposition of the MSE)} \\ &= E\left[(\hat{\theta}-E(\hat{\theta})^2\right]+\left[E(\hat{\theta})-\theta\right]^2+2\underbrace{E\left(\hat{\theta}-E\left(\hat{\theta}\right)\right)E\left(\hat{\theta}\right)}_{=0} \\ &= V(\hat{\theta})+\text{bias}^2 \end{aligned}$$

This is called the bias-variance decomposition.

Example (Normal estimation $\mathcal{N}\left(\mu,\sigma^2\right)$)
The maximum likelihood estimators we just derived are: $\hat{\mu}=\overline{X}$, $\hat{\sigma}^2=\frac{1}{n}\sum_i\left(x_i-\overline{X}\right)^2$. *They have the following mean and variance:*

$$\begin{cases} E\left(\overline{X}\right)=\mu\,, \\ V\left(\overline{X}\right)=\frac{\sigma^2}{n}\,, \end{cases} \qquad \textit{Unbiased.}$$

$$\begin{cases} E\left(\hat{\sigma^2}\right)=\frac{n-1}{n}\sigma^2\,, \\ V\left(\hat{\sigma}^2\right)=\frac{2(n-1)}{n^2}\sigma^4\,, \end{cases} \qquad \textit{Biased.}$$

Hence:

$$\begin{aligned} MSE\left(\hat{\mu},\mu\right) &= bias^2\left(\hat{\mu},\mu\right)+\text{Var}\left(\hat{\mu}\right) \\ &= \text{Var}\left(\hat{\mu}\right)=\frac{\sigma^2}{n}\,. \\ MSE\left(\hat{\sigma},\sigma\right) &= \left(\frac{n-1}{n}\sigma^2-\sigma^2\right)+\frac{2\left(n-1\right)}{n^2}\sigma^4 \\ &= \frac{(2n-1)}{n^2}\sigma^4\,. \end{aligned}$$

You can see that the estimator of the variance is slightly biased. Should we correct for that? *Prima facie* that seems like a good idea, but let's consider all error sources.

Bias-Variance Trade Off

Consider the following estimation for σ^2:

$$\hat{\sigma}^2=\frac{1}{n-1}\sum_i\left(x_i-\overline{x}\right)^2=\frac{n}{n-1}\hat{\sigma}^2_{MLE}\,.$$

Then

$$E(\hat{\sigma^2})=\sigma^2\,, \qquad \text{(No bias)}$$

and

$$\text{MSE}\left(\hat{\sigma}^2, \sigma^2\right) = V(\sigma^2) = \frac{2}{n-1}\sigma^4 .$$

But

$$\text{MSE}\left(\hat{\sigma}^2_{MLE}, \sigma^2\right) = \left(\frac{2n-1}{n^2}\right)\sigma^4 .$$

Considering that $\frac{2n-1}{n^2} < \frac{2}{n-1}$, we have that:

$$\text{MSE}(\ \hat{\sigma}^2_{MLE}\ , \sigma^2\ \ ) < \text{MSE}(\ \hat{\sigma}^2\ \ , \sigma^2\ \ ) . \tag{5.21}$$

$$\uparrow \text{bias}\ \downarrow \text{var} \qquad\qquad \downarrow \text{bias}\ \uparrow \text{var} \tag{5.22}$$

So even though it would be tempting to get rid of the bias, this would increase the overall error.By trading off variance for bias, a lower MSE can be reached. The MLE is optimal in the sense that it has the lowest MSE of all estimators; it sometimes achieves this by introducing some bias, so bias is not inherently evil. As usual in Life, there is no free lunch: one cannot possibly lower bias and variance, so tradeoffs must be made.

Consistency

Both estimators above have an MSE that will tend to zero as $n \to \infty$: this is a property called *consistency*: as we collect more and more observations, we eventually beat down our uncertainty about the true parameters to zero. Put differently, a consistent estimator is one that converges to the true parameter value as the number of observations gets large ("asymptotically"). We'll see in Chap 9 that some spectral estimators don't do that, and that will be grounds for dismissal.

Efficiency

Suppose θ is an unknown parameter which is to be estimated from measurements x, distributed according to some probability density function $f(x;\theta)$. The variance of any unbiased estimator $\hat{\theta}$ of θ is then bounded from below by the inverse of the *Fisher information* $I(\theta)$. That is:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \tag{5.23}$$

where the Fisher information $I(\theta)$ is defined by:

$$I(\theta) = \text{E}\left[\left(\frac{\partial \ell(x;\theta)}{\partial \theta}\right)^2\right] = -\text{E}\left[\frac{\partial^2 \ell(x;\theta)}{\partial \theta^2}\right] \tag{5.24}$$

and $\ell(x;\theta) = \log f(x;\theta)$ is the natural logarithm of the likelihood function and E denotes the expected value (over x).

The efficiency of an unbiased estimator $\hat{\theta}$ measures how close this estimator's variance comes to this lower bound; estimator efficiency is

defined as

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\text{var}(\hat{\theta})} \tag{5.25}$$

or the minimum possible variance for an unbiased estimator divided by its actual variance. The Cramér – Rao bound thus gives $e(\hat{\theta}) \leq 1$.

One can verify that the MLE is efficient; that is, it is the best point estimator data can buy.

Note *The method of moments yields estimators that are neither consistent nor efficient, so you should always prefer MLE. Most numerical estimation routines (including Python's fitter). MLEs have good theoretical properties: they are invariant under rescaling, consistent, efficient, and converge in distribution to a normal (asymptotically, that is, for a large number of observations). In practice, finding the maximum may not be as straightforward as in the normal cases shown above, but even in a case where no closed-form expression exists, it usually involves fairly modest computations. If it does not, the Expectation-Maximization algorithm will quickly become your friend.*

V Bayesian Estimation

The Big Idea

The following section is very strongly inspired from *Wikle and Berliner* (2007)

Gelman et al. (2013) define Bayesian inference as

> ". . . the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations".

The basic idea is that we want the distribution of the parameters given the observations, $p(\theta|X)$[5], but often we can only compute $p(X|\theta)$, and have some knowledge about θ. To invert this situation we use Bayes' theorem:

[5] Here p is a generic notation for a probability distribution

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{5.26}$$

There are four essential ingredients in this recipe, which all have great conceptual importance:

Data distribution $p(X|\theta)$. Statisticians often refer to this as a "sampling distribution" or "measurement model". It is simply the distribution of the data, given the unobservables. When viewed as a function of θ for fixed X, it is known as a likelihood function, $L(X|\theta)$, as in classical maximum likelihood estimation (Sect. III). A key is that one thinks of the data conditioned on θ. For example, if X represents imperfect observations of temperature, and θ the true (unobservable) temperature, then $p(X|\theta)$ quantifies the distribution of measurement

errors in observing temperature, reflecting possible biases as well as instrument error.

Prior distribution $p(\theta)$: This distribution quantifies our *a priori* understanding of the unobservable quantities of interest. For example, if X corresponds to temperature, then one might base this prior distribution on historical information (climatology) or perhaps from a forecast model. In general, prior distributions can be informative or non-informative, "subjective" or "objective". The choice of such distributions is an integral part of Bayesian inference[6].

[6] Because it involves a human decision, some frequentists sneer at this as a despicable parody of statistics – forgetting that they often have to make all kinds of unrealistic assumptions, most of them highly subjective, to be able to compute an MLE.

Marginal distribution $p(X)$. This may be rewritten by conditioning on the parameters and integrating over them[7]:

$$p(X) = \int p(X|\theta)p(\theta)d\theta \tag{5.27}$$

We assume continuous θ but note that there are analogous forms (sums) for discrete θ. This distribution is also known as the prior predictive distribution. Alternatively, for the observations X, $p(X)$ can be thought of as the "normalizing constant" in Bayes' rule. Unfortunately, it is only for very specific choices of the data and prior distribution that we can solve this integral analytically.

[7] effectively using the law of total probability Eq. (2.18)

Posterior distribution $p(\theta|X)$: This distribution of the unobservables given the data is our primary interest for inference. It is proportional to the product of the likelihood and the prior. The posterior is the update of our prior knowledge about θ as summarized in the prior $p(\theta)$ given the observations X. In this sense, the Bayesian approach is analogous to the scientific method: one has prior belief (information), collects data, and then updates that belief given the new data (information). Much about the rational human mind works this way (remember how you learned how to walk?).

In many cases, the normalizing constant $p(X)$ is of secondary importance, so it is common to rewrite Eq. (5.26) as:

$$\boxed{p(\theta|X) \propto p(X|\theta)p(\theta)} \tag{5.28}$$

Which makes plain that the posterior distribution (what we want) is the product of the likelihood (what we observed, given prior information) by the prior information.

Bayesian coin flips

We return to the examples of Sect. III, now with a Bayesian viewpoint. Given n coin flips (or some other kind of Bernoulli trial), how do we estimate the underlying parameter p?

Writing the likelihood is trivial once you know p (see Eq. (5.6)). Of course, we don't know p (that's the whole point), but we can set priors on it, and then apply Bayes' theorem to obtain the posterior $\mathbb{P}(p|\boldsymbol{x})$. That's exactly what's done in this excellent blog post. To simplify the estimation, it employs an oft-used Bayesian stratagem: conjugate priors. While the Bayesian recipe is universal, it can be computationally tedious. Conjugate priors help simplify calculations a great deal; a conjugate prior is one that keeps the same functional form once multiplied by the likelihood (see "*A note on priors*"). For a binomial likelihood, those priors (and posteriors) look like the beta distribution, whose density f writes:

$$f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1} \tag{5.29}$$

where Γ is the Gamma function and α, β are two real, positive parameters. α may be interpreted as number of times success is observed and β as the number of times failure is observed. The beta distribution is extremely flexible and limited to the interval $[0,1]$, so it is especially useful to model a probability like p (Fig. 5.4).

The companion widget allows to experiment with various choices of α and β[8]. By playing with the widget, you may assess when the choice of the prior parameters matters, and when it does not.

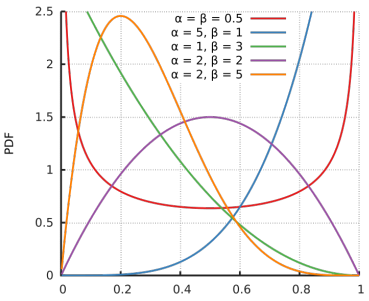


Figure 5.4: Beta distribution for various values of its parameters α and β. Note that it is always zero outside the interval [0,1]. Source: Wikipedia

[8] The parameters that characterize the prior and are called hyperparameters. In principle these can also be associated with distributions (called hyperpriors), though nothing so complicated is required here

Normal model with known variance, unknown mean

Let us apply the Bayesian framework to a somewhat artificial example, simplified to the extreme for the purpose of analytical tractability. Say we are interested in the temperature T at some location. Assume we have the prior distribution (e.g., from historical observations): $T \sim \mathcal{N}(\mu, \tau^2)$. Conditioned on the true value of the state process, $T = m$, assume we have n independent but noisy observations $X = (X_1, \cdots, X_n)$, with standard deviation σ, and thus the data model :

$$X_i|\{T = m\} \sim \mathcal{N}(m, \sigma^2) \tag{5.30}$$

Let's assume for simplicity that we know σ (it comes from a calibrated thermometer of known precision) and we want the distribution of m conditional on observations, $p(m|X)$. What we do know is:

$$p(X|m) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x_i - m)^2}{\sigma^2}\right) \tag{5.31a}$$

$$p(X|m) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2\right] \tag{5.31b}$$

And the prior is, by assumption:

$$p(m) = \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(m-\mu)^2}{2\tau^2}} \tag{5.32}$$

So Bayes' rule [Eq. (5.28)] gives:

$$p(m|X) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\left(\frac{x_i - m}{\sigma}\right)^2 + \left(\frac{m-\mu}{\tau}\right)^2\right]\right\} \tag{5.33}$$

Notice that this is just the product of two Gaussian distributions. It can be shown (by completing the square) that the normalized product is also Gaussian[9], $T|X \sim \mathcal{N}(\mu_p, \sigma_p^2)$, with the following mean and variance:

[9] in other words, the normal prior and normal likelihood conspire to make a normal posterior. We have just found another conjugate prior.

$$\mu_p = \frac{\sum_{i=1}^{n} x_i/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \tag{5.34a}$$

$$\sigma_p^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \tag{5.34b}$$

Put differently: the posterior variance is the harmonic mean of the observation and prior variance, weighted by the number of observations. In that case, the prior amounts to one extra observation, and it will quickly get overwhelmed by the actual observations as long as σ is comparable to or smaller than τ.

The posterior mean (the conditional expectation of the temperature, given the observations) may be re-expressed as:

$$E(T|X) = w_x \bar{x} + w_\mu \mu \tag{5.35}$$

with $\bar{x} = \sum_{i=1}^{n} x_i/n$ (the sample mean of the observations), $w_x = \frac{n\tau^2}{n\tau^2+\sigma^2}$ and $w_\mu = \frac{\sigma^2}{n\tau^2+\sigma^2}$[10]. That is, the posterior mean is a weighted average of the prior mean and the sample mean of the observations. Note that as $\tau^2 \to \infty$ (vague prior), the data model overwhelms the prior and

[10] Verify that $w_x + w_\mu = 1$

$$p(m|X) \to \mathcal{N}(\bar{x}, \sigma^2/n) \tag{5.36}$$

(as per the central limit theorem). Alternatively, for fixed τ^2, but very large amounts of data (i.e., $n \to \infty$) the data model again dominates

the posterior density. On the other hand, if τ^2 is very small, the prior is critical for comparatively small n (it is highly informative). Though these properties are shown for the normal data, normal prior case, it is generally true that for very large datasets, the data model is the major controller of the posterior.

To illustrate this, assume the prior distribution is $m \sim \mathcal{N}(20, 3)$, and the data model is $X_i|m \sim \mathcal{N}(m, 1)$. We have two observations $x = (19, 23)$. The posterior mean is $20 + (6/7)(21 - 20) = 20.86$ and the posterior variance is $(1 - 6/7)3 = 0.43$. Fig. 5.5 shows these distributions graphically. Since the data are relatively precise compared to the prior, we see that the posterior distribution is closer to the likelihood than the prior. Another way to look at this is that the data weight $w_x = 6/7$ is close to one, so that the data model is weighted more than the prior.

Next, assume the same observations and prior distribution, but change the data model to $X_i|m \sim \mathcal{N}(m, 10)$. The data weight is now $w_x = 6/16$ and the posterior distribution is $\mathcal{N}(20.375, 1.875)$. This is illustrated in Fig. 5.6. In this case, the weight is closer to zero (since the measurement error variance is relatively large compared to the prior variance) and thus the prior is given more weight.

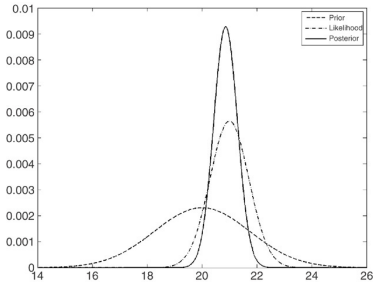


Figure 5.5: Posterior distribution with normal prior and normal likelihood; relatively precise data. From *Wikle and Berliner* (2007)

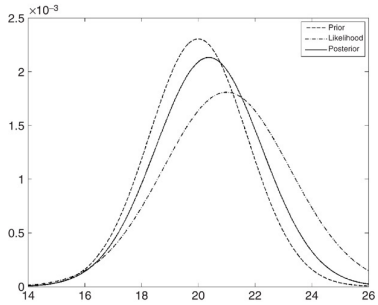


Figure 5.6: Posterior distribution with normal prior and normal likelihood; relatively imprecise data. From *Wikle and Berliner* (2007)

Normal model with unknown mean and variance

In the case where σ is unknown, things get a little more difficult, but still tractable. *Gelman et al.* (2013, chap2) show that a convenient prior for σ is the scaled inverse χ^2 distribution, parameterized by its degrees of freedom ν_0 and a scale parameter σ_0[11]. They show a very neat result: in such a case the role of prior can be thought of as providing the information equivalent to ν_0 independent observations with average precision σ_0. This illustrates once more that the role of prior, mathematically, is to provide a starting point for the inference. In cases where the observations are too few, or too poor, the choice of the prior will be important to the final outcome (the posterior distribution); in cases where observations are numerous and/or good, its role is diluted by the observations.

[11] the inverse χ^2 distribution is a special case of the inverse gamma distribution, plotted in Fig. 5.7

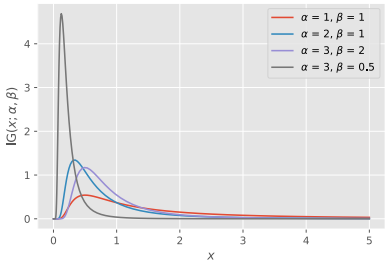


Figure 5.7: PDF of the inverse gamma distribution for various values of its parameters α and β. If $X \sim$ Scale-inv-$\chi^2(\nu, \sigma^2)$ then $X \sim$ Inv-Gamma $\left(\frac{\nu}{2}, \frac{\nu\sigma^2}{2}\right)$.

A note on priors

Three notable properties of prior distributions deserve mention here:

- **Conjugacy**: As we've seen, conjugate priors are priors that, combined with the likelihood, produce a posterior distribution that has the same form as the likelihood. For instance, the normal model above with a normal prior on the observations' mean yielded a normal posterior. For a binomial likelihood you'd choose a beta prior. The main justification for using such priors is that it enables a closed-form expression for the posterior; otherwise, the posterior has to be

found numerically, sometimes at considerable expense. Clearly the choice of a conjugate prior is a matter of convenience, and may not always reflect our actual scientific knowledge about the parameters we wish to discover from the data. They also have great pedagogical value, but should be avoided in scientific applications unless they have a compelling scientific rationale.

- **Objectivity**: Uninformative priors, sometimes called flat priors, diffuse priors, vague priors or (more boldly) "objective priors", attempt to provide as little information as possible to let the data speak. In ideal circumstances (with good enough data), one may choose an essentially flat prior (e.g. a uniform distribution) for a parameter, and the likelihood will overwhelm it, leading to an estimate that is very close to objective. Note that conjugacy and objectivity are not mutually exclusive. Any prior may be made uninformative by expanding its scale parameter to large values.

- **Han Solo:** For an entertaining view on how to choose a prior for the odds of Han Solo to make it through an asteroid field, read this blog post by Count Bayesie. The bottom line is that sometimes it is sensible to be subjective, especially if you know Hollywood.

Non-analytical cases

As said above, in most real-world cases the posterior distribution does not have a closed-form solution. In such cases, one must evaluate the various integrals numerically, often over multiple dimensions, and this is a topic of considerable complexity. Some of this can be coded in your favorite language[12], though more specialized packages like STAN are probably preferable for a first-timer. Either way, most applied Bayesians spend the majority of their time hunched over a computer, as illustrated in Fig. 5.8.

[12] *e.g.* PyMC3 in Python

Bayesian pros and cons

Spinal health notwithstanding, there are many advantages to the Bayesian method of inference. First, it allows to use scientific knowledge as encoded by a prior, hence using all relevant knowledge to the analysis of a dataset. As such, it mimics the scientific mind, updating intuition and prior experience in the light of observations. Second, the posterior distribution is a much richer output than an MLE point estimate, and immediately characterizes everything we want to know about the model parameters: their most likely value (posterior mode), their central tendency (posterior mean), their uncertainty (posterior variance or IQR). In the case of a flat prior, the posterior mode corresponds to the MLE,

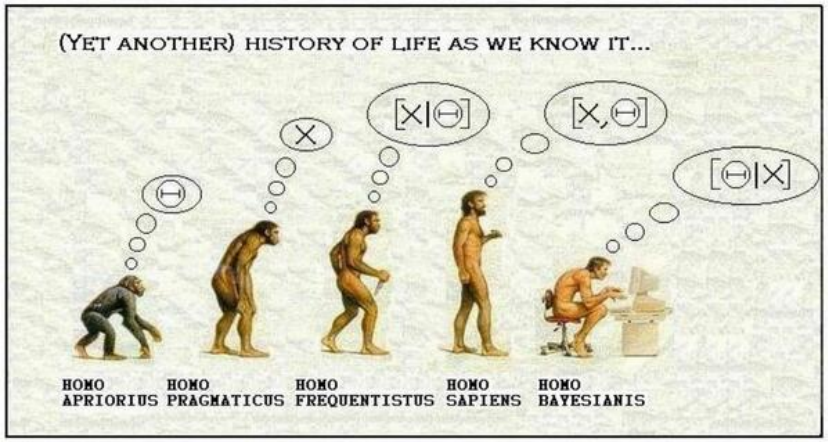


Figure 5.8: The evolution of statisticians: *Homo apriorus* only considers his prior knowledge $p(\theta)$, without regard for observations; *Homo Pragmaticus* only considers the observations (X), which is not much better; *Homo frequentistus* considers the likelihood (the conditional distribution of observations given the parameters), $f(X|\theta)$; *Homo sapiens*, being wiser, considers the *joint distribution* of parameters and observations; finally, *Homo Bayesianis*, the wisest of all, considers the posterior distribution of parameters given the observations $p(\theta|X)$, but needs an awful lot of coding to do that.

but the posterior also provides uncertainty quantification for the same price.

One particularly neat use of the posterior is the ability to *forecast* new observations. The *posterior predictive distribution* is the distribution of a new data point $\tilde{x}$, marginalized over the parameters:

$$p(\tilde{x}|X) = \int_\theta p(\tilde{x}|\theta)p(\theta|X)\,\mathrm{d}\theta \tag{5.37}$$

Notice how the left hand side depends solely on the observations, but no longer on the parameters. That is, past observations have been digested by the statistical model so it can produce a probabilistic forecast of a new estimate. In the normal case with known variance:

$$p(\tilde{x}|X) \propto \int_{\mathbb{R}} \exp\left[-\frac{1}{2}\left(\frac{\tilde{x}-m}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{m-\mu_p}{\sigma_p}\right)^2\right] dm \tag{5.38}$$

so it can be seen that integrating over the previously unknown parameter m has made it disappear, and we are now predicting future (testable) observations given only past observations, which is quite appealing to a scientist.

So what can possibly go wrong with the Bayesian approach? As said earlier, accommodating scientifically-defensible priors may require very complex integrals, and therefore very heavy computations (e.g. Monte Carlo Markov Chains). These are far beyond the scope of this book, but you may want to dig in this direction if Bayesians have taken over your field of study. *Tarantola* (2004) is an excellent introduction for geophysicists.

Chapter 6

CONFIRMATORY DATA ANALYSIS

> One must credit an hypothesis with all that has had to be discovered in order to demolish it.
>
> *Jean Rostand*

Now we're getting to what most laypeople think of when they hear the word 'statistics': putting measurements to the test, and finding whether the results are *significant*. This is usually taken as a synonym of rigor, but we shall see that care is needed: sometimes, applying the wrong test or testing the wrong hypothesis is worse than not doing any statistics at all.

I Preamble

Two tribes war over the ownership of a piece of land, which each claims to have first occupied before the other. They ask a geochronologist to arbitrate their dispute. The geochronologist finds wood and bone artefacts that allow the site to be radiocarbon-dated and the claims evaluated: specifically, tribe A claims to have occupied the region since 622 AD, while tribe B claims they got there in 615 AD.

The following dates come back:
$$\begin{cases} \hat{t}_A = 650 \pm 50\text{y}\,, \\ \hat{t}_B = 750 \pm 50\text{y}\,. \end{cases} \quad (1\sigma)$$

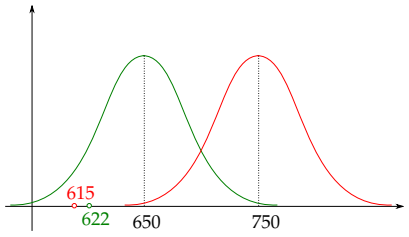


Figure 6.1: Age distributions for artifacts from the two tribes. In green we have the data from tribe A with mean of 650 years and standard deviation of 50 years. In red we have the data from tribe B with mean of 750 years and standard deviation of 50 years.

The geochronologist asks you three questions:

1. How confident can I be in each date? → **Confidence or credible Intervals**

2. Are the observations compatible with their claims? → **Hypothesis Test**

3. How confident should I be that tribe A got there first? → ***p*-value**

II Confidence Intervals

How confident are we that the true age lies within a certain range? Assume $t_A \sim N(\hat{\mu}_A, \sigma_A)$, where $\hat{\mu}_A = 650$ y and $\sigma_A = 50$ y. We want to find t_{min} and t_{max} such that :

$$\mathbb{P}(t_{min} \leq t_A \leq t_{max}) = 95\%$$

Transformation to a Z Score (standard normal)

Defining: $Z_A = \frac{t_A - \hat{\mu}_A}{\sigma_A} \quad \rightarrow \quad N(0,1)$,

$$\mathbb{P}(z_{min} \leq Z_A \leq z_{max}) = P(|Z_A| \leq z_\alpha) \qquad \text{(Symmetry)} \qquad (6.1)$$

Find z_α such that $P(|Z_A| \leq z_\alpha) = (100 - \alpha)\%$, where α is the **confidence level**.

$$\alpha = 0.05 \quad \Rightarrow \quad \begin{cases} Z_{0.025} = \Phi(0.025) = -1.96\,, \\ Z_{0.975} = \Phi(0.975) = +1.96\,. \end{cases}$$

A normal RV spends 95% of its time within 1.96 of the mean. That is, it lies within $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ with 95% confidence.

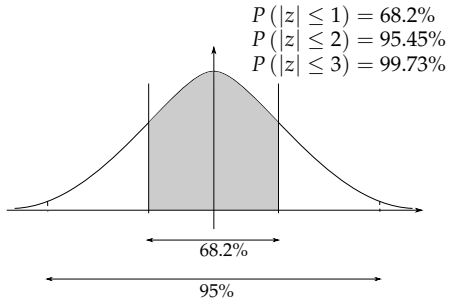


Figure 6.2: Confidence level of each confidence interval chosen.

Tribe A: $Z_A = \frac{t_A - \hat{\mu}_A}{\sigma_A} \quad \Leftrightarrow \quad t_A = \hat{\mu}_A + \sigma_A Z_A$,

$$\Rightarrow P(\hat{\mu}_A - 1.96\sigma_A \leq t_A \leq \hat{\mu}_A + 1.96\sigma_A) = 95\%\,.$$

A 95% C.I. for the arrival of tribe A is $[552, 748]$. A 95% C.I. for the arrival of tribe B is $[652, 848]$.

General Case

Find $z_{\alpha/2}$, $z_{1-\alpha/2}$ such that,

$$\mathbb{P}(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha\,. \qquad (6.2)$$

- Non linear problem (involves inverse CDF).
- Asymmetric if skewed distribution (*e.g.* Weibull, Gamma, Poisson).

Interpretation

A confidence interval is such that, if we were to repeat the experiment a great many times, the true value t_A would be in this interval 95% of the time. Notice that it is subtly different from saying we are "95% sure that t_A in this interval". Such is the difference between (frequentist) confidence intervals and (Bayesian) credible intervals[1].

[1] An irony of orthodox statistics is that, if you quiz most practitioners on what a CI is, the vast majority of them (though perhaps not 95%) will give you the definition of a credible interval. It's not too surprising, because it's what people actually want to know, and if they truly understood CIs the way frequentists do, they would probably abandon the concept.

III Testing Archeological Hypotheses

Working hypothesis: **Tribe A arrived before Tribe B.** Do the data support this idea, and how confident are we about the statement?

Known Variance Case (Z-test)

Let us assume that we know the measurements' precision:

$$\sigma_A = \sigma_B = \sigma = 50y\,.$$

Now, Tribe A claims that they arrived at $t_A = 622$. So we define:

Null Hypothesis

$$\mu_A = 622 \quad (H_0)$$

And we proceed to evaluate the probability of observing the data under this hypothesis. If H_0 is true,

$$\begin{aligned}
\mathbb{P}(t_A \geq 650) &= \mathbb{P}\left(\frac{t_A - 622}{50} \geq \frac{650 - 622}{50}\right) && \text{(6.3a)}\\
\mathbb{P}(t_A \geq 650) &= \mathbb{P}\left(z_A \geq \frac{650 - 622}{50}\right), z_A \sim \mathcal{N}(0,1) && \text{(6.3b)}\\
\mathbb{P}(t_A \geq 650) &= \mathbb{P}\left(z_A \geq 0.56\right) = 1 - \Phi(0.56) && \text{(6.3c)}\\
\mathbb{P}(t_A \geq 650) &\simeq 29\% && \text{(6.3d)}
\end{aligned}$$

where, once again, we have transformed our normal variable to a Z score so we can compute everything in terms of Φ, the standard normal CDF. If H_0 were true, we thus would observe ages as large as 650 about 29% of the time from chance alone ; this is not a rare occurrence, and means that the data do not allow to reject the null hypothesis at the 5% level. Equivalently, we find *insufficient evidence to exclude this possibility* ("not guilty" is not the same thing as innocent). Put differently, if H_0 were true, then t_A should lie within $[622 - 1.96 \times 50, 622 + 1.96 \times 50] = [524, 720]$. $\hat{t}_A = 650$ is in this interval, but $\hat{t}_B$ is not. Note:

- This is called a **Z test** on account of the use of the standard normal.
- A 5% test level means "we have a 5% chance of observing a result as extreme as this by random chance". (see Sect. V)
- $t_B \notin$ interval means that we could resoundly (at the 5% level) reject the hypothesis that tribe B got there even as late as 622 AD; the tribe's claim claim that $\mu_B = 615$ is even less convincing (the associated p-value would be smaller still).

Indeed, the probability computed above is a p-value, a fundamental notion of confirmatory data analysis, and one so misused that the American Statistical Association issued a cautionary statement on it in 2016. Some journals have even taken the drastic measure of banning them altogether. So, with that context: *a p-value is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true*. A p-value is emphatically **not** the probability that H_0 is true (or false, for that matter). Rather, it's used to quantify how much evidence the data provide against H_0. The smaller the p-value, the more suspicious we are that the data could have been generated under the null hypothesis. We say that the test is rejected if the p-value falls under the (prescribed[2]) test level: $p \leq \alpha$. If such is not the case, we say... nothing. We shut up and we collect more data. Or we go talk to a Bayesian statistician and try to squeeze more information out of our data.

[2] It bears repeating that α is chosen by the analyst, and there is nothing magical about 5%. If a test passes at the 5.1% but not the 5% level, it does not mean that the result is "insignificant". It only means that under the null hypothesis, results as extreme as the one observed happen about 5% of the time from chance alone.

Unknown Variance Case (T-test)

Assume now that we don't know the measurement's precision; we have to estimate it. To that end, our geochronologist takes repeated samples $\{t_{A_1}, t_{A_2}, \ldots, t_{A_n}\}$ and estimates:

$$\overline{t}_A = \frac{1}{n_A} \sum_{i=1}^{n_A} t_{A_i} \qquad \text{(Sample mean)}$$

with $n_A = 12$ and *idem* for $\overline{t}_B$ with $n_B = 9$.

Question *Are the data still compatible with $t_A = 622$?*

Question *What is the probability that $t_A > t_B$?*

Assuming again that $t_A \sim N\left(\mu_A, {\sigma_A}^2\right)$, then:

$$E\left(\overline{t}_A\right) = \mu_A\,,$$
$$V\left(\overline{t}_A\right) = \frac{\sigma_A^2}{n_A}\,, \qquad \text{(Central limit theorem)}$$

if σ^2 is known. If it is not known, estimate it via the sample variance:

$$S_A^2 = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} \Big(\underbrace{t_{A_i} - \overline{t}_A}_{\substack{iid\ \mathrm{RVs} \sim \mathcal{N}\left(0, \sigma_A^2\right)}} \Big)^2 \tag{6.4}$$

Note that $t_{A_i} - \overline{t}_A \sim \mathcal{N}\left(0, \sigma_A^2\right)$, so S_A^2 is the sum of n squared, normal RVs with zero mean and identical variance σ_A^2 (precisely what we seek

to estimate). We can of course scale them so that they have unit variance, for simplicity. $(n-1)\frac{S^2}{\sigma^2}$ is then the sum of n iid standard normal random variables, **squared**. It turns out that solving this problem was sufficiently important that $(n-1)\frac{S^2}{\sigma^2}$ got its own distribution, known as a *chi-squared distribution*, dependent only on the number of elements in the sum (ν), and denoted χ^2_ν. It is a special case of Gamma distribution (cf Fig. 6.3), with the following properties:

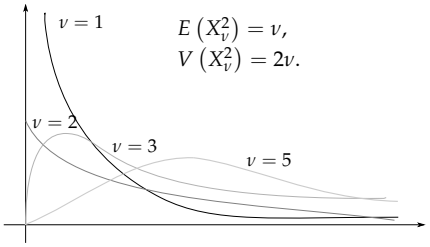


Figure 6.3: χ^2 distribution for different values of ν. We can see that for large ν, it converges to a Gaussian, $\mathcal{N}(\nu, 2\nu)$, as per the central limit theorem. Note that ν need not be an integer.

$$\chi^2_\nu = \sum_{i=1}^{n} Z_i^2, \qquad \begin{cases} Z_i \sim \mathcal{N}(0,1) & IID \\ \nu = n-1 = \text{"degrees of freedom"} \end{cases} \tag{6.5}$$

$$\chi^2_\nu \sim \Gamma\left(\frac{\nu}{2}, 2\right). \tag{6.6}$$

$$i.e. \qquad f_{\chi^2_\nu}(x) = \frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{\Gamma\left(\frac{\nu}{2}\right) 2^{\frac{\nu}{2}}} \tag{6.7}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\underbrace{x_i}_{\mathcal{N}(\mu,\sigma^2)} - \overline{x})^2 \qquad \Rightarrow \qquad (n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\underbrace{Z_i^2}_{\mathcal{N}(0,1)} \sim \chi^2_{n-1}.$$

Now you may ask: why do we only have $n-1$ degrees of freedom when we added up n independent numbers? The answer is that 1 degree of freedom went into estimating the mean, so we lost that.

Now recall the properties of the sample variance:

1. $E\left(S^2\right) = \sigma^2$, Unbiased Estimator.
2. $V\left(S^2\right) = \frac{2\sigma^4}{n-1}$, (uncertainty about S also decreases as $\sqrt{n}$)

So it would make sense to consider the *test statistic*:

$$\hat{T}_A \equiv \frac{\overline{t}_A - \mu_A}{S_A/\sqrt{n_A}} \tag{6.8}$$

Dividing numerator and denominator by σ_A, we see that it is basically the ratio of a standard normal to the square root of a chi-squared variable. Informally, one may write:

$$T_A \sim \frac{Z}{\sqrt{\chi^2_{n-1}}}$$

So what, you say? Again, solving this problem was, at some point, sufficiently important that someone went through the trouble of working out this distribution analytically, naming it **Student's T (or t) distribution**[3]: t_ν, depicted in Fig. 6.4. In Python, it may be accessed via `scipy.stats.t()`. In the case of the mean of n IID normal RVs, $\nu = n-1$.

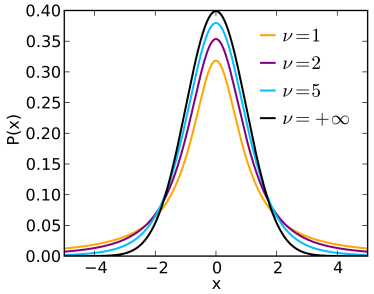


Figure 6.4: PDF of Student's T distribution, which is close to Gaussian with fatter tails: that is, having to estimate the variance introduces uncertainty about the estimate of the mean. However, for $\nu \to \infty$, $t_\nu \to N(0,1)$, in another splendiferous manifestation of the central limit theorem. Credit:Wikipedia

[3] In the English-language literature it takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student". Gosset worked at the Guinness Brewery in Dublin, Ireland. One version of the origin of the pseudonym is that Gosset's employer forbade members of its staff from publishing scientific papers, so he had to hide his identity. Another version is that Guinness did not want their competitors to know that they were using the t-test to test the quality of raw material. Either way, beer and statistics do mix. Credit: Wikipedia

T-TEST: We distinguish two kinds:

One Sample T-Test

Let us define the null hypothesis H_0: "$\mu_A = 622$" and the alternate Hypothesis: H_a: "$\mu_A > 622$"[4]. Can the data distinguish between the observed and hypothesized means? Let us compute the T statistic[5]:

$$\hat{T} = \frac{\bar{t}_A - \mu_A}{S_A/\sqrt{n_A}} = 1.94$$

$$\downarrow$$

Is it significantly different from 0?

Then compute $\mathbb{P}\left(t > \hat{T}\right) = 1 - F_{t_\nu}(1.94)$ where F_{t_ν} is the CDF of the t-distribution with ν degrees of freedom. This is the p-**value**[6].

- $n = 4, \mathbb{P}\left(T > \hat{T}\right) = 13\%$. We fail to reject H_0[7].
- $n = 12, \mathbb{P}\left(T > \hat{T}\right) = 4\% \rightarrow$ We reject H_0.

Put differently, with $n = 4$, there is insufficient evidence to distinguish between 650 and 622 at the 5% level; with $n = 12$ there is sufficient evidence to do so. So the data suggest that tribe A is exaggerating a little, but is fundamentally honest. On the other hand, tribe B is either lying or delusional. Let us ask a more relevant question for our dispute: "Did tribe B arrive **significantly** after tribe A?"

Two Sample T-Test

When the means of two samples must be compared, with unknown variances, we use the two sample T-test. That is, we compare $\bar{t}_A$ and $\bar{t}_B$, with the null hypothesis H_0: "$\mu_A = \mu_B$" and the alternate hypothesis H_a: "$\mu_B > \mu_A$"[8]. We compute the two-sample T statistic:

$$T = \frac{\bar{t}_B - \bar{t}_A}{\sqrt{\frac{S_B^2}{n_B} + \frac{S_A^2}{n_A}}} \sim t_{\nu'}\,.$$

The value of ν' depends on the case:

Equal Variance $\quad \nu' = \dfrac{n_A n_B}{n_A + n_B}\,,$ (Harmonic Mean)

Unequal Variance $\quad \nu' = \dfrac{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}{\left(\frac{S_A^2}{n_A}\right)^2 \frac{1}{n_A - 1} + \left(\frac{S_B^2}{n_B}\right)^2 \frac{1}{n_B - 1}}\,.$

[4] this is a one-sided test; we could also test for $\mu_A < 622$, but it wouldn't make sense here since we can safely assume that either tribe is giving us its earliest possible date of arrival at the site

[5] a deterministic function of the data

[6] Here, one would compute 1-`scipy.stats.t.cdf`$(1.94, n_A - 1)$

[7] this double-negative is essential to hypothesis tests: we start from a presumption of innocence (H_0 is true) and see if the data can convince us otherwise. The choice of H_0 is therefore crucial, and we should be careful that it does not stack the cards against a particular result, as is sometimes the case

[8] By this point we have dealt with tribe B enough that we suspect that it did not get there first, so we consider a one-sided alternate hypothesis

$$E.g.: \begin{cases} n_A = 12 \\ n_B = 9 \end{cases} \rightarrow T \simeq -4.536.$$

As expected, the number is negative (since $\bar{t}_B > \bar{t}_A$), but is it significantly different from zero? For this we compute the p-value:

$$p\text{-value} = \texttt{scipy.stats.t.cdf}(\hat{T} = -4.536, \nu = 108/21) \approx 3 \times 10^{-3} \ll \alpha = 5 \times 10^{-2}$$

That is, if the two distributions had equal means, it would be exceedingly unlikely to observe a deviation this big. We can thus reject the null hypothesis with very high confidence: we conclude that tribe A colonized the land first. Note that we would not be able to make this claim with the same level of confidence with $n_A = n_B = 2$ (HW: what p-value would we get?), so it was critical to collect more measurements.

Application: $p_{val} = f(n_A, n_B)$ can tell you how many measurements you need to achieve a certain level of confidence in a result. When designing experiments, this tell us how many are needed to claim a "**significant difference**" – a useful approach to convince a program manager that you need monies to go collect data in faraway locales, or buy a new instrument. This touches on the rich topic of *experiment design*.

IV The Logic of Statistical Tests

To summarize, there are five ingredients to a statistical test:

1. Identify the test and the test statistic (*e.g.* T-test).
2. Define the null hypothesis (*e.g.* "H_0: $\mu = \mu_0$").
3. Define an alternate hypothesis (*e.g.* "H_a: $\mu > \mu_0$").
4. Obtain the *null distribution* (distribution of the test statistic assuming H_0 is true).
5. Compute the *p-value* (probability of occurrence of values as extreme as the observed test statistic) and compare to the test level α.

   $p < \alpha \Rightarrow$ reject null hypothesis (guilty).
   $p \geq \alpha \Rightarrow$ insufficient evidence (not guilty $\neq$ innocent).

Example
A Palm Spring resort claims that $f_c = \frac{6}{7}$ of its days are cloud-free. Yet, observations over 25 days indicate that only $\frac{15}{25}$ days are cloud-free. Is the claim supported by the evidence?

1. *Test statistic:* $f = \frac{k}{N} = \frac{15}{25}$.
2. *Null hypothesis:* $f = f_c = \frac{6}{7} \simeq 0.857$.

3. *Alternate hypothesis:* $f < f_c$ *(claim is overblown).*

4. *Null distribution: Sequence of* 25 *Bernoulli trials with probability* f $X \sim \mathcal{B}(25, f_c)$*, that is:*

$$\mathbb{P}(X = k | f = f_c) = \binom{25}{k} f_c^k (1 - f_c)^{25-k} . \qquad \textit{Binomial Distribution} \tag{6.9}$$

5. *p-value:*

$$\begin{aligned} \mathbb{P}(X \leq 15) &= \sum_{k=0}^{15} \binom{25}{k} f_c^k (1 - f_c)^{25-k} \\ &\simeq 0.0015 \qquad \textit{Very unlikely} \end{aligned}$$

We conclude, with high confidence, that the data are inconsistent with the claim that $f = \frac{6}{7}$.

V Test Errors

How might a test lead us astray? There are two main errors that we should watch for. A type-one error is a *false positive*: finding an effect when there is no effect. Medical examples of this include determining a practice to be unsafe when it is safe, or an intervention non-beneficial when it is in fact beneficial. A type-two error is the reverse, a *false negative*: a practice is determined to be safe when it is unsafe, or an intervention beneficial when it is not. More precisely:

Type I error $= \alpha = \mathbb{P}\left(\text{rejecting } \mathrm{H}_0 / \mathrm{H}_0 \text{ true}\right) =$ **False Positive**.

Type II error $= \beta = \mathbb{P}\left(\text{accepting } \mathrm{H}_0 / \mathrm{H}_a \text{ true}\right) =$ **False Negative**.

No simple relationship between α and β, but they are impossible to jointly minimize $\rightarrow$ there is a **trade off**. We summarize the decisions (and their correctness) in a table:

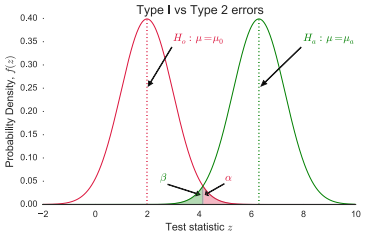


Figure 6.5: This plot represents the false positive, α (red shading), and false negative, β (green shading). We can see that if β increases, α will decrease, though there is in general no simple equation to relate the two.

Decision	H_0 is true	H_0 is false
Reject H_0	False positive (Type I error)	True negative
Fail to reject H_0	True positive	False negative (Type II error)

Table 6.1: Possible outcomes of a statistical test, and associated errors

A type-three error, in contrast, is one where the analysis itself is framed incorrectly and thus *the problem is mischaracterized*. This one is a more of a logical fallacy than a statistical error, and cannot therefore be spotted by statistics, but it is very common: keep your eyes peeled for it, and try your best never to commit that sin!

The **power** of the test is defined as:

$$1 - \beta = \mathbb{P}\left(\text{rejecting } H_0 / H_a \text{ is true}\right) \tag{6.10}$$

- Measures how discriminatory it is.
- Always depends on the number of observations.

Typically, the more observations, the more power to discriminate, but this obviously depends on their quality; sometimes one good observation is worth 10 bad ones. But if they are truly independent, the central limit theorem tells us that a million bad ones can beat a few good ones!

VI Common Parametric Tests

We now visit common parametric tests, so named because they rely on parametric distributions. In fact, all these rely exclusively on the normal distribution, more specifically normal IID observations (some adjustments can be made for dependent data).

Z-Test

Comparing 2 means where the variance is known: see Sect. III

T-Test

Comparing 2 means when both means and variance are unknown. see Sect. III

F-Test

This test compares the variances of two samples from two populations.

$$\begin{cases} \text{Sample 1}, \sigma_1, n \text{ observations} \\ \text{Sample 2}, \sigma_2, m \text{ observations} \end{cases} \qquad F_{m,n} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \frac{\chi^2_{n-1}}{\chi^2_{m-1}}\,.$$

[9] Its CDF maybe accessed via `scipy.stats.f.cdf(F, n, m)`

(Ratio of two sample variances normalized by their variances)

$\mathbf{H}_0$: "$\sigma_1 = \sigma_2$" $\rightarrow$ the 2 samples have identical variance. so the ratio should be close to unity. How close is "close"? In other words, how large a deviation should we observe before we pronounce the two variances as distinct? We quantify this via the F distribution, which is known but complicated (non-analytical), and depends only on two degrees of freedom n and m[9].

Testing for Goodness of Fit

Frequently, one is led to the question: does a theoretical distribution accurately represent a given empirical CDF? Here we give two approaches to answer this (there are many more).

χ^2 Test

The idea is to plot the histogram of the theoretical and empirical PMFs (*e.g.* Fig. 6.6). If the fit is good, then the observed number of events in bin k (O_k) should be close to the expected number (E_k). To quantify this, we compute the statistic:

$$\Xi^2 = \sum_{k=1}^{N_b} \frac{(E_k - O_k)^2}{E_k}, \tag{6.11}$$

It turns out that

$$\Xi^2 \sim \chi^2_{\nu-1}, \tag{6.12}$$

where $\nu = N_b - n_p$ with n_p the number of parameters estimated. One can therefore use the χ^2 inverse CDF to test whether the fit is acceptable for a given number of bins[10]. As the number of observations increase, so does the number of allowable bins, and $\Xi^2 \sim \mathcal{N}(0,1)$.

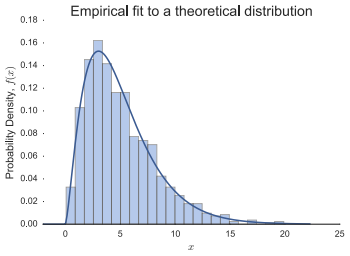


Figure 6.6: An example of a theoretical PDF (solid line) fit to an empirical PMF (columns)

[10] `scipy.stats.chisquare.ppf()`

Kolmogorov-Smyrnov Test

The KS test[11] relies on the existence of a universal distribution for the following statistic:

$$D = \max_x |F_n(x) - F(x)| \qquad \begin{array}{l} F_n(x) = \text{empirical CDF} \\ F(x) = \text{reference CDF} \end{array} \tag{6.13}$$

The critical value for rejection at level α is defined as

$$C_\alpha = \frac{k_\alpha}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}} \tag{6.14}$$

where n is the sample size and k_α is a function of α only.

One advantage is that it is universal: for any reference distribution $F(x)$, this can tell us where the sample that generated $F_n(x)$ was drawn from the same population. One problem is that the test may not be stringent enough if some data has been used to estimate parameters (an example of double-dipping). Also, for some specific distributions, one may obtain more powerful tests (*i.e.* tests with greater statistical power $1-\beta$) by exploiting the functional form of the distribution.

[11] `scipy.stats.kstest()` for a fit to a theoretical distribution, `scipy.stats.ks_2samp()` to determine if 2 samples come from the same distribution

The Anderson-Darling test[12] is an avatar of the KS test against a few common continuous distributions (e.g. normal, exponential, gumbel, logistic). It is more powerful (statistically speaking) because less general. There are other specialized tests to determine if a sample is consistent with having come from a normal distribution (e.g. the Jarque-Bera test). Use them with caution, as they are often a referendum on sample size: with low sample size, every normal-ish dataset will pass the test with flying colors; with high enough sample size even data that are actually normal have a decent chance of being flagged as non-normal.

[12] `scipy.stats.anderson()`

VII Non-parametric tests

Parametric tests are incredibly useful, but rely heavily on the IID Gaussian assumption, so they may break down outside of this rather restrictive context. In this day and age those assumptions seem rather quaint, but remember that they were born out of necessity (not ignorance) at a time when analytical approximations were the only salvation. In the computer age, we can devise *non-parametric tests* that do not rely on such assumptions. The key to such tests is to generate a large ensemble of *surrogate data*, from which we obtain the null (sampling) distribution. Once this distribution is obtained, we can use it to do all manner of uncertainty quantification, like forming confidence intervals and testing hypotheses. Here we discuss four main ideas, which all fall under the more general umbrella of *Monte Carlo methods*[13]:

1. permutation
2. reordering
3. resampling
4. direct simulation

Note that first three assume *exchangeability*, which is closely related to the IID assumption, but they need not assume Gaussianity.

[13] One abuse of language is that – in the Earth Sciences at least – Monte Carlo is often used interchangeably with resampling plans, whereas Monte Carlo methods are a broader class of computational algorithms that rely on repeated random sampling to solve complex integrals not amenable to closed form solutions. Monte Carlo Markov Chains (MCMC) are a prime example of this in the Bayesian world (Chap 5, Sect. V)

Permutation Tests

Imagine that we have a sample of size $n = n_1 + n_2$

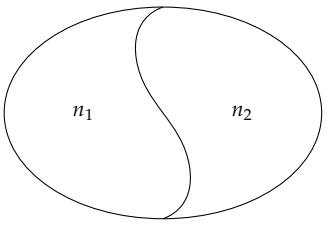


Figure 6.7: Samples from two populations of different sizes

We can generate N surrogate replicates by sampling without replacement

$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2} = \frac{(n_1 + n_2)!}{n_1!\, n_2!} \tag{6.15}$$

For example, imagine that we have two climatic scenarios, $1 \times CO_2$, $2 \times CO_2$, and the corresponding 2×16 simulations. We would like to know if California rainfall is significantly different between the two ensembles. One way to judge this significance is to choose a test statistic (*e.g.* the difference in maximum winter rainfall between one group of simulations and the next, a non-normal statistic for which parametric assumptions would fail), create a large ensemble of permutations, and use it to obtain the sampling distribution. If the observed value of the statistic seems unlikely under these permutations, this would be a strong indication that the two populations are distinct.

Reordering Tests

Skip (not common in Earth science)

Resampling plans

There are two main ones:

Bootstrap : thus called because it is a way to magically generate a large ensemble from a limited sample, hence allowing you to pull yourself up by your bootstraps. The idea is to generate surrogate data by sampling with replacement from the original data (Fig. 6.8); there are n^n possible combinations.

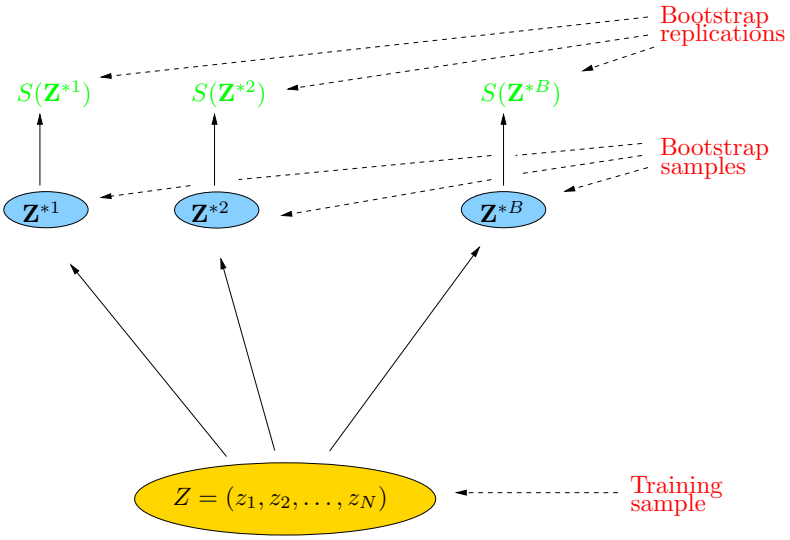


Figure 6.8: Schematic representation of bootstrap sampling for a sample Z and a test statistic $S(Z)$. The ensemble thus created allows to estimate a sampling distribution, or various numerical summaries, like a sample mean, variance, etc. It can be shown that this distribution asymptotically converges to the true one.

The idea is to generate B bootstrap samples, compute the test statistic for each of them, then sort and find the $B \times \alpha/2$ smallest and $B \times (1 - \alpha/2)$ largest values, from which we get the α-level confidence interval. However, the draws are obviously not independent, as each datum has probability e^{-1} of appearing in each sample. In some cases, consecutive observations might not be exchangeable, but large enough blocks of them might reasonably be assumed to be. This is a job for the *block bootstrap*, an example of which is shown in Fig. 6.9.

What value of B should we pick? As a rule of thumb, distributions start looking reasonable around $B = 200$, but if possible, why not go for 1,000 or 10,000? More would probably be a waste of CPU time.

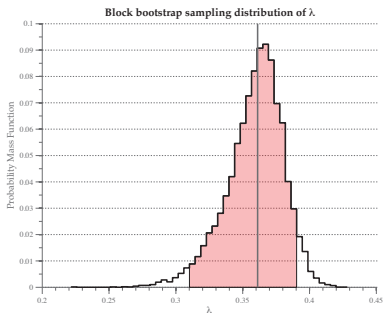


Figure 6.9: Uncertainty quantification for a parameter λ following a non-normal distribution. The sampling distribution (staircase) was obtained by 2,000 bootstrap realizations using 20-year long chunks of a 10,000 year long timeseries at biannual resolution. The shaded area depicts a 95% confidence interval for $\hat{\lambda}$ under this distribution; the mean is indicated by a vertical line.

Jackknife ("leave one out"): the idea is recompute a quantity by leaving out each datum in turn, that is, defining each

$$\text{sample } j = \{1, 2, \underset{j-1}{\ldots}, \underset{j}{\square}, \underset{j+1}{\ldots}, n\} \tag{6.16}$$

and repeating this n times (one for each observation). For a sample of size n, we have only n possibilities, each sample having $n-2$ common points with another. The jackknife is in fact a special case of the bootstrap; it is $\sim$ 20 years older, and much cheaper. In this day and age it is not particularly recommendable, but it does prove useful to gauge the sensitivity of a result to one particular observation. It typically yields much less accurate confidence intervals or p-values than the bootstrap, however

There are canned codes to perform these resampling plans automatically, but they are a black box, so only use them if you're confident that you know what each is doing. Even better, write your own[14], and use the tools you already know to compute the test statistic of your choice, its empirical CDF, quantiles, and relevant numerical summaries.

[14] using `np.random.choice()`, or something similar, to generate the samples

Direct simulation

For our purposes, if we've determined a particular statistical model (*e.g.* AR(1), see Chapter 8) to be an appropriate null hypothesis for the process that generated the data, one can simulate a large number of samples (say $N_{mc} > 200$) from that model, compute the ECDF of a statistic of interest and estimate the probability of observing the observed data under this null distribution. We will use this in the *isopersistent* and *isospectral* tests (Lab 7).

Example: significance of correlations

A common problem is to estimate the significance of correlation coefficients. A famous example is the high correlation ($r = 0.52$) between the number of GOP senators and ... sunspots! (Fig. 6.10). Is the correlation significant? Does it imply some causation? If so, in which direction? I'll let you think about that.

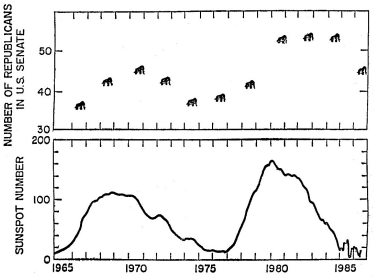


Figure 6.10: Republican Senators and Sunspots, 1960-1986. Figure attributed to Richard Lindzen, MIT. For more details, see Fun with correlations. For a hilarious account of extremely high correlations that cannot possibly harbor any causal relationships, see Tyler Vigen's spurious correlations

Standard theory says that for 2 normal IID series, the test statistic $t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_\nu$ with $\nu = n-2$ degrees of freedom. So it is common to test against the null hypothesis $|r| = 0$ using a Student's distribution as the null distribution. This is all fine and good except when the IID assumptions are not met, which happens all the time in in Earth sciences. In particular, many geophysical datasets exhibit high serial correlation, meaning that consecutive observations are not independent (*e.g.* the temperature on Tuesday depends on the temperature on Monday). This means that the *effective degrees of freedom* are usually much lower than $\nu = n-2$, so even a relatively large value of r may not be significant. Therefore, assuming $\nu = n-2$ would result in overly confident assessments of the significance of a correlation. Unfortunately, this test comes "out of the box" with most numerical packages,

so many claims of significance made in the literature by unsuspecting users are questionable.

We will see this in excruciating detail in Lab 7, in which we will use non-parametric tests to solve this problem (and obtain more stringent bounds on significance in serially-correlated data). GEOL425L alumn (and Fall 2017 TA) Jun Hu wrote a nice paper on this topic (**?**).

VIII Bayes: return of the reverend

Let us revisit the example of Sect. IV about a resort's advertising claims for cloud-free days[15]. We've just seen the orthodox (frequentist) practice of defining a null hypothesis about some parameters, using it to obtain a null distribution, and then computing the likelihood of observing values of the test statistic as extreme as those we actually obtained. Phew, is it just me or is it an awfully convoluted and perversely counter-intuitive way of going about things? We have to teach this because such tests show up all over the literature, but we wish there was something better. Thankfully there is.

[15] This excellent example is shamelessly reproduced from *Wilks* (2011, pp 198-199)

What would a Bayesian do? A Bayesian would reason in terms of the posterior distribution [Eq. (5.26)]. In this case, our likelihood is similar to Eq. (6.9). For the parameter $\theta = f$, we write

$$\mathbb{P}(X = k|f) = \binom{N}{k} f^k (1-f)^{N-k} \tag{6.17}$$

(probability of observing k could free days in N, given some probability of occurrence f for such days. We want $\mathbb{P}(f|X)$ (the distribution of f given the observations), so the only missing piece is the prior distribution $p(f)$.

We have potentially an infinity of choices. Let's start by being magnanimous and assuming that there is no reason to prefer any outcome, so we may use a flat prior ($p(f) = 1 \, \forall f \in [0,1]$). It turns out that the uniform distribution so described is a special case of the *beta distribution*, introduced in Sect. V and depicted in Fig. 5.4. The flat prior corresponds to the choice $\alpha = 1, \beta = 1$ (Fig. 5.4).

Of course, the beta distribution is a *conjugate prior* to the binomial likelihood. Thus the posterior writes:

$$\mathbb{P}(f|X) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+k)\Gamma(N-k+\beta)} f^{k+\alpha-1}(1-f)^{N-k+\beta-1} \tag{6.18}$$

so it is also a beta distribution, but with parameters $\alpha' = \alpha + k$ and $\beta' = N - k + \beta$. In our case, with $k = 15$ cloud-free days observed over the span of $N = 25$ days, we get $\alpha' = 16$ and $\beta' = 11$.

This distribution is illustrated in Fig. 6.11, where it can be seen that the data allow to lift the prior from a state of complete uncertainty (f could be anywhere in the $[0,1]$ interval), to one centered around $f = 15/25 = 0.6$ (dashed line), but fairly broad; the associated 95% credible interval[16] is [0.406,0.766].

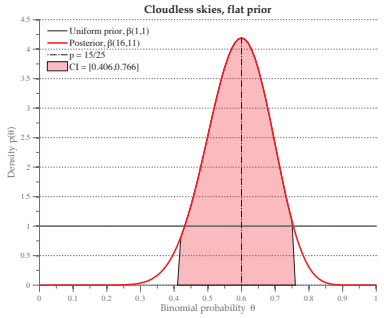


Figure 6.11: Bayesian inference on the probability of cloudless skies with a uniform prior on f. The shaded area depicts a 95% credible interval

Now, someone who is a bit more skeptical about the veracity of advertising claims may assume that there is only a 5% chance that the true probability is above the claimed value of $f_c = 0.857$. In addition, they may assume that values of f above and below 0.5 are equally plausible (symmetry). These two constraints together suggest $\alpha = 4, \beta = 4$ for a more informative prior. The result of using this prior is illustrated in Fig. 6.12, where it can be seen that the mode is pulled slightly to the left compared to the flat prior case (0.581 instead of 0.6), which simply reflects the fact that this prior is more concentrated around 0.5. The associated 95% credible interval, [0.406,0.736], is slightly narrower (notice how only the high values got trimmed), also because extreme values near 1 are weighted down by this prior.

Does it make a difference which prior we choose? In both cases, we see that the claimed probability $f_c = 0.857$ is far outside the 95% credible interval. We can easily compute $\mathbb{P}(f \geq f_c)$ (by numerical integration if necessary, or using the incomplete beta function) and one would see that in either case it is smaller than 5×10^{-4}, so we could reject the claim at that level of confidence. Hence the answer is: no, in this case the prior did not matter to the actual question, because the evidence from the data is so strong that it overwhelms the prior. The choice would be more impactful in less constrained situations.

Finally, let's leverage one advantage of posterior distributions, mentioned in Chap 5, Sect. V: quantifying uncertainties in predicted outcomes. The posterior predictive distribution does just that, and in this case it takes the form of a beta-binomial (aka Pólya) distribution. Assume we want to predict (probabilistically) what will happen over the next $N^+ = 5$ days, using the posterior distribution in Fig. 6.11. There are $N^+ + 1 = 6$ possible outcomes ($k^+ = \{0, \cdots 6\}$), the probability of which is shown in Fig. 6.13 (red curve). It is instructive to compare this distribution to one where we assume that $f = 0.6$ is known exactly. In both cases the most likely outcome is $k^+ = 3$ (since $3/5 = 0.6$), but the beta binomial (red) allocates less probability to the central value, and more to the extremes, in light of not knowing f exactly.

[16] the Bayesian counterpart to confidence intervals, credible intervals quantify the probability of the true parameter being in some range, conditional on the modeling assumptions.

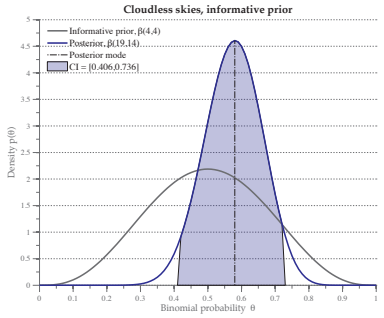


Figure 6.12: The shaded area depicts a 95% credible interval

IX Further considerations

Now, having said all that, the reader is now urged to read a few things:

- Science's Significant Stats Problem, by Tom Siegfried does a great job of explaining why the malpractice or over-insistence on frequentist hypothesis tests to determine "statistical significance" may lead to wildly erroneous scientific claims. We will never say it enough: statistical significance is not practical significance, and it depends crucially on defining a sensible hull hypothesis.
- Scientific method: Statistical errors by Regina Nuzzo very much reaffirms these points, but provides more historical background on how P-values came to pervade the experimental sciences... mostly for the wrong reasons. It is now at the point that many statisticians argue that P-values simply do not provide a useful measure of evidence against a null hypothesis[17]
- *Gelman and Stern* (2006) have a neat article showing that "The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant". We highly recommend it.
- What would a Bayesian do? To compare the merits of two relative hypotheses, Bayes Factors are very useful (*O'Hagan*, 2006). A related idea is the likelihood ratio, which as the name implies only involves the likelihoods, not the priors.
- A classical test is primarily a measure of how much data one has (that is, how much finely one can discriminate between a result originating from chance alone or from a "real" effect)[18]. The choice of significance level is itself very subjective (if something is significant at the 5.1% level but not the 5.0% level, do you throw the baby out with the bathwater?). *What matters is in the end for your research is that you estimate the parameters that interest you and report their uncertainties, estimated as transparently as possible so people can decide whether they should believe you or not*.
- No matter what you hear, there is no such thing as objectivity in human endeavors, so settle for a defensible kind of subjectivity!

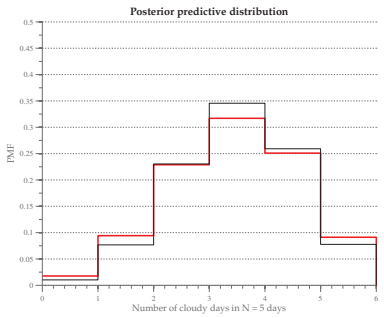


Figure 6.13: Using Bayesian inference to predict future observations. The red curve is the posterior predictive distribution using the posterior in Fig. 6.11. The black curve is the posterior predictive distribution is we knew for sure that $f = 0.6$. One can see that both distributions are very similar, but the red curve is slightly more spread out, in recognition of the fact that f is not known exactly (though its most likely value is also $f = 0.6$).

[17] http://journals.sagepub.com/doi/abs/10.1177/0959354307086923

[18] see *e.g.* Normality tests

Part II

Living in the temporal world

Chapter 7

FOURIER ANALYSIS

Fourier analysis aims to represent a time-ordered signal[1] into periodic (sinusoidal) components – in effect, to decompose sounds into their constituent notes. There are several good reasons for doing so:

1. Periodicity is a source of predictability. For instance, the annual cycle is a type of climate change that far exceeds anthropogenic global warming in many locations, yet human and natural systems have learned to adapt to it because it is so regular.
2. Harmonic signals (of the form $e^{i\omega t}$) are smooth, orthonormal, objective, invariant by integration or differentiation; they are a very convenient basis over which to represent almost any function.
3. Thanks to the Fast Fourier Transform, this decomposition is cheap.

Let us start with clarifying what we mean by timeseries.

[1] That is, a timeseries

I Timeseries

Up to this point we have considered data as an amorphous scramble whose order was unimportant. Obviously, in many instances the order of the data will matter just as much as their values themselves, and timeseries analysis is all about extracting patterns out of this order.

Notion of timeseries

A timeseries $\{t, h(t)\}$ is a collection of ordered observations. h is the dependent variable (the quantity of interest), t the independent variable. While timeseries analysis was ostensibly developed with time as the independent variable, it could just as well be a spatial variable (longitude, latitude, depth). We distinguish:

Continuous timeseries $h(t)$ continuous, analog signal (e.g. old seismometer on paper)

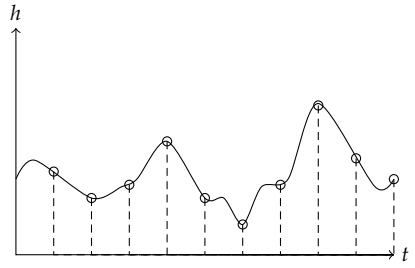


Figure 7.1: An example timeseries $h(t)$, sampled at regular increments.

Discrete timeseries in which case $\{t, h(t)\} = \{(t_1, h_1), \cdots, (t_n, h_n)\}$. Here $h_n = h(n\Delta)$ is the n^{th} **sample**. All digitized signals are discrete, and in our digital world this will be of most interest.

Timeseries analysis

Methods for time series analysis may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and recently wavelet analysis; the latter include auto-correlation and cross-correlation analysis. The bridge between the time and frequency domains is the *Fourier transform*, which can be either discrete or continuous. Methods that use the Fourier transform usually fall under the purview of Fourier analysis, which seeks to *represent all signals as a superposition of pure oscillations*.

If the record is of finite length T, one may easily **periodize** it, so that $h(t)$ becomes *periodic* of period T (cf Fig. 7.2).

How can we represent (decompose) the signal h in terms of simpler components? This amounts to a **projection** of h onto those elemental components.

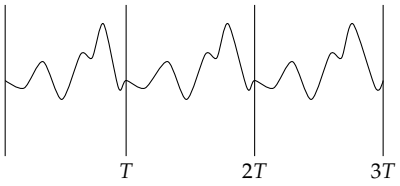


Figure 7.2: Periodization of a finite-length signal

Geometric analogy

In a plane P (orthonormal basis) any vector $\vec{u}$ can be represented as $\vec{u} = u_i\vec{i} + u_j\vec{j}$ or, equivalently, as

$$\vec{u} = \begin{pmatrix} u_i \\ u_j \end{pmatrix}$$

where u_i and u_j are called *components* of the vector $\vec{u}$.

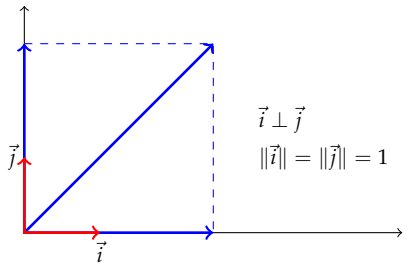


Figure 7.3: Projection

Question *How to find $\vec{u}$'s components (i.e. coordinates)?*

The answer comes from the **inner product** for vectors:

$$\langle \vec{u}, \vec{v} \rangle : P \times P \to \mathbb{R}$$

Recall that:

$$\langle \vec{u}, \vec{v} \rangle = u_i v_i + u_j v_j \quad \left(= \sum_{k=1}^{n} u_k v_k \right) \tag{7.1}$$

As seen in Appendix B, continuous functions on $[0, T]$ form a vector space, and we defined an **inner product for functions**:

$$\langle h, g \rangle = \int_0^T h(t) g(t) dt \tag{7.2}$$

Now,

$$\langle \vec{u}, \vec{i} \rangle = u_i \underbrace{\langle \vec{i}, \vec{i} \rangle}_{1} + v_j$$

so

$$u_i = \langle \vec{u}, \vec{i} \rangle.$$

In a space provided with an orthonormal basis, a vector's components are simply its projections onto the basis vectors (which are given by the **inner product**). One may use this property to decompose a signal onto *basis functions*, as we will do in what follows.

Back to timeseries

In general, one can show that:

$$\mathcal{L}^2_{[0,T]} = \left\{ \text{timeseries } h(t), \text{period } T, \underbrace{\int_0^T |h(t)|^2 dt}_{\mathcal{L}^2 \text{ norm}} < \infty \right\} \tag{7.3}$$

is a **vector space**.

Note *Verify this at home.*

Question *Do we know of any basis?*

Say $T = 2\pi$ without loss of generality. In Appendix B we see that

$$\left\{ \frac{1}{\sqrt{2\pi}} \right\} \cup \left\{ \frac{1}{\sqrt{\pi}} \cos(n\theta) \right\}_{n=1}^{n=\infty} \cup \left\{ \frac{1}{\sqrt{\pi}} \sin(n\theta) \right\}_{n=1}^{n=\infty} \tag{7.4}$$

form a basis of this space. Moreover, sines and cosines are orthonormal:

$$\frac{1}{\pi} \int_0^{2\pi} \cos(n\theta) \sin(n\theta) = 0 \qquad \forall n, m > 0 \tag{7.5a}$$

$$\frac{1}{\pi} \int_0^{2\pi} \cos(n\theta) \cos(m\theta) = \delta_{n,m} \begin{cases} 1 \text{ if } n = m \\ 0 \text{ otherwise} \end{cases} \qquad \text{(same with sines)} \tag{7.5b}$$

Thus, **sines and cosines form an orthonormal basis of** $\mathcal{L}^2_{[0,2\pi]}$. We may thus use them to decompose any timeseries into pure oscillations. This is equivalent to decomposing a sound into its component notes (harmonics), hence the name *harmonic analysis* often used interchangeably with Fourier analysis.

II Fourier Series

Figure 7.4: Joseph Fourier, whose forays into the diffusion of heat led him to invent how to represent any periodic function as a trigonometric series. In the process, he gave birth to mathematical physics.

Joseph Fourier is perhaps best known for his work on trigonometric representations of periodic signals, but this really was a side project of his. He wanted to understand how heat flows through continuous media, and found that heat flow was proportional to the gradient of temperature (i.e. temperature is the potential of heat). He was the first to formulate heat flow as a diffusion problem

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T \tag{7.6}$$

In trying to solve this equation, he was looking for solutions that were proportional to sines and cosines, because he could then easily turn this partial differential equation into an algebraic one. This in turn led to the question: is this legit? Can any periodic function be represented this way? It took a long time for mathematicians to demonstrate this rigorously, but physicists went ahead with it anyway. As usual, their intuition was right. The key was to use an infinity of waves.

Fourier's theorem

Fourier's theorem states that any integrable[2], T-periodic function can be **represented** as an infinite superposition of sines and cosines called a **Fourier series**:

$$h(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\omega t) + b_k \sin(k\omega t) \qquad \omega = \frac{2\pi}{T} \tag{7.7}$$

Since $\cos(k\omega t)$ and $\sin(k\omega t)$ are orthogonal, coefficients a_k and b_k can be obtained by **projecting** on the basis functions as before:

$$a_k = \frac{2}{T}\int_0^T h(t)\cos(k\omega t)dt \qquad k \in \mathbb{N} \tag{7.8}$$

$$b_k = \frac{2}{T}\int_0^T h(t)\sin(k\omega t)dt \qquad k \in \mathbb{N}^* \tag{7.9}$$

Recall that sines and cosines may be viewed as the real and imaginary parts of a complex exponential (Appendix C). That is: $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, so one may also use a **complex representation** of $h(t)$:

$$h(t) = \sum_{n=-\infty}^{+\infty} H_n e^{in\omega t} \tag{7.10}$$

Again, these circular functions are orthogonal[3]:

$$\frac{1}{T}\int_0^T e^{in\omega t} e^{-im\omega t} dt = \delta_{n,m}$$

so the n^{th} complex Fourier coefficient H_n writes:

$$H_n = \frac{1}{T}\int_0^T h(t) e^{-in\omega t} dt \tag{7.11}$$

[2] $\int_0^T |h(t)|dt < \infty$

[3] The complex dot-product of two functions $h(t)$ and $g(t)$ is $\int h(t)g^*(t)dt$, where the asterisk indicates the complex conjugate of g.

It is easy to show that:

$$H_n = \begin{cases} \frac{a_n - ib_n}{2} & n \geq 0 \\ \frac{a_n + ib_n}{2} & n < 0 \end{cases} \tag{7.12}$$

Note *Do this at home, also expressing a_n and b_n as a function of H_n.*

Properties of the Fourier series

- **Hermitian property**: if $h(t)$ is real, then $H_{-n} = H_n^*$;
- **even** functions[4] are only comprised of **cosine** terms;
- **odd** functions[5] are only comprised of **sine** terms;
- Fourier series converge in a least square sense: Defining

  $$S_n(t) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(\omega k t) + b_k \sin(\omega k t) \quad \text{(Fourier expansion of order } n\text{)}$$

  Then $\|S_n(t) - h(t)\|_2 \to 0 \quad \text{as} \quad n \to \infty$. So while **any** periodic function can be represented in this way, it may need many terms if the original signal looks nothing like sines or cosines. The universality of Fourier decomposition may thus come at the expense of efficiency – this is a recurrent tradeoff in science.

[4] $h(-t) = h(t)$

[5] $h(-t) = -h(t)$

Note

- *$h(t)$ must be continuous but not necessarily smooth (Fig. 7.5);*
- *Oscillations near break points may reach high amplitudes (Gibbs phenomenon, lab. 6);*
- *These series have great theoretical use to find solutions of (linear) ordinary and partial differential equations via the principle of superposition.*

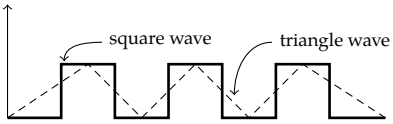


Figure 7.5: Example of non-smooth series

III Fourier transform

The Fourier transform generalizes Fourier series to *any* continuous function. We go from quantized frequencies to a continuum of frequencies.

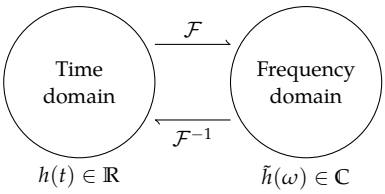


Figure 7.6: Fourier transform

Definition

A **Fourier transform** can be viewed as the continuous limit of a Fourier series. $(h, \tilde{h})$ are called a **Fourier transform pair** and they are given by:

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \tilde{h}(\omega) e^{+i\omega t} d\omega \qquad \omega = 2\pi f \tag{7.13}$$

$$\tilde{h}(\omega) = \int_{-\infty}^{+\infty} h(t) e^{-i\omega t} dt \tag{7.14}$$

Notation: $h \circ\!\!-\!\!- \tilde{h}$ means "$\tilde{h}$ is the Fourier transform of h". Conversely, by $\tilde{h} -\!\!-\!\!\circ h$ we mean "h is the inverse Fourier transform of $\tilde{h}$". Those two statements are equivalent[6]

[6] There are many different definitions of the Fourier transform in the literature: pick one and be consistent. In particular, it is often more convenient to make calculations with an angular frequency (ω) formulation, though the frequency f is more intuitive (since it is the inverse of a period). In this chapter we will mostly deal with ω, but in applications we will mostly think in terms of f. The two are related by $\omega = 2\pi f$.

Examples

Dirac δ-function

- Definition:

$$\delta(t) = \begin{cases} 0 & \forall t \neq 0 \\ \int_{-\infty}^{+\infty} \delta(t) dt = 1 \end{cases} \tag{7.15}$$

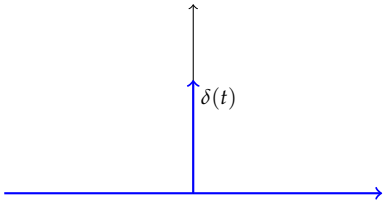


Figure 7.7: The Dirac δ-function describes perfect localization and represents an idealized point mass.

- Sampling property: for any smooth function h

$$\int_{\mathbb{R}} \delta(t) h(t) dt = h(0)$$
$$\int_{\mathbb{R}} \delta(t - t_0) h(t) dt = h(t_0)$$

- Fourier transform:

$$\Delta(\omega) = \int_{-\infty}^{+\infty} \delta(t) e^{-i\omega t} dt = 1$$

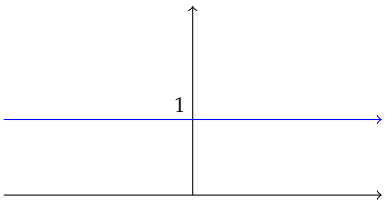


Figure 7.8: The Dirac δ-function has no localization in the frequency domain.

In general, good localization in time is associated with poor localization in frequency. On the opposite end, harmonic functions (sines and cosines) have perfect localization in frequency space, but no localization in time.

Negative Exponential

$$h(t) = H(t)e^{-at}$$

- $H(t)$ = Heaviside jump (Fig. 7.9)

- $$H(t) = \begin{cases} 1 & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{h}(\omega) = \int_{-\infty}^{+\infty} H(t)e^{-at}e^{-i\omega t}dt = \frac{-1}{a+i\omega} \tag{7.16}$$

Whence

$$|\tilde{h}(\omega)|^2 = \frac{1}{a^2+\omega^2} \tag{7.17}$$

Figure 7.9: Heaviside function, akin to flipping a switch on at $t = 0$.

Boxcar Function (aka Gate Function)

$$b(t) = [H(t-\tau) - H(t+\tau)]$$
$$\int_{\mathbb{R}} b(t)dt = \frac{1}{2\tau}$$
$$\tilde{b}(\omega) = \frac{\sin(\omega\tau)}{\omega} = \tau\text{sinc}(\omega\tau)$$

Definition 5 *The **cardinal sine** function* $sinc(x) \equiv \frac{\sin(x)}{x}$ *is characterized by a central peak at the origin and many oscillatory* **side lobes** *that decay hyperbolically away from it. (Fig. 7.10). The function integrates to* π*. It is single-handedly responsible for* leakage *– one of the worst problems to befall the time series analyst (Section V).*

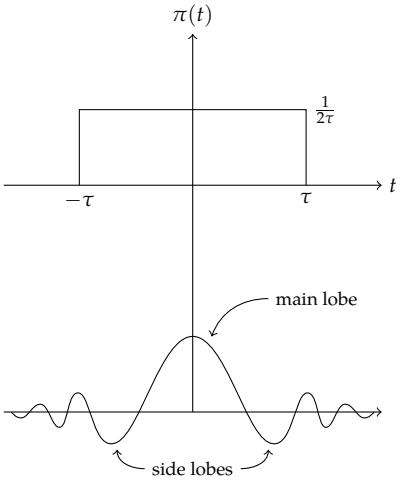


Figure 7.10: The boxcar function and its Fourier transform

Monochromatic waves

Pure vibrations, aka harmonic functions, or monochromatic waves, are a fancy names for sines and cosines.

$$c(t) = \cos(\omega_0 t) \quad \rightarrow \quad \tilde{c}(\omega) = \pi\left[\delta(\omega+\omega_0) + \delta(\omega-\omega_0)\right]$$
$$s(t) = \sin(\omega_0 t) \quad \rightarrow \quad \tilde{s}(\omega) = \frac{\pi}{i}\left[\delta(\omega+\omega_0) - \delta(\omega-\omega_0)\right]$$

Perfect localization in frequency. No localization in time.

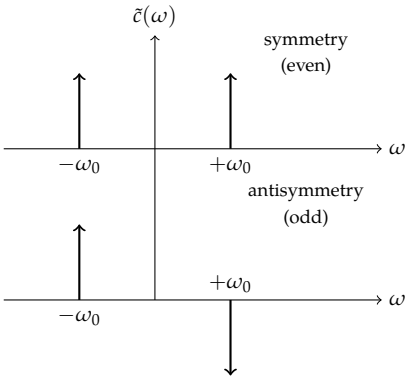


Figure 7.11: Fourier transform of monochromatic waves

Gaussian function

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-t^2}{2\sigma^2}}, \qquad \int_{\mathbb{R}} g(t)dt = 1$$

- $\mathcal{F}$ (Gaussian) = another Gaussian!

$$\tilde{g}(\omega) = e^{-\frac{\omega^2\sigma^2}{2}} \qquad \int_{\mathbb{R}} \tilde{g}(\omega)d\omega \propto \sigma^{-1}$$

with $\sigma' = 1/\sigma$.

- scale in time = (scale)$^{-1}$ in frequency

- $\begin{cases} \text{fatty curve} & \leftrightarrow & \text{slim curve} \\ \text{slim curve} & \leftrightarrow & \text{fatty curve} \end{cases}$

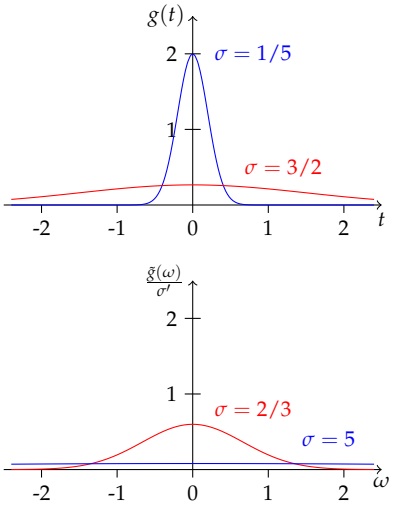


Figure 7.12: The Gaussian transform pair. Note again that localization in the time domain is inversely related to localization in the frequency domain

Properties of the continuous Fourier transform

- **Linearity**: $\mathcal{F}(ah + bg) = a\mathcal{F}(h) + b\mathcal{F}(g)$ A linear system is one where Fourier's principle of superposition works. Linearity $\Leftrightarrow$ Fourier transform.
- **Unicity**: Every $h(t)$ admits a unique $\tilde{h}(\omega)$. Conversely, every $\tilde{h}(\omega)$ admits a unique $h(t)$. A function can thus be *identified* by its Fourier transform.
- **Hermitian property**: $h(t) \in \mathbb{R} \quad \Leftrightarrow \quad \tilde{h}(-\omega) = \tilde{h}(\omega)^*$
- **Parity**:

  $h(t)$ even $\Leftrightarrow$ $\tilde{h}(\omega)$ even

  $h(t)$ odd $\Leftrightarrow$ $\tilde{h}(\omega)$ odd

- **Scaling**: for (a, b) real constants

  $h(at) \circ\!\!-\!\!- \frac{1}{|a|}\tilde{h}\left(\frac{\omega}{a}\right) \quad \rightarrow$ time scaling

  $\frac{1}{|b|}h\left(\frac{t}{b}\right) \circ\!\!-\!\!- \tilde{h}(b\omega) \quad \rightarrow$ frequency scaling

- **Shifting**:

  $h(t - t_0) \circ\!\!-\!\!- e^{i\omega t_0}\tilde{h}(\omega) \quad \rightarrow$ time shifting

  $h(t)e^{-i\omega_0 t} \circ\!\!-\!\!- \tilde{h}(\omega - \omega_0) \quad \rightarrow$ frequency shifting

- **Derivation**:

  $\frac{\partial h}{\partial t} \circ\!\!-\!\!- i\omega\tilde{h}(\omega)$

- **Integration**:

$$\int h dt \circ\!\!-\!\!- \frac{-i}{\omega}\tilde{h}(\omega)$$

This means that differential equations can be solved as polynomials in frequency space (very useful since there is such a large apparatus to solve polynomial equations).

- **zero frequency**

$$\int_{-\infty}^{+\infty} h(t)dt = \tilde{h}(0)$$

corresponds to the average value of h[7].

[7] the "DC" in AC/DC

Important theorems

Parseval's theorem

The energy of a signal $h(t)$:

$$E \equiv \int_{-\infty}^{+\infty} |h(t)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} |\tilde{h}(\omega)|^2 d\omega \tag{7.18}$$

may be measured in the time or frequency domain. This leads to the definition of a fundamental quantity:

Definition 6 *The **Power Spectrum** or **Spectral Density** is defined as:*

$$\frac{1}{2T}\int_{-T}^{+T} |h(t)|^2 dt \qquad PSD/unit\ time \tag{7.19a}$$

$$S(\omega) \equiv |\tilde{h}(\omega)|^2 \tag{7.19b}$$

It is an energy density per frequency band, and quantifies the partition of variance (energy) among frequency components (periodicities). It tells us fundamental things about the behavior of a physical system from which the time-series originated. For instance, how much temperature variance is accounted for by the annual cycle? How is energy transferred between different scales of motion? What processes generated the measurement? What is the energy of the low-frequency seismic waves ($T > 4$ min)? Are there scale invariances in the system? Is the system particularly sensitive to forcing at a given frequency?). Obtaining reliable estimates of the spectral density is a holy grail of **spectral analysis** *(Chapter 9).*

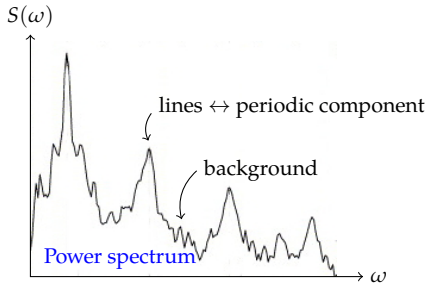


Figure 7.13: Spectral analysis and harmonic analysis.

Convolution theorem

Definition 7 *The convolution of two functions h, g is:*

$$h * g = g * h = \int_{\mathbb{R}} h(\tau)g(t-\tau)d\tau \tag{7.20}$$

This product verifies many of the properties of the usual product: commutativity, associativity. Its neutral element is the Dirac δ-function. The convolution represents the action of one linear system over another (cf Filters, Chapter 10)

Example
A recorded seismological signal $r(t) = s(t) * b(t) * g(t) = (s * b * g)(t)$ where s is the seismometer, b is the building motion, and g is ground motion. Convolution formalizes the idea of the composition of various (linear) systems influencing each other.

An essential property of linear systems is that

$$h * g \circ\!\!-\!\!- \tilde{h}(\omega) \times \tilde{g}(\omega) \tag{7.21}$$

- convolution in time $\leftrightarrow$ multiplication in the frequency domain.
- multiplication in time $\leftrightarrow$ convolution in the frequency domain.

This theorem is absolutely fundamental and is a key reason why the Fourier transform is so used. In practice, convolution is almost always done in the frequency domain.

Correlation theorem

$$C(h,g) = \int_{\mathbb{R}} h(t+\tau)g(t)dt \qquad \tau = \text{“lag”} \tag{7.22}$$

$$C(g,h) = \int_{\mathbb{R}} h(t)g(t+\tau)dt \tag{7.23}$$

$$C(h,g) \circ\!\!-\!\!- \tilde{h}(\omega)\tilde{g}(-\omega) = \tilde{h}(\omega)\tilde{g}(\omega)^* \text{ for } h,g \text{ real} \tag{7.24}$$

Definition 8 *The **autocorrelation** of a function h is the function a:*

$$a(\tau) = C(h,h) = \int_{\mathbb{R}} h(t)h(t+\tau)dt \tag{7.25}$$

is a measure of its memory (persistence), indicating how much a function resembles itself at lag τ.

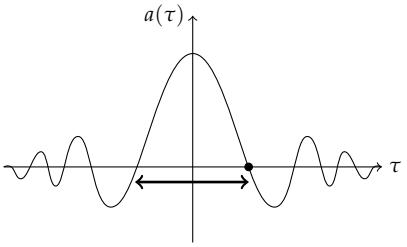


Figure 7.14: One way to define decorrelation time, is the time it takes for the autocorrelation function to first cross zero. Alternatively, one may use the e-folding time associated with the envelope of this curve. In both cases it is a measure of the memory, or persistence, of a timeseries

Note - *a is even ($a(\tau) = a(-\tau)$) when h is real.*

- *the autocorrelation is the cross-correlation of a function with itself.*

Wiener-Khinchin theorem

The Fourier transform of the autocorrelation is the squared amplitude (**power spectral density**) of $\mathcal{F}(h)$.

$$a(t) \circ\!\!-\!\!- \tilde{h}(\omega)\tilde{h}^*(\omega) = |\tilde{h}(\omega)|^2 = S(\omega) \tag{7.26}$$

This is the foundation for classical (Blackman-Tukey) spectral analysis, but we will shortly see that much better methods exist.

IV Discrete Fourier transform

Definitions

In practice, all measured signals are discrete if processed on a computer:

$$h(t) \to h_k = h(k\Delta) \quad k \in 0, 1, \dots N-1, \quad N \text{ even}$$

where Δ is the *sampling interval*. Ideally, N spans the entire domain of h.

Definition 9 *The Discrete Fourier Transform of h is:*

$$H_n = \sum_{k=0}^{N-1} h_k W_N^{kn} \tag{7.27}$$

where $n \in \{0, N-1\}$ and W_N is an N^{th} root of unity. (Appendix C, section II). We write $h_k \overset{N}{\longmapsto} H_n$.

Note the relationship to the CFT:

$$\tilde{h}(f_n) = \int_{-\infty}^{+\infty} h(t) e^{-2\pi i f_n t} dt \simeq \sum_{h=0}^{N-1} h_k e^{-2\pi i f_n t_k} \Delta \simeq \Delta \sum_{k=0}^{N-1} h_k e^{-\frac{i2\pi}{N} kn} = \Delta H_n \tag{7.28}$$

There is however a fundamental difference: the DFT is limited to a finite range of frequencies; in particular:

Definition 10 (Nyquist frequency) *In the frequency domain $f_n = \frac{n}{N\Delta}$, $n = -\frac{N}{2}, \dots 0$, the following expression*

$$f_{\pm\frac{N}{2}} = \pm \frac{1}{2\Delta} \tag{7.29}$$

*is called the **Nyquist frequency** and is the highest frequency resolvable by the dataset. If the dataset has energy above this frequency, it will be* aliased *to lower frequencies (Sect. V).*

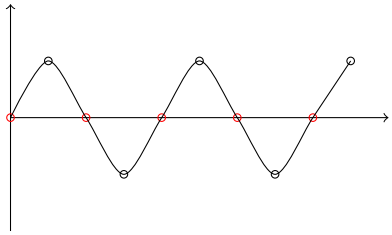


Figure 7.15: Sampling at the Nyquist frequency, where a digitized sinusoid would take values of $(-1, 0, +1)$. It should be clear that sampling at half that rate would result in mistaking the sinusoid for a constant (red dots).

Periodicity

$$H_{-n} = H_{N-n} \qquad n \in [1, N-1] \quad \to \quad N \text{ periodic} \tag{7.30}$$

Ordering of frequencies

$$\begin{aligned} 0 < f < f_{Ny} \quad &\Leftrightarrow \quad 1 \le n \le \tfrac{N}{2} - 1 \\ -f_{Ny} < f < 0 \quad &\Leftrightarrow \quad \tfrac{N}{2} + 1 \le n \le N - 1 \\ f = \pm f_{Ny} \quad &\Leftrightarrow \quad n = \tfrac{N}{2} \end{aligned}$$

Invertibility

Can we retrieve h_k from H_n, and vice versa? Yes, via the inverse DFT:

Definition 11 *Inverse Discrete Fourier Transform:*

$$h_k = \frac{1}{N}\sum_{n=0}^{N-1} H_n W_N^{-kn} \tag{7.31}$$

We write $H_n \overset{N}{\longmapsto} h_k$

$$\text{iDFT}\left(\text{DFT}(h)\right) = \sum_{n=0}^{N-1}\frac{1}{N}\sum_{l=0}^{N-1} h_l e^{-\frac{2\pi}{N}iln} e^{\frac{2\pi ikn}{n}} = \sum_{l=0}^{N-1}\frac{h_l}{N}\sum_{n=0}^{N-1} e^{-\frac{i2\pi n}{N}(l-k)} \tag{7.32}$$

In the last term, we recognize the sum of the N^{th} roots of unity:

$$\sum_{n=0}^{N-1} W_N^{n(l-k)} = \frac{1-(W_N^{l-k})^N}{1-W_N^{l-k}} = \begin{cases} 0 & l \neq k \\ N & l = k \end{cases}$$

(See Appendix C).

Put differently,

$$\frac{1}{N}\sum_{n=0}^{N-1} W_N^{n(k-l)} = \delta_{kl} \qquad \delta = \text{Kronecker Delta} \tag{7.33}$$

So

$$h_k = \sum_{n=0}^{N-1}\frac{1}{N} H_n e^{+\frac{2\pi ik}{N}n} = \sum_{l=0}^{N-1} h_l \delta_{kl} = h_k \tag{7.34}$$

The two operations really are reciprocal (we start with h_k and we get it back without any loss). The converse is also true, applying the inverse DFT to H_n, then the DFT, gives you H_n back (try it at home). This is fundamental, since it means that we can go from time domain to frequency domain *without any loss of information*. So even though the range of frequencies is limited, this transformation defines a one-to-one mapping between (discrete) time and frequency domains.

Properties

- **Parseval's theorem (Discrete form)**

$$\sum_{h=0}^{N-1} |h_k|^2 = \frac{1}{N}\sum_{n=0}^{N-1} |H_n|^2 \tag{7.35}$$

- **Discrete Convolution** For two sequences (r_k, s_k) such that $r_k \overset{N}{\longmapsto} R_n$ and $s_k \overset{N}{\longmapsto} S_n$, their discrete convolution is the product of their DFTs.

$$(r * s)_j \equiv \sum_{k=-\frac{N}{2}+1}^{\frac{N}{2}} s_{j-k} r_k \overset{N}{\longmapsto} S_n R_n \tag{7.36}$$

Note *Care is required with the edges* $\rightarrow$ *see lab 6.*

To which we could add the discrete form of the cross-correlation and autocorrelation functions.

Fast Fourier Transform

Cooley and Tukey developed in 1968 an algorithm to compute the Discrete Fourier Transform very efficiently.

Matrix formulation

The key is to recognize the DFT as a matrix multiplication, and the inverse DFT as a matrix inversion problem (Appendix B).

Definition 12 *Fourier matrix*

$$\mathbf{H}_n = \mathbf{F}\mathbf{h}_k \qquad \textit{with } \mathbf{F}_{ij} = W_N^{(i-1)(j-1)},\ \mathbf{h}_k = \frac{1}{N}\mathbf{F}^*\mathbf{H}_n$$

Because of the wonderful properties of the roots of unity, F is an orthonormal (circular) matrix, so that its inverse is it conjugate: $F^{-1} = F^*$. This makes the inverse DFT trivially simple to compute. Further, when N is a power of two, huge computational gains can be achieved by exploiting the properties of symmetry between the roots and the transformed values. Padding to the next power of two (*e.g.* $1000 \rightarrow 1024$) is therefore extremely common.

Numerical cost

The FFT makes many costly operations cheap in the Fourier domain :

Operation	Formulation	Direct cost	FFT cost
DFT	$\sum_{k=0}^{N-1} h_k W_N^{kn}$	N^2	$2N \log_2 N$
Convolution	$\sum_{n=0}^{N-1} h_k g_{n-k}$	N^2	$3N \log_2 N$
Correlation	$\sum_{n=0}^{N-1} h_k g_{n+k}$	N^2	$3N \log_2 N$

This doesn't look like much, but as N becomes large, the FFT cost becomes negligible in front of the direct cost. Hence, in practice convolutions are always done in the Fourier domain: $(s * r)_k \xrightarrow[N]{} (S_n \times R_n)$.

V The Curse of Discretization

Shannon's sampling theorem

If a signal $h(t)$ is band-limited[8], and is sampled at intervals Δ fine enough that $\Delta \leq \frac{1}{2f_c}$, then $h(t)$ can be *completely recovered* by the convolution:

$$h(t) = \sum_{k=0}^{N-1} h_k \,\mathrm{sinc}\left(\frac{t - k\Delta}{\Delta}\right) \tag{7.37}$$

where sinc is the *cardinal sine* function Δ is chosen so that $\Delta = \frac{1}{2f_c} = \frac{T_c}{2} \quad \rightarrow \quad 2\frac{\text{samples}}{\text{period}}$.

You heard it right: not just approximately recovered, but *completely* recovered. If a signal is band-limited, it only contains a countable amount of information, and therefore can be entirely summarized with a (discrete) sequence of numbers. The applications of this principle are all around us, in audio engineering, video& sound processing, telecommunications, cryptography, etc.

Alternatively, this theorem states that the highest recoverable frequency is the **Nyquist frequency** $f_N = \frac{1}{2\Delta}$. If $h(t)$ is *not* band-limited, power outside $[-f_N, f_N]$ will be *aliased* (falsely translated) into this interval (see below).

[8] $\tilde{h}(f) = 0$ outside of some interval $[-f_c, f_c]$, where f_c is called the *cutoff frequency*

Fourier sampling theory

The Discrete Fourier Transform allows to move freely between time and frequency domains, but there are three important differences with the Continuous Fourier Transform:

Aliasing: $\Delta > 0 \quad \Rightarrow$ limited spectral resolution $(f_N = \frac{1}{2\Delta})$
$\rightarrow$ only frequencies up to f_N are faithfully captured.
$\rightarrow \quad |f| > f_N \quad \Rightarrow$ frequencies higher than f_N parade as low frequencies (bad, Fig. 7.18)

Cyclicity: the cyclicity of the roots of unity leads to $H_{n+N} = H_n$ – the spectrum repeats itself indefinitely; more precisely, it gets folded over the Nyquist frequency.

Leakage: Finite time window $\Rightarrow$ *Leakage* of energy from lines to broader bands (bad)

Both leakage and aliasing are linked to the {boxcar $\leftrightarrow$ sinc} transform pair. Let us analyze these plagues in detail.

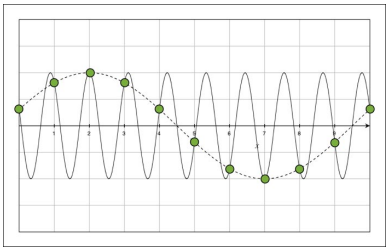
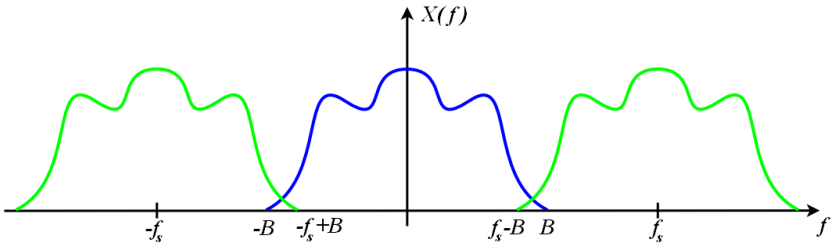
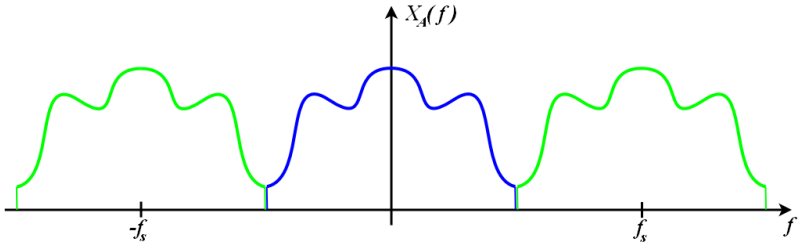


Figure 7.16: When sampling at coarse intervals, higher-frequency harmonics are erroneously projected on low-frequency harmonics. This is the essence of aliasing. Credit: xyo balancer blog

Aliasing

The discretization is akin to **folding** the time spectrum at integer multiples of f_N.





Unresolved frequencies ($|f| > f_N$) come back to haunt the time spectrum like a, well, specter (grey area in Fig. 7.18). These high-frequency components thus parade as low-frequency components. This leads to a distortion of the spectrum, and may lead to misinterpretation of the variability. For a stark example of this, see *Wunsch* (2000).

Solution (instrument) = band-limit the signal by analog filtering prior to digitization (e.g. seismometer). Unfortunately, this is not always possible, e.g. if your "instrument" is a paleo record.

Leakage

Leakage arises because sampling over a finite time window amounts to multiplication by a gate function (*e.g.* Fig. 7.19, between the red lines):

$$B_T(t) = \begin{cases} 1 & \forall |t| \leq \frac{T}{2} \\ 0 & \text{elsewhere} \end{cases}$$

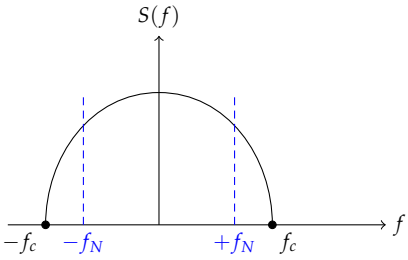


Figure 7.17: An example of true spectrum, aliased by sampling at too low a frequency.

Figure 7.18: Consequences of the spectral folding inherent to the DFT. The blue curve is the spectrum of a band-limited signal in $[-B, B]$, the green curves its replica about the sampling frequency ($f_s = 1/\Delta$). If f_s is lower than B, the folding results in high frequency signals being mislabeled as lower-frequency components, thus distorting the spectrum (lower panel). Credit: Wikipedia

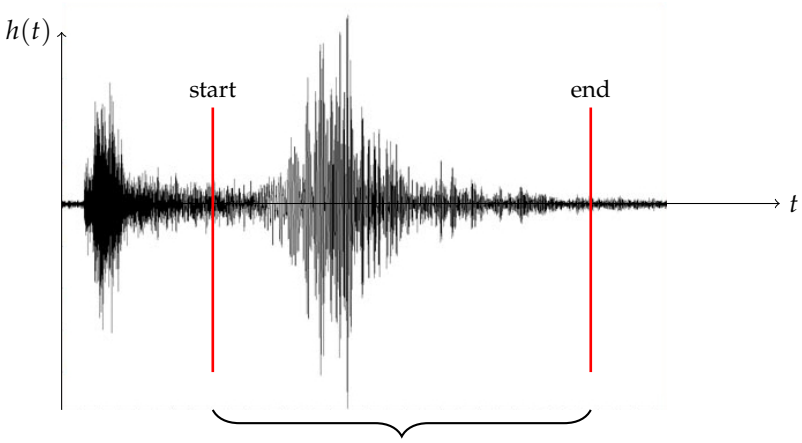


Figure 7.19: An example seismogram, sampled over a finite time. Credit: http://thegeosphere.pbworks.com/w/page/24663884/Sumatra

Owing to the convolution theorem (Eq. (7.21)), multiplication by a gate function in the time domain is tantamount to convolution by its Fourier transform in the frequency domain. This function is $\tilde{B}_T(f) = T\text{sinc}(\pi T f)$ (Fig. 7.20, bottom right), far from being the perfect delta function we wish for (Fig. 7.20, top right). Thus energy leaks from the narrow central peak to a blurry, broad peak. Adding insult to injury, the side lobes spread this energy even further to the sides in an oscillatory fashion, where it may pollute other harmonics. As $T \to \infty$, this effect goes to zero[9], so one should always strive for long records.

[9] the cardinal sine function converges to a delta function in that limit, meaning that sampling in the frequency domain becomes perfectly localized in the frequency domain

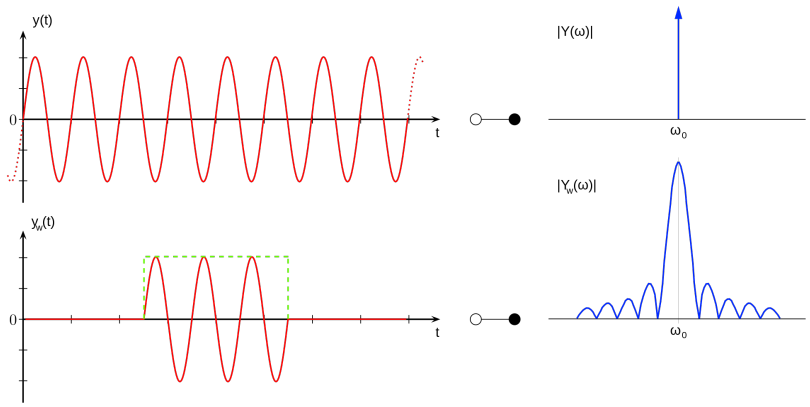


Figure 7.20: Spectral leakage for a perfect sinusoid. (top) ideal sampling situation, when T is large; (bottom), T spans only two periods of the oscillation, resulting in convolution by a rather broad cardinal sine in the frequency domain. The resulting spectrum (lower right) is very much distorted.

In summary, both aliasing and leakage are consequences of the finiteness of N.

finite length $T = N\Delta < \infty \quad \Rightarrow \quad$ **leakage**

finite resolution $\Delta > 0$ (+ spectrum cyclicity) $\quad \Rightarrow \quad$ **aliasing**

Because they are bound by the relation $N = \frac{T}{\Delta}$, one cannot increase both T and decrease Δ for a constant N: this tradeoff is the **curse of discretization**.

Spectral range

In practice, a spectrum can only be estimated on the interval $[f_R, f_N]$

$$f_R = \text{Rayleigh frequency} = \frac{1}{N\Delta}$$

(aka gravest tone, lowest note, fundamental harmonic).
Since $f_n = \frac{n}{N\Delta} = nf_R$, f_R is also the spacing between frequency points, *i.e.* the spectral resolution: $\Delta f = f_R$. All other frequencies are integer multiples of f_R.

$$f_N = \text{Nyquist frequency} = \frac{1}{2\Delta} = \frac{N}{2} f_R$$

(highest note that can be recorded without aliasing the signal). If sampling is chosen so that the signal has no energy beyond the Nyquist frequency, then all is good. Otherwise, aliasing is always present, and impossible to remove. This is one reason some scientific results may change when more high-resolution data have been collected.

Bottom line: if N is fixed, Δ can be increased at the expense of T, but there is no free lunch. Hunting for periodic signals in discrete observations thus requires to deal with these fundamental constraints. This is the object of spectral analysis (Chapter 9)

Chapter 8

TIMESERIES MODELING

Classic statistical tests assume IID data. When such conditions are met, life is beautiful and those tests are useful. In the geosciences, this is rarely the case. We thus wish to provide geophysically-relevant null hypotheses so that we can correctly judge the signficance of spectral peaks, or of correlations between timeseries. Additionally, we will need such models to estimate features of a timeseries believed to follow such models (cf the Maximum Entropy spectral method, Chap. 9).

I The AR(1) model and persistence

Perhaps the overarching quality of geophysical timeseries is their persistence: very often, neighboring values tend to be highly correlated. This may be because low-frequency dynamics are at play (*cf.* Fig. 8.1), or because the way we are measuring the signal introduces this persistence. For instance, it is well known that watersheds tend to integrate climate fluctuations, so that watershed measurements (e.g. streamflow) exhibit more persistent behavior than the input climate (*Hurst*, 1951). Paleoclimate examples abound. Persistence has many consequences, which will shall explore now.

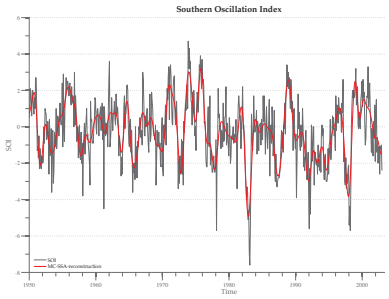


Figure 8.1: The Southern Oscillation (SOI) times series (black), whose low-frequency behavior is highlighted by singular spectrum analysis (red)

Stationarity

In all the following, we consider stationary stochastic processes. A *stochastic process* is a sequence of random variables indexed by time; that is, we consider each temporal observation as the realization of a random variable. A *strictly stationary* process is one whose joint probability distribution ($\mathbb{P}(X_{t_0}, X_{t_1}, \cdots, X_{t_n})$) does not change when shifted in time ($t \rightarrow t + L$). Accordingly, all its moments (e.g. mean, variance) are constant in time. Here, we only need a weaker form, called *wide-sense stationarity*, which states that the autocorrelation depends only on the relative lag separating two events – not their absolute time. That is:

$$\forall L \in \mathbb{R}, \gamma(\tau) = \mathrm{Cov}(X_t, X_{t+\tau}) = \mathrm{Cov}(X_{t+L}, X_{t+L+\tau}) \qquad (8.1)$$

The AR(1) model

The simplest and most popular way to represent persistence is via the autoregressive model of order one, aka AR(1):

$$X_t = \phi X_{t-1} + \varepsilon_t \qquad \phi = \rho_1 = \text{“Lag-1 autocorrelation”} \tag{8.2}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. It is common to consider a normally distributed series $X \sim \mathcal{N}(0, \sigma^2)$. The coefficient ϕ measures the *memory* of the system, i.e. the degree of **persistence** from one value to the next:

- $\phi = 0 \quad \rightarrow \quad$ independent, Gaussian i.i.d. process.
- The higher the ϕ, the smoother the timeseries (Fig. 8.6).
- for $\phi > 0$, $\text{Var}(X_t) = (1-\phi^2)\sigma_\varepsilon^2 < \sigma_\varepsilon^2 \quad \rightarrow \quad$ the variance of the noise is reduced by knowledge of past observations.
- ϕ may be estimated by regressing X_t onto X_{t-1} (Fig. 8.2), hence the name *autoregression*.$\rightarrow \quad X_t$ tends to resemble X_{t-1}, with a scatter governed by σ_ε^2.
- the autocorrelation function (ACF) of an AR(1) model is simply $\rho(k) = \phi^k$ for each lag k. We distinguish two cases, $\phi > 0$ and $\phi < 0$.

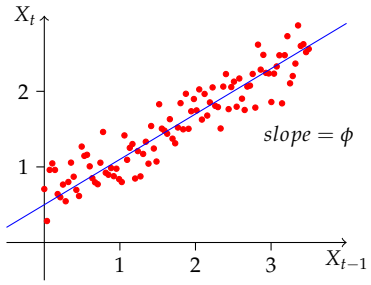


Figure 8.2: Regression of timeseries values over their predecessors. This is one way to estimate ϕ.

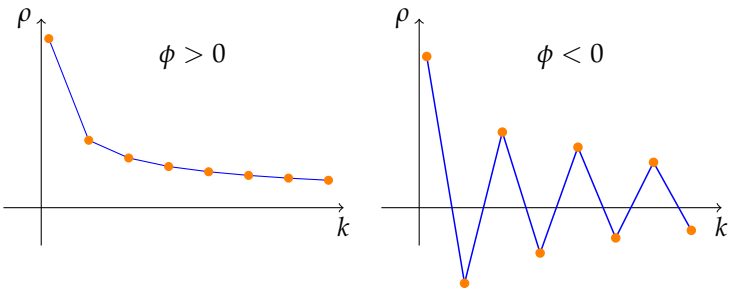


Figure 8.3: Autocorrelation function of an AR(1) series, for positive and negative values of ϕ. k is the lag, $\rho(k)$ the autocorrelation at this lag.

Theoretical Spectrum

The theoretical spectrum of an AR(1) process is as follows.

$$S(f) = \frac{\sigma^2}{1 - 2\phi\cos\left(2\pi\frac{f}{f_N}\right) + \phi^2} \tag{8.3}$$

which is represented in Fig. 8.4. However, as *Ghil et al.* (2002) point out, “a single realization of a noise process can [...] have a spectrum that differs greatly from the theoretical one, even when the number of data points is large. It is only the (suitably weighted) average of such sample spectra over many realizations that will tend to the theoretical spectrum of the ideal noise process. Indeed, the Fourier transform of a

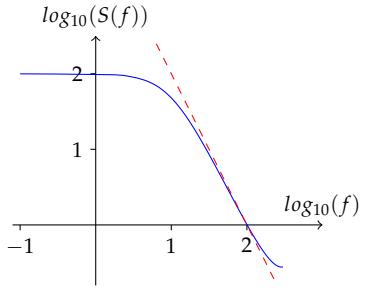


Figure 8.4: Spectrum of an $AR(1)$ model with $\sigma = 1$, $f_N = 10$ and $\phi = 0.9$. Notice the slope (-2) at the inflection point. Generally, any process obeying $S(f) \propto f^{-2}$ is called "red noise" (Sect. III)

single realization of a red noise process can yield arbitrarily high peaks at arbitrarily low frequencies; such peaks could be attributed, quite erroneously, to periodic components." This is illustrated in Fig. 8.6, where it can be seen that many spectra exhibit peaks above the spectrum's theoretical value. We will see in lab 8 how to determine the significance of such peaks against a AR(1) nulls.

The effects of persistence

Since X_t depends on ϕX_{t-1}, then X_t depends on $\phi^2 X_{t-2}$, ..., and so on until $\phi^k X_{t-k}$. The influence of old values decays exponentially, but there's an *infinite memory* of the system (the first value is never completely forgotten). This means the X_t's are *no longer independent* – we say that they are *serially dependent*. Hence, all tests requiring i.i.d. data (e.g. Correlation between two timeseries) will fail on such data. For a t-test, the effective number of degrees of freedom may be estimated as:

$$N_{\text{eff}} \simeq N \left(\frac{1-\phi}{1+\phi} \right) \qquad \text{effective sample size} \tag{8.4}$$

For $\phi = 0.8$, which is far from unusual, N_{eff} is smaller than N by nearly an order of magnitude! One can then plug this d.o.f into a t-test with $T = r\sqrt{\frac{N_{\text{eff}}-2}{1-r^2}} \sim t_{N_{\text{eff}}-2}$, and one often finds very, very different (much less significant) results that if one went along willy-nilly with the naïve assumption that $N_{\text{eff}} = N$ (Fig. 8.5). Alternatively, one may use non-parametric tests based on simulated timeseries (Lab 8).
Persistence means that every spectrum will look "red", that is, more energetic at low than high frequencies (*e.g.* Fig. 8.4).

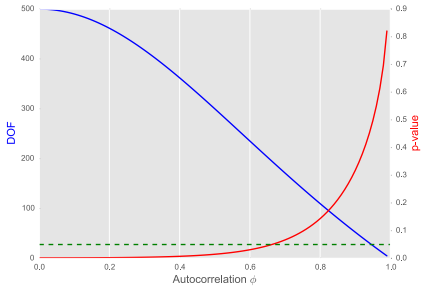


Figure 8.5: The p-value and numbers of degrees of freedom (DOF) for a t test of a relatively low correlation (0.13) between two AR(1) time series (500 samples each) with autocorrelation ϕ. The green dashed line is the 5% threshold for significance. From *Hu et al.* (2017)

II Linear Parametric Models

AR(1) models are part of a general class of timeseries models called linear parametric models, comprising autoregressive models, moving average models, and their union, ARMA models.

Autoregressive Models AR(p)

A random process X is said to follow an autoregressive model of order p if:

$$X_t - \mu = \sum_{i=1}^{p} \phi_i (X_{t-i} - \mu) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad \phi_i \in \mathbb{R}. \tag{8.5}$$

where $\mu = E(X)$. X_t depends only on the last p observations, plus an innovation term ε_t. This means that X_t is conditionally independent of all past observations prior to X_{t-p}, given $\{X_{t-1}, X_{t-2}, \ldots X_{t-p}\}$.

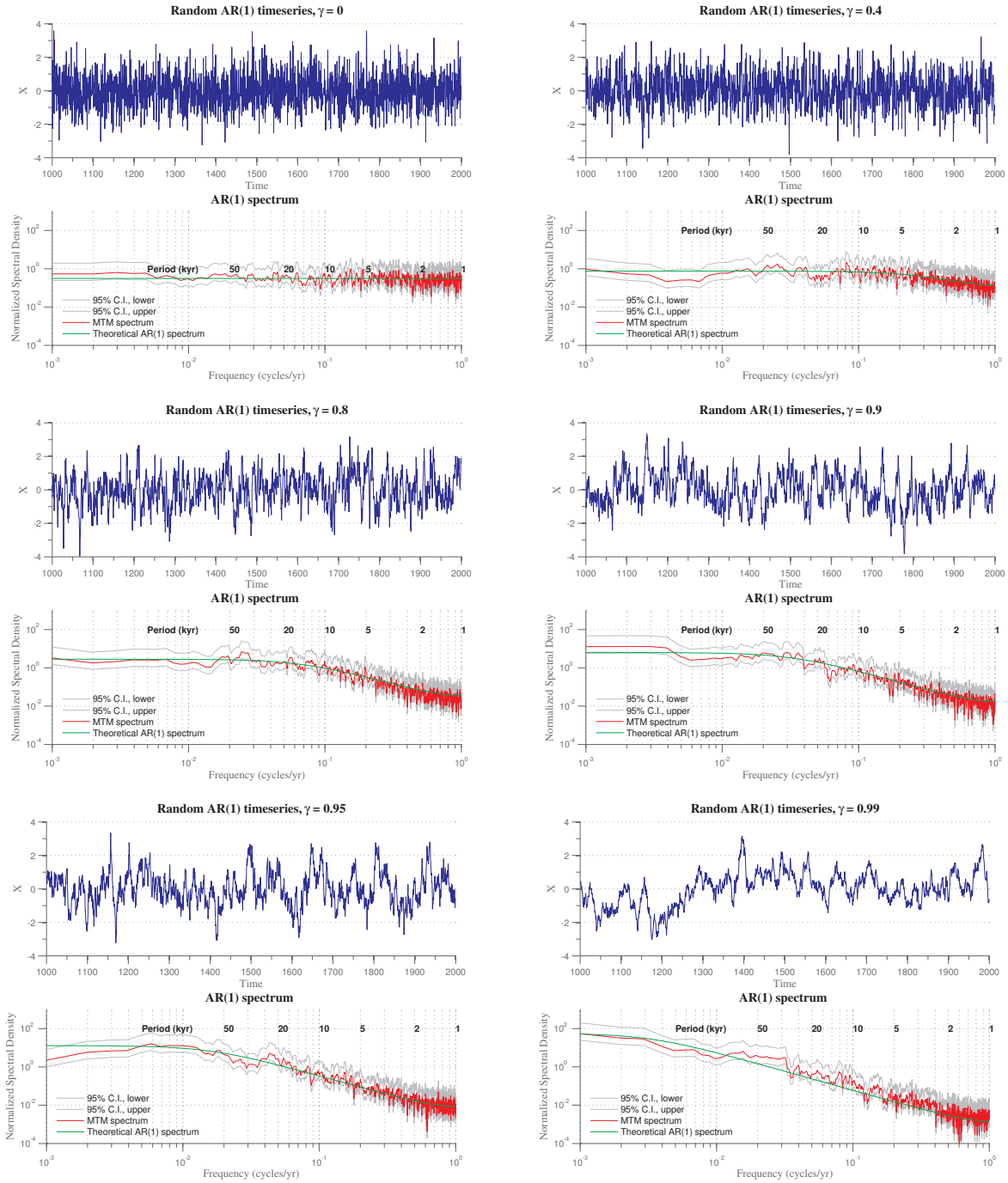


Figure 8.6: The spectral effects of persistence, illustrated for 8 values of the lag-1 autocorrelation parameter, here denoted by γ.

Example
AR(1) model:

$$\begin{aligned} X_t - \mu &= \phi(X_{t-1} - \mu) + \varepsilon_t \\ &= \phi(\phi(X_{t-2} - \mu) + \varepsilon_{t-1}) + \varepsilon_t \\ &\vdots \\ X_t &= \sum_{k=1}^{\infty} \phi^k \varepsilon_{t-k} + \mu \end{aligned}$$

Although X_t depends on *all* past values of the noise, it is *conditionally independent* of them, given X_{t-1}; values prior to $t-1$ do not *add* any information. This is also known as a *Markov process*.

Moving Average Models MA(q)

$$X_t = \mu + \sum_{i=0}^{q} \theta_i \varepsilon_{t-i} \qquad \theta_0 = 1,\ \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{8.6}$$

Amounts to a discrete convolution with Gaussian noise, with filter coefficients θ_i. This is also known as a finite impulse response filter (FIR).

ARMA models

In general, a broad class of timeseries can be approximated by an ARMA(p, q) model, which fuses AR and MA models:

$$X_t - \mu = \sum_{i=1}^{p} \phi_i (X_{t-i} - \mu) + \sum_{i=0}^{q} \theta_i \varepsilon_{t-i} \tag{8.7}$$

Fitting a timeseries model

Autoregressive model of order p Assume without loss of generality that $\mu = 0$. Then:

$$X_t = \varepsilon_t + \sum_{k=0}^{K} \alpha_k X_{t-k}$$

$$X_t \to \underbrace{E(X_t X_t)}_{\gamma(0)=1} = E(\varepsilon_t X_t) + \sum_{k=0}^{K} \alpha_k E\underbrace{(X_t X_{t-k})}_{\gamma_k} = 0 + \sum_{k=0}^{K} \alpha_k \gamma_k$$

$$E(\underbrace{X_t X_{t-1}}_{\gamma(1)}) = E(\varepsilon_t X_{t-1}) + \sum_{k=0}^{K} \alpha_k E(\underbrace{X_{t-1}, X_{t-k}}_{\gamma(k-1)}) \qquad \text{if stationary}$$

$$\gamma(1) = \sum_{k=0}^{K} \alpha_k \gamma(k-1)$$

$$\gamma(n) = \sum_{k=0}^{K} \alpha_k \gamma(k-n) \quad \to \quad \text{discrete convolution}$$

Moving average of order q Since $X_t = \sum_{k=0}^{K} \beta_k \varepsilon_{t-k}$,

$$X_t = \alpha X_{t-1} + \varepsilon_t = \alpha^n X_{t-n} + \sum_{k=0}^{\infty} \alpha^k \varepsilon_{t-k} \qquad \begin{array}{l}\text{where } X_{t-1} = \alpha X_{t-2} + \varepsilon_{t-1},\\ Xd_{t-2} = \alpha X_{t-3} + \varepsilon_{t-2}, \text{ etc.}\end{array}$$

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

$$E(\cdot \times X_t) \to \underbrace{E(X_t X_t)}_{\sigma^2 = \gamma(0)} = \sum_{i=1}^{p} \phi_i \underbrace{E(X_t X_{t-i})}_{\text{lag } i \text{ autocorrelation } (\gamma(i))} + E(\varepsilon_t X_t)$$

$$E(\cdot \times X_{t-1}) \to \underbrace{E(x_{t-1} X_t)}_{\gamma(1)} = \sum_{i=1}^{p} \phi_i \underbrace{E(X_{t-1} X_{t-i})}_{\gamma_{t-i}} + E(\varepsilon_t X_{t-1}) \qquad \begin{array}{l}\text{using } \textit{wide-sense stationarity}\\ (\gamma \text{ is function of lag only})\end{array}$$

$$E(\cdot \times X_{t-k}) \to E(X_{t-k} X_t) = \sum_{i=1}^{p} \phi_i \gamma(i-k) = \gamma(k) \qquad \text{define } \rho_i = \gamma(i-1)$$

Linear system of equations (Yule-Walker equations)

$$\begin{cases} \rho_1 = \gamma_0 = \phi_1 + \phi_2 \rho_1 + \phi_3 \rho_2 + \ldots + \rho_{k-1} \phi_k \\ \rho_2 = \phi_1 \rho_1 + \phi_2 + \phi_3 \rho_1 + \ldots + \rho_{k-2} \phi_k \\ \vdots \\ \rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \ldots + \phi_k \end{cases} \quad \to \quad \boldsymbol{\rho} = \mathbf{M}\boldsymbol{\rho} + \boldsymbol{\phi}\mathbf{1} \quad \to \quad \begin{array}{l}\exists \text{ strong}\\ \text{relationship}\\ \text{between}\\ \text{coefficients}\end{array}$$

$$\text{where} \quad \mathbf{M} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 \ldots \phi_k & \\ \vdots & \vdots & \ddots & \vdots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

$$\text{Hence} \quad \rho_m = \sum_{k=1}^{K} \phi_k \rho_{m-k}$$

This is solved recursively starting from $\rho_0 \equiv 1$. In practice, one estimates the Autocorrelation function (ACF) and uses this recursive formula to estimate the ϕ's.

Conditions necessary to ensure stationarity

Example

$$\begin{array}{lll} AR(1) & \to & |\phi| < 1 \\ AR(2) & \to & \phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1 \\ & & |\phi_2| < 1 \end{array}$$

Question *How to choose p?*

There is a common tradeoff in statistics: a high p means higher fidelity, but fewer and fewer observations to estimate each coefficient. In other

words, there's a *tradeoff between model accuracy and complexity*. The following numerical criteria can be minimized to identify an optimal order balancing fidelity and complexity:

- AIC (Akaike Information Criterion, aka Final Prediction Error)
- BIC (Bayesian Information Criterion, aka Minimum Description Length)

Both penalize the misfit and model complexity (the number of parameters, p) in different ways, and will generally yield different estimates for the optimal p. Such criteria can both either overfit or underfit, depending on the situation. There is no foolproof method[1].

[1] The timeseries analysis (tsa) module of the excellent statsmodels Python package allows to fit any number of such timeseries models, and the fitted models will include an estimate of AIC and BIC. See this blog post for an example on Sunspot number data

Theoretical spectrum of an AR(p) process

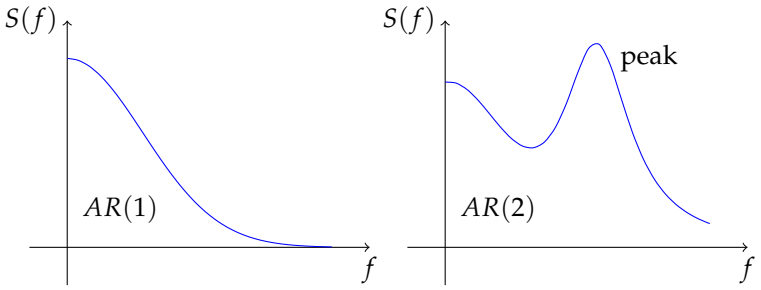


Figure 8.7: Examples of spectra that can be fit by AR(1) and AR(2) spectra. In general, AR(p) models can fit at most $p-1$ peaks

A big advantage of modeling with AR(p) processes is that their theoretical spectrum is known, and depends solely on the ϕ_j's:

$$S(f) = \frac{\phi_0}{|1 + \sum_{j=1}^{p} \phi_j e^{2\pi i j f}|} \tag{8.8}$$

They can fit a wide array of spectral shapes; this formula is the basis for Maximum Entropy spectral estimation (Chapter 9)

III Noise color

White noise

By analogy with the color white, which blends all colors of the rainbow in equal amounts, white noise exhibits power at all frequencies. The spectrum is flat ($S(f) \propto f^0 = 1$), which corresponds to pure Gaussian noise (an AR(0) process).

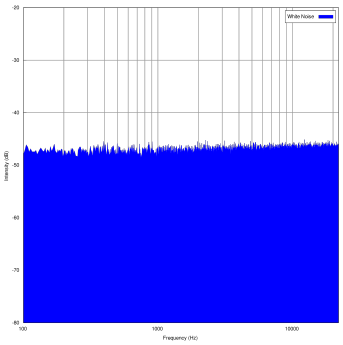


Figure 8.8: White noise

Red noise

Red noise refers to processes which exhibit a spectral slope $S(f) \propto f^{-2}$. This is certainly the case for an AR(1) process near the inflection point of its spectrum (Fig. 8.4). More generally, red noise can be generated by integrating white noise over time. An example of this is the celebrated model of *Hasselman* (1976), which shows how simple ocean mixed-layer physics may "redden" random weather fluctuations to produce power at low-frequencies. This physical plausibility is one reason why red noise is such a common null hypothesis. Note that the term red noise may be loosely applied to many spectra that exhibit decreasing power with increasing frequency. It is sometimes called "Brown" noise, not for the color brown but rather as a corruption of Brownian motion. Often used exchangeably with the term "AR(1) process".

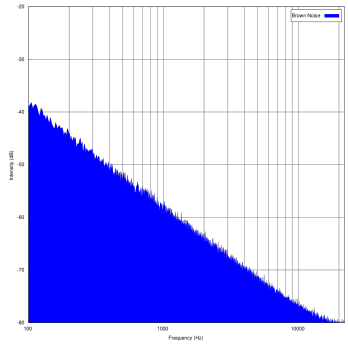


Figure 8.9: Red noise

Blue noise

Blue noise is used loosely to refer to $S(f) \propto f^2$ or $S(f) \propto f^2$, that is power increasing with frequency (Fig. 8.10). Blue noise is not common in the geosciences, and that's all the space we'll devote to it.

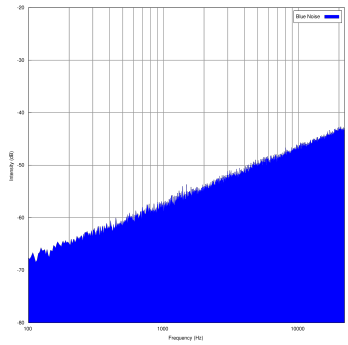


Figure 8.10: Blue noise

Chapter 9

SPECTRAL ANALYSIS

The object of spectral analysis is to estimate the partitioning of energy between frequency bands. Owing to Parseval's theorem (Eq. (7.18)), this means evaluating the spectral density $|\tilde{h}(\omega)|^2$ of a signal $h(t)$. Doing so from a sequence of finite and possibly noisy measurements, we must *estimate* this density and characterize its associated uncertainties. The notions we saw in Part I will therefore come to good use.

Before we venture too far into technical details, however, let us pause and reflect on the scientific motivations underlying such a complicated analysis.

I Why spectra?

What do we hope to learn from the spectrum? Why is it worth calculating in the first place? To see this, let us consider three Earth science examples.

Earth's normal modes

First, a simple example from low-frequency seismology[1]: if you examine the Fourier amplitude spectrum of a long vertical-component seismogram from a big earthquake ($M > 7$) at frequencies of 1-5 mHz, you will see distinct spectral lines that can be associated with the eigenfrequencies of Earth's spheroidal normal modes. That is, the Earth is like a bell that rings at certain notes (harmonics) when hit by the hammer of large Earthquakes. These harmonics reveal fundamental aspects of its inner structure.

Each of these discrete modes can be matched with spherical harmonics (cf Appendix C) of fixed angular order, as illustrated in Fig. 9.1.The amplitude of the spectral peaks constrains the earthquake source excitation. These data are especially important for very large earthquakes, which have linear dimensions comparable to the mode wavelength. Also, the peaks move around a bit (in frequency) from earthquake to earth-

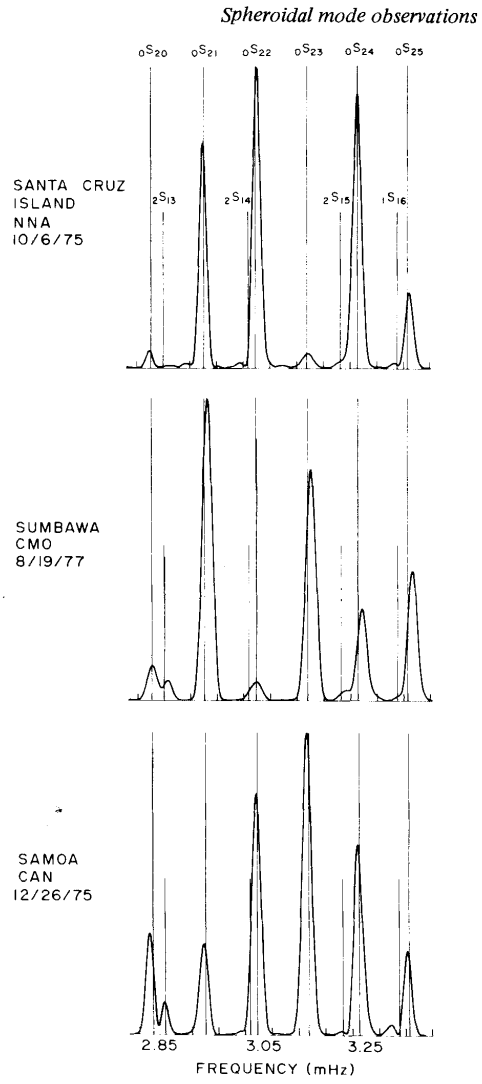


Figure 9.1: Expression of Earth's normal modes in seismograms. Fig. 6 of *Silver and Jordan* (1981)

[1] courtesy of Prof T.H. Jordan

quake and station to station. This variation arises because the mode degeneracy is split by 3D Earth structure, and it can be precisely interpreted and inverted. These "apparent eigenfrequency" data are used in most global 3D earth models.

The continuum of climate variability

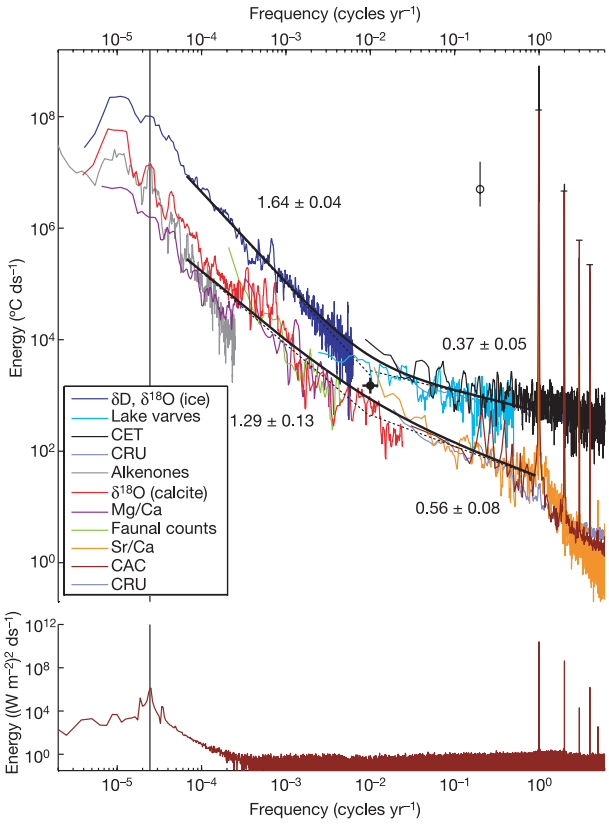


Figure 9.2: **Patchwork spectral estimate using instrumental and proxy records of surface temperature variability, and insolation at 65N**. Power-law estimates between 1.1–100 yr and 100–15,000 yr periods are listed along with standard errors, and indicated by the dashed lines. The sum of the power-laws fitted to the long- and short-period continuum are indicated by the black curve. The vertical line-segment indicates the approximate 95% confidence interval, where the circle indicates the background level. The mark at 1/100 yr indicates the region mid-way between the annual and Milankovitch periods. At bottom is the spectrum of insolation at 658 N sampled monthly over the past million years plus a small amount of white noise. The vertical black line indicates the 41-kyr obliquity period.

In most real-world spectra the peaks are superimposed on a "background". Peaks correspond to components that are periodic or nearly so. They often reveals eigenmodes (as in the previous case) or the mainly linear response of a system to periodic forcing (e.g. seasons). The background connecting the peaks is informative of the processes that transfer energy between scales. As pointed out by *Lovejoy* (2015), it is often at least as interesting as the peaks.

Fig. 9.2 (reprinted from *Huybers and Curry*, 2006) shows a composite spectrum attempting to portray surface temperature variability from timescales of 1 month to 100,000 years. One can see that most of the energy is associated with the annual cycle and its higher order harmonics (vertical lines on the right). The second-most important source of variability are broadband peaks near the so-called Milankovitch periodici-

ties near 23, 41, and 100 kyr, clearly visible in the insolation spectrum on the bottom. While those frequencies show up strongly in the climate response as well (top), the main point of the article was to draw attention to the slope of the background joining those peaks, and how this slope appears to change at the centennial scale (black cross). The authors hypothesized that this change in spectral slope (aka scaling exponent) is indicative of different processes accomplishing energy transfers between space and time scales.

Testing theories of geophysical turbulence

Theories of fluid flow in the atmosphere and oceans make predictions about rates of energy transfer between spatial and temporal scales that are most easily summarized in the spectral domain. In particular, such theories predict the existence of scaling regimes: a power-law behavior (i.e. linear behavior in a log-log plot) of the spectrum of various state variables (temperature, velocity, passive tracers), which is indicative of scale invariance. As an example, consider radiance measurements from satellites as presented by *Lovejoy et al.* (2008).

Fig. 9.3 shows the "along track" 1D spectra from the Visible Infrared Sounder (VIRS) instrument of the Tropical Rainfall Measurement Mission (TRMM) at wavelengths of 0.630, 1.60, 3.75, 10.8, 12.0 mm, i.e. for visible, near infrared and (the last two) thermal infrared (IR) radiation. The first two bands are essentially reflected sunlight, so that for thin clouds the signal comes from variations in the surface albedo (influenced by the topography and other factors), while for thicker clouds it comes from nearer the cloud top via (multiple) geometric and Mie scattering. The units are such that $k = 1$ is the wavenumber corresponding to the size of the planet (20 000 km). The high-wavenumber fall-off is due to the finite resolution of the instruments.

A scaling behavior is evident from the largest scales (20 000 km) to the smallest available – there are no peaks in this spectrum, and all the interesting information is contained in the slope of the "background". This slope turns out to be incompatible with classical turbulence cascade models which assume well-defined energy sources and sinks, with a source and sink-free "inertial" range in between. Thus, these estimates of spectral scaling enable one to directly test theories of atmospheric turbulence. Such a test would be impossible (and meaningless) with unprocessed timeseries data; nothing would be learned by staring at wiggles, but staring at their spectra does teach us a lot.

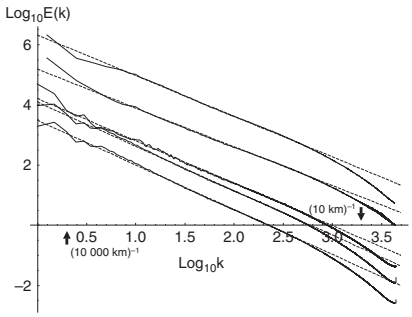


Figure 9.3: Scaling behavior. Spectra from channels 1-5 (at wavelengths of 0.630, 1.60, 3.75, 10.8, 12.0 μm from top to bottom, displaced in the vertical for clarity). The straight regression lines have spectral exponents $\beta = 1.35, 1.29, 1.41, 1.47, 1.49$ respectively, close to the value $\beta = 1.53$ corresponding to the spectrum of passive scalars (5/3 minus intermittency corrections). Note that the analysis was performed over space, not time, so the relevant spectral index is the wavenumber k, which is to wavelength what angular frequency is to the period. Adapted from *Lovejoy et al.* (2008)

Summary

From these examples we conclude that spectra are characterized by two main features – peaks and background – and that the properties of both (location and height for peaks, scaling law for the background) are deeply informative of the physics and dynamics underlying the measurements. Put it another way, spectra can reveal things that nothing else can reveal, and we hope that you are now convinced of their virtues. Let us thus estimate spectra until we grow numb.

II Signals, trends and noise

At this time it is useful to define a few terms of timeseries lore. We shall consider each timeseries $X(t)$ as the sum of one or more **periodic**[2], **stationary components** (or "signals"); a **trend** $q(t)$, and **noise** ε_t. The periodic components are the reason we're doing spectral analysis in the first place – it will tell us how big they are and what their frequency is.

[2] not necessarily harmonic

Trend is a very fashionable[3] word these days, and it is used and abused into meaninglessness. Here, it will refer to a slowly-evolving, non-stationary component[4]. For instance: a linear increase or decrease; a sinusoidal component so slow that its period is not resolved by the dataset; a nonlinear trend like the exponential increase of the Keeling curve (Fig. 15.7) or the red line in Fig. 9.4).

[3] shall we say trendy?

[4] Please only use it for this purpose, and banish the heinous "cyclical trend" from your vocabulary.

As for **noise**, it clearly involves a subjective definition. Under this name we usually subsume any variable component in which we are not interested. Some of this noise may be composed of actual measurement errors (what you would think of as noise), but some of it could be what another analyst would call "signal". If you study climate, daily fluctuations are "noise"; if you study weather, they are your bread and butter. One commonly says that *"one analyst's signal is another analyst's noise"*. This noise is often modeled as a Normal random process (aka white noise) with zero mean, $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$.

To summarize, our model is:

$$X(t) = q(t) + \sum_{k=0}^{N-1} \left[a_k \cos\left(\frac{2\pi k}{N} t\right) + b_k \sin\left(\frac{2\pi k}{N} t\right) \right] + \varepsilon_t \qquad (9.1)$$

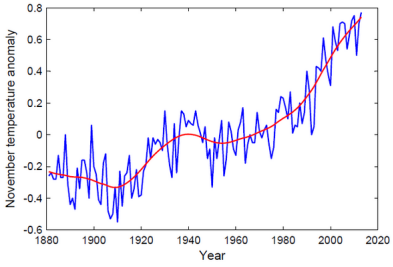


Figure 9.4: Global mean temperature for all November months since 1880 in the NASA GISS dataset. The non-linear trend is highlighted in red.

III Data pre-processing

In Chapter 7 we saw that Fourier analysis implicitly assumes a periodic timeseries, and that all digital estimates of Fourier components come with the curses of leakage and aliasing. Accordingly, some preparatory steps must be taken to minimize the most common problems.

Sampling

Analog signals may be pre-filtered to make them *band limited* prior to digitization. This operation brings any period to zero beyond a cutoff frequency f_c. Sampling usually refers to digitization at *even* increments Δ. If even increments are not possible (*e.g.* in a sediment core, time may not be linearly related to depth), there are two possibilities:

1. *interpolate* on regular time grid, as seen in the next Chapter (this might introduce spurious features) and use the methods of this chapter.
2. *use methods designed for irregularly-spaced data* like the Lomb-Scargle periodogram[5](*Mudelsee et al.*, 2009) or the weighted wavelet Z-transform (*Foster*, 1996).

[5] also known as Least-Squares Spectral Analysis

Detrending

Example
In the global warming example of Fig. 9.4, it would make sense to remove the slowly-evolving component (red line) if interested in periodic or quasi-periodic oscillations

Detrending always needs a very good justification, because one might be inadvertently throwing out the signal with the "trend". For instance, if the instrumental record were 5 times as long, you might find that the red line is actually a quasi-periodic low-frequency oscillation as well. If you choose to detrend, you must therefore avoid making any statement about the lowest frequency variability.

Tapering

Tapering refers to bringing the edges of a timeseries to zero. There are two reasons to do so:

Edge effects We have seen that DFT periodizes the signal $h(t)$ (so the sequence H_n is N-periodic). If $h(N-1) \neq h(0)$, the jump $h(N-1) - h(0)$ will generate a discontinuity at the junction (*cf.* Fig. 9.5), so higher-order harmonics will pollute the spectrum via Gibbs' phenomenon: this an example of *edge effect*. To some extent, detrending will help with that. Multiplying by a *taper*, however, will ensure that both edges are zero exactly, eliminating this edge effect.

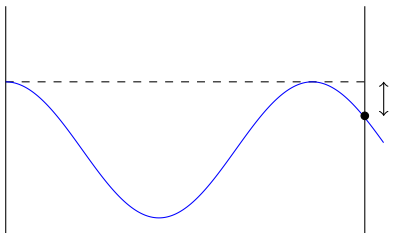


Figure 9.5: Edge effects may appear ins the simplest settings.

Minimizing leakage recall that spectral leakage is a consequence of observing a system over a finite sample, which is akin to multiplication by a boxcar (whose Fourier transform is proportional to $\mathrm{sinc}(\pi f t)$). Tapers are often designed to *minimize leakage* outside the main lobe. But there is always a trade-off. Tapering will broaden central peaks

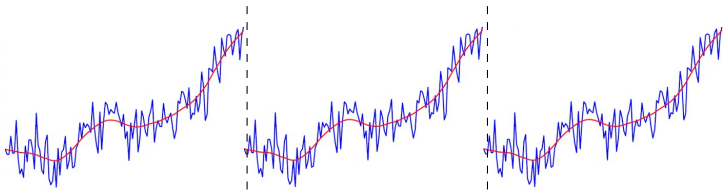


Figure 9.6: Periodization of a non-stationary signal. In this case, the large trend generates discontinuities that would mar the spectrum, if no care is taken.

and give a smoother appearance to the spectrum; this is good or bad, depending on what you want.

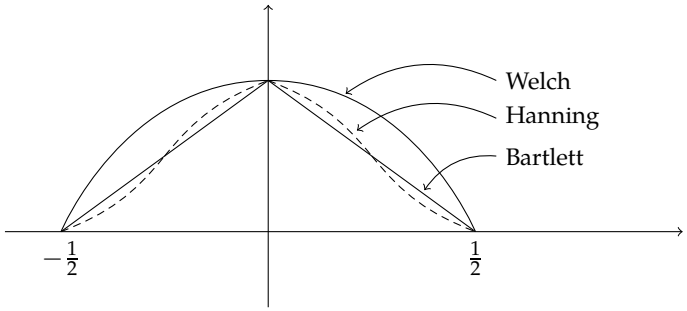


Figure 9.7: Window carpentry

There are several choices available (Fig. 9.7), depending on what you're trying to achieve:

- Hanning window $w(n) = \sin^2\left(\frac{\pi n}{N-1}\right)$
- Bartlett (triangular) window:

$$w(t) = \begin{cases} \frac{2}{T}\left(t + \frac{T}{2}\right) & t \in \left[-\frac{T}{2}, 0\right[\\ -\frac{2}{T}\left(t - \frac{T}{2}\right) & t \in \left[0, \frac{T}{2}\right] \end{cases} \tag{9.2}$$

  (convolution of two boxcars)

- Gaussian window
- Parzen window.

There is a whole menagerie of tapers, illustrated in Fig. 9.7. Designing windows with certain spectral properties was once referred to as "window carpentry". Fortunately, nowadays, the Multi-Taper Method (MTM, Sect. V) solves this conundrum for us in an optimal way.

Zero-padding

Because the FFT algorithm works optimally on datasets with a sample size equal to a power of two, it is common to pad with zeroes to reach the next power of two: $N \rightsquigarrow N' = 2^{m}$[6]. This will make the Fast Fourier Transform algorithm faster (though it is not mandatory); it may reduce

[6] `math.ceil(math.log(number,2))`

frequency spacing, yielding a smoother spectrum; and it theoretically does not add or destroy information. However, this is only true if the series is stationary and there is no jump at the end points; so it must always be used in conjunction with a taper.

IV Classical Spectral Estimation

Correlogram *vs.* Periodogram

The simplest way to compute a spectrum is the *periodogram*:

$$\hat{S}(f_n) = P_n = |H_n|^2 \tag{9.3}$$

for the n^{th} frequency component, f_n. Unfortunately, this turns out to be a very bad idea. In the old days, the *correlogram* (Blackman and Tukey, 1958) was more reliable. Recall the Wiener-Khinchin theorem, with $a(\tau) = \int_{\mathbb{R}} h(t)h(t+\tau)dt$, then:

$$\tilde{a}(\omega) = \tilde{h}(\omega)\tilde{h}(\omega)^* = |\tilde{h}(\omega)|^2 = S(\omega) \tag{9.4}$$

In practice, $a(\tau)$ is estimated for each discrete lag k :

$$\hat{a}(k) = \frac{\frac{1}{N+|k|-1}\sum_{j=0}^{N-|k|} h_j h_{j+k}}{\frac{1}{N-1}\sum_{j=0}^{N-1} h_j^2}, \qquad k \in [-(N-1), +(N-1)], \tag{9.5}$$

Then apply the Discrete Fourier Transform[7].

[7] this estimate requires tapering as well, to minimize edge effects, as well as centering, so that the mean is zero.

Statistical considerations

Periodogram and correlogram are **estimates** of the true spectrum $S(f)$. Ideally, this estimator $\hat{S}(f)$ is **unbiased**: $E(\hat{S}(f)) = S(f)$. Now, the problem is with its variance; it turns out that

$$V(\hat{S}) = S$$

which is independent of N. Hence the estimate won't improve as $N \to \infty$: it is *inconsistent*. Put another way, increasing N only adds more frequency points f_n at which to estimate S, but at each of these points the number of observations going into the estimate $\hat{S}(f_n)$ does not change, so it does nothing to reduce uncertainties.

Another way of saying this is that the precision decreases with the height of a peak, so the more energetic components (i.e. the ones that dominate the spectrum) are the ones we are most uncertain about, which is vexing to the point of annoyance.

Confidence intervals

Recall that:

$$H_n = \sum_{k=0}^{N-1} h_k e^{-i\frac{2\pi}{N}kn} \text{ is a complex number} \tag{9.6}$$

$$|H_n|^2 = \Re(H_n)^2 + \Im(H_n)^2 \tag{9.7}$$

If

$$h_k \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

then

$$|H_n|^2 \approx X_1^2 + X_2^2, \qquad \text{where } X_1, X_2 \sim \mathcal{N}(0, \sigma^2)$$

so

$$\begin{aligned} \frac{|H_n|^2}{\sigma^2} &\sim \chi_2^2 \qquad \text{chi-squared with 2 d.o.f.} \\ &\sim \text{Exp}\left(\frac{1}{2}\right). \end{aligned}$$

So

$$P\left(\frac{|H_n|^2}{\sigma^2} \geq g_\alpha\right) = 1 - \left[1 - e^{\frac{-g_\alpha}{2}}\right]^N \tag{9.8}$$

where g_α is a number estimated from the quantile function of an exponential distribution with parameter $1/2$, for a given confidence level α. Because the exponential distribution is highly skewed to the right, the corresponding confidence intervals (C.I.) are skewed towards the high end.

The Windowed Correlogram

Although more computationally intensive, the correlogram can be made consistent through a cunning use of $0 < M \ll N$ windows. This is the essence of Blackman-Tukey spectral analysis, which dominated the field for a long time.

$$\hat{S}_{BT}(f) = \left| \sum_{k=0}^{N-1} w_k a_k e^{-2\pi i f \Delta k} \right| \tag{9.9}$$

where:

- w_k = window function (taper);
- a_k = auto-correlation at lag k.

With a judicious choice of window:

$$V\left(\hat{S}_{BT}\right) \approx \frac{3M}{2N}\hat{S}_{BT} \overset{N\to\infty}{\longrightarrow} 0 \tag{9.10}$$

so it is now *consistent* (hurray!). But that, of course, comes at a price: the bias-variance trade-off rears its ugly head again: increasing M will increase variance but decrease leakage; decreasing M will decrease variance but increase leakage. So once again there is a compromise to be made. No free lunch.

Ghil et al. (2002) write: "it turns out that the [...] windowed correlogram method is quite efficient for estimating the continuous part of the spectrum, but is less useful for the detection of components of the signal that are purely sinusoidal or nearly so (the lines). The reason for this is twofold: the low resolution of this method and the fact that its estimated error bars are still essentially proportional to the estimated mean $\hat{S}(f)$ at each frequency f".

Welch's averaged periodogram

Idea: chop series into K segments of length $\frac{K}{N}$:

$$V\left(\hat{S}_{\text{Welch}}(f)\right) = \frac{S}{K} \tag{9.11}$$

$$\text{for i.i.d. Gaussian } h_k, \quad \xi = \nu\frac{\hat{S}(f)}{S(f)} \sim \chi^2 \tag{9.12}$$

where $\nu =$"equivalent d.o.f." , which depends on K and shape of tapers. Typically, $\nu = 2K - 2$. Then:

$$P\left(\chi^2_{\nu\left(\frac{\alpha}{2}\right)} \leq \nu\frac{\hat{S}(f_n)}{S(f_n)} \leq \chi^2_{\nu\left(1-\frac{\alpha}{2}\right)}\right) = 100 \times (1-\alpha)\% \tag{9.13}$$

This yields a parametric 95% confidence interval for the spectral estimate. If the data are not i.i.d. Gaussian, you would be better inspired to use a non-parametric test (see examples below). In Python, this method maybe accessed as `scipy.signal.welch`.

V Advanced Spectral Estimation

The previous methods were all fine and good in the 1960's, but in this day and age we can do much better.

Maximum entropy method

The first one uses the notions of Chapter 8 to approximate a time series $h(t)$ by an autoregressive model of order p, $AR(p)$. It thus performs best when estimating line frequencies for a time series that is actually generated by such a process.

The maximum entropy method (MEM) determines the spectral density $\hat{S} = S_{\text{MEM}}$ *that is associated with the most random, or least predictable AR process that has the same auto-correlation function (ACF)* $\hat{a}(k)$ *(Eq. (9.5))* as the dataset. In terms of information theory, this corresponds to the concept of *maximum entropy*, hence the name of the method (*Ghil et al.*, 2002) Fig. 9.8.

In practice, one estimates p coefficients $(\phi_1, \phi_2, \ldots, \phi_p)$ from the ACF. Knowledge of these coefficients enables to compute the spectrum according to the theoretical spectrum of an AR process (Eq. (8.8)). Any good routine[8] will do all this for you given a model order p.

[8] http://thomas-cokelaer.info/software/spectrum/html/user/tutorial_pburg.html

---

Notes on MEM

- Because everything comes down to an $AR(p)$ process, one must be very careful in choosing p. One objective way to do this is via several *information criteria* (*Neumaier and Schneider*, 2001; *Stoica and Selen*, 2004) like Akaike's Information Criterion (AIC, *Akaike*, 1974) and the Bayesian Information Criterion (BIC, *Schwarz*, 1978)[9]. In many cases, however, these tend to over- or under-estimate p depending on the timeseries' properties (*Ghil et al.*, 2002).
  Possible solutions are
  - try a few different p's and look for robust features;
  - compare with other methods (Welch, BT, MTM).
- The number of special peaks grows with p so it's easy to "find" as many peaks as you want. Nonetheless, MEM is extremely good in detecting split peaks, if you know where they should be.

[9] aka Minimum Description Length, or Final Prediction Error

---

Multi-Taper Method

MEM can detect amazingly fine features (e.g. split peaks), but the choice of p tends to be arbitrary; further the method only makes sense if an

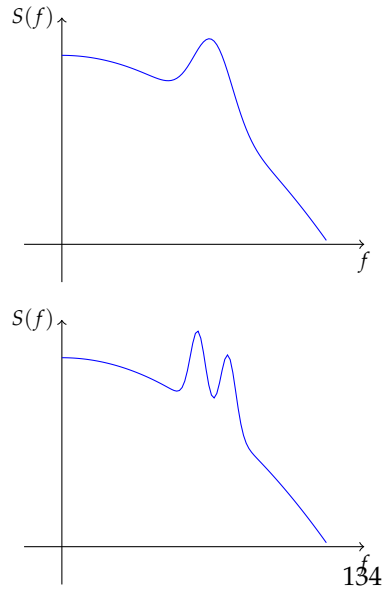


Figure 9.8: Smoothed peak (obtained with Blackman-Tukey) versus split peaks obtained with MEM

AR model is an adequate approximation of the series. The Blackman-Tukey method is even worse, since the choice of taper is rather *ad hoc*. In 1982, D. Thomson found an optimal solution, known as the Multi-Taper method (MTM, *Thomson*, 1982)

MTM uses K orthogonal tapers $(w_k(t), k = 1, \cdots, K)$ belonging to a family of special functions (Discrete Prolate Spheroidal Sequences, or Slepian functions, Fig. 9.9). These $w_k(t)$ solve the variational problem of minimizing leakage outside of a frequency band $f_0 \pm pf_R$ where $f_R = 1/(N\Delta)$ is the Rayleigh frequency and p is some nonzero integer. Averaging over tapered spectra yields a more stable estimate, as $\mathrm{Var}(S(f_0)) \propto 1/K$. In practice only the first $2p-1$ tapers have usefully small leakage so we can take $K \leq 2p-1$. The **bandwidth** $2pf_R$ is the width of a harmonic line with this method. There is a trade-off between minimizing the variance $(\propto 1/K)$ and minimizing the leakage $(\propto 2p)$ but this tradeoff is *optimal* by design (e.g. the choice $p = 2$ and $K = 3$ is quite reasonable). MTM estimates both:

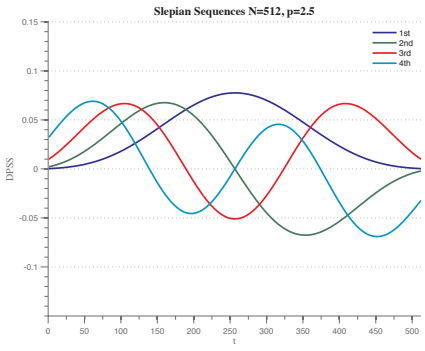


Figure 9.9: Slepian functions used to produce the tapers $w_k(t)$. As a rule, the k^{th} taper has k bumps.

- the background (as the Blackman-Tukey method does),
- the lines (as the Maximal entropy method does).

It is adaptive and explicit, unlike many of its competitors. The spectrum is obtained by averaging all the tapered spectra:

$$S_{\text{MTM}}(f) = \frac{\sum_{k=1}^{K} \mu_k \mid Y_k(f) \mid^2}{\sum_{k=1}^{K} \mu_k} \tag{9.14}$$

where $Y_k(f)$ is the DFT of $h(t) \times w_k(t)$ and μ_k its corresponding weight, chosen adaptively by the algorithm. In Python[10], you must provide n_w (the time-bandwidth product), which can take the values $[2,\ 5/2,\ 3,\ 7/2,\ 4]$. Small values of the parameter mean high spectral resolution, low bias, but high variance. Large values of the parameter mean lower resolution, higher bias, but reduced variance. No free lunch, as usual. The choice is subjective, but since it formulates an explicit tradeoff, it may be argued rationally.

[10] https://krischer.github.io/mtspec/

In our opinion, MTM is usually the best option, but it must be tried with several different bandwidths to make sure the results are robust. If a feature appears with only one method or one choice of parameter, you should always regard it as suspect, and conduct additional analysis to confirm it.

Uncertainty Quantification

So, you've estimated a spectral density, ideally using several methods, or with one method but different choice of its parameters (e.g. time-bandwidth product, AR model order, etc). What now? The most common use of spectral analysis is to identify periodic components, and hav-

ing done so, decide whether they are *significant*. Aaaargh, the dreaded S word again! We are now back in the realm of Chapter 6.

White nulls

As discussed earlier, the frequentist view is that if $X(t)$ is IID normal (white noise) then the Fourier coefficients are distributed as χ^2 variables with 2 degrees of freedom, so one can test against the null hypothesis that the spectrum was generated by such a process; most spectral analysis routines will provide confidence intervals for $S(f)$ based on this null distribution. For instance, MTM also enables to test for the significance of peaks using an F ratio against a null hypothesis of white noise. There are some extensions (*Mann and Lees*, 1996) to test against a red noise null hypothesis (AR(1) process), which is often much more appropriate in the geosciences.

Autoregressive nulls

As you saw in Chapter 8 and in lab 7, individual AR processes may exhibit spectra that differ considerably from their theoretical shape, featuring prominent bumps that one might erroneously interpret as 'real' (Fig. 8.6). It is thus crucial to guard against these pitfalls.

One solution is to simulate a large number of AR nulls (say, with an autocorrelation parameter comparable to that of $X(t)$), compute their spectrum as you did for $X(t)$, and empirically estimate the quantiles of this distribution. Such an approach is illustrated in Fig. 9.10, which shows the result of estimating uncertainties in the estimate of δ^{18}O in a speleothem from Vanuatu (*Partin et al.*, 2013). The thin gray curves encompass 95% of the AR(1) spectra, showing that, although several broadband peaks are present, few lie outside the AR(1) envelope. This implies that the record is generally consistent with red noise, with a few notable exceptions in the interannual (ENSO) and multidecadal bands.

This is a good occasion to point out some good practices on graphically displaying spectra. The output of an MTM routine [11] will be phrased in terms of angular frequency[12]. It is often much more meaningful to phrase things in terms of the period of oscillation, as done on this graph. You will also notice that both axes are logarithmic, without which the spectrum would be squished to the left-hand corner, and none of the interesting details would appear. Sometimes a semi-log scale may be more sensible; the choice should ultimately serve to highlight interesting parts of the spectrum, and what's "interesting" is a subjective matter. Apply common sense.

One drawback of the log-log representation is that the area occupied by the low-frequencies appears disproportionately large, so that integrating the area under the curve does not represent the energy of the sig-

[11] for Python, see the nitime package

[12] from 0 to π, the latter corresponding to the Nyquist frequency

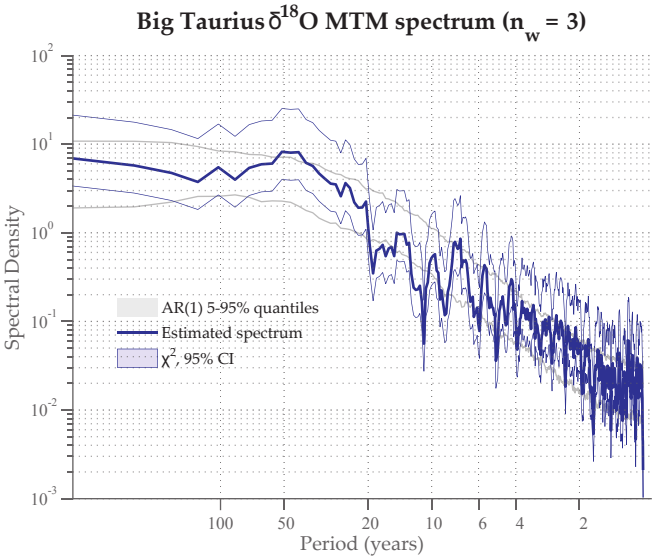


Figure 9.10: Example of 95% confidence interval for $\hat{S}$. The thin blue lines depict χ^2-based intervals, while the thin gray lines depict a 95% confidence region from an ensemble of 1,000 AR(1) timeseries (Chapter 8), which serve as a null hypothesis for the significance of spectral peaks.

nal in that frequency band. A solution to this problem uses a variance-preserving log-scale (Fig. 9.11). This representation plots $\log(f \times S(f))$ as a function of $\log(f)$, which conserves variance in each band and thus allows to "integrate by eye" (A. Wittenberg, pers. comm., 2012)

For reasons elaborated in Chapter 8, an AR(1) null hypothesis is often sensible in the Earth Sciences, but is by no means an absolute rule; the key is to realize that every estimated spectrum must be gauged with respect to a suitable null hypothesis before claiming that some of its features are "significant". Also pursuant to the discussion at the end of Chapter 6, recall that statistical significance is not the same as practical significance. In some cases, an overly stringent null hypothesis would have us reject things that we know to be true! For instance, we've seen some econometrists argue that temperature trends should be gauged against some nulls called "fractional Brownian motion", though it's been shown with physically-based models (*Mann*, 2011) that this statistical vitriol can dissolve even the most real trends into "noise". In summary, you should always choose a null hypothesis based on scientific, not just statistical, considerations.

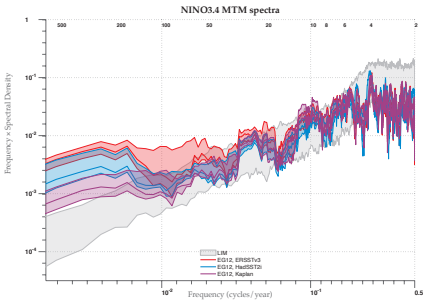
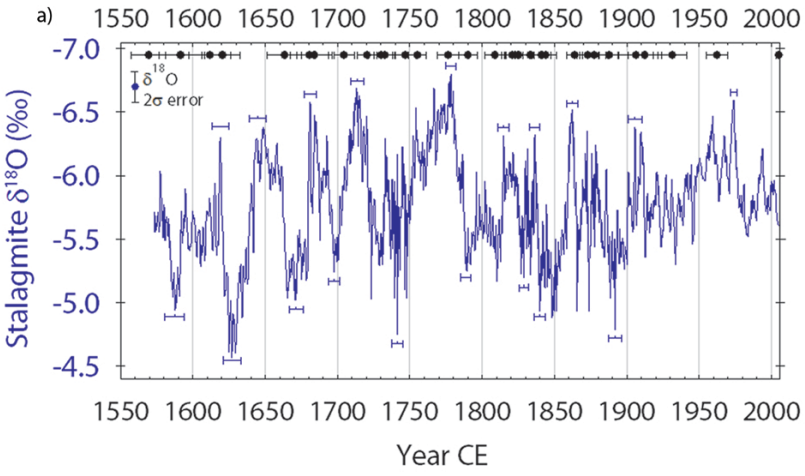


Figure 9.11: Comparisons of the MTM spectra of 3 reconstructions of the NINO3.4 index (a common metric of ENSO) with a simple stochastic null hypothesis (*Ault et al.*, 2013).

Age-uncertain spectral analysis

As an example of a scientifically-motivated null hypothesis, let us consider the study of *Partin et al.* (2013), who used oxygen isotope measurements in a South Pacific speleothem to learn about low-frequency hydroclimate variability. The measured timeseries may be seen in Fig. 9.13. As in most climate proxy records, the time axis is uncertain to some degree, in the sense that many possible functions could fit the age-depth constraints (Fig. 9.12). One approach to this is to use ensemble methods (Monte Carlo simulations), generating many possible realizations (say, N = 10,000) of the timeseries that are permissible within age uncertainties.



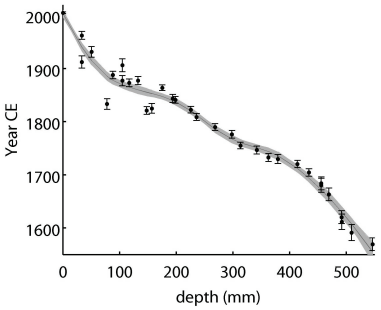


Figure 9.12: Age model for the Big Taurius record of *Partin et al.* (2013), based on 31 U-Th ages and a top date of 2005. The grey envelope represents 10,000 possible spline fits of the age model, whose median is depicted by the dark curve.

Figure 9.13: Time series of stalagmite δ^{18}O from Taurius Cave in Espiritu Santo, Vanuatu from 1557 to 2003 C.E. U-Th ages are represented by black points on top, with 2σ analytical error bars. Blue error bars along the record are 2σ errors for individual δ^{18}O measurements calculated using 10,000 realizations of the age model, and are included to depict range of uncertainty associated with peaks or valleys. From *Partin et al.* (2013)

One may then compute the resulting spectra for each realization of the age model, and see how the spectrum of the original timeseries (blue curve in Fig. 9.13) fares in relation to this ensemble (Fig. 9.14). The analysis shows that age uncertainties would tend to blur high-frequency peaks rather than creating them, lending credence to the notion that the peaks observed in Fig. 9.10 are real.

Robust spectral analysis

Building on the MTM, *Chave et al.* (1987) extended the theory to non-Gaussian signals and devised non-parametric confidence intervals via the jackknife method. This is particularly handy if your data are highly non-normal and if no parametric null hypothesis applies to your problem. Some excellent Matlab code may be found on the page of Frederik J. Simons, along with many useful goodies. In Python, the Python `nitime` library makes this easy as well.

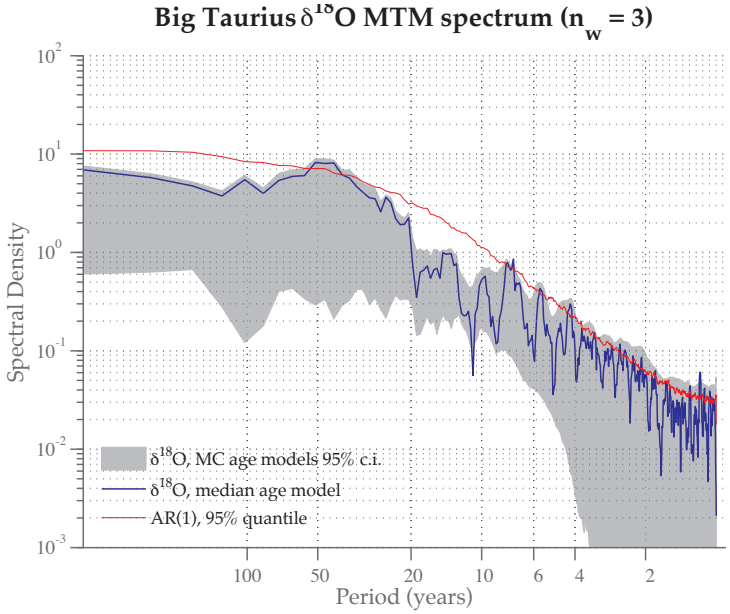


Figure 9.14: MTM spectral analysis of the Fig. 9.13 series using the median age model (blue). Gray shading delineates 95% confidence interval obtained from 10,000 realization of the age model. From *Partin et al.* (2013)

VI Cross-spectral analysis

Consider the Fourier transform pairs $x(t) \circ\!\!-\!\!- X(f)$, $y(t) \circ\!\!-\!\!- Y(f)$. Remember the Wiener-Khinchin theorem

$$S(f) = \mid X(f) \mid^2 \; -\!\!-\!\!\circ \gamma_{xx}(t) \tag{9.15}$$

where $\gamma_{xx}(t)$ is the autocorrelation function (ACF).

1. What if we wanted to examine the spectral features of the cross - correlation? Remember the definition of γ_{xy}

$$\gamma_{xy} = \int_{-\infty}^{+\infty} x(t)y(t+\tau)dt. \tag{9.16}$$

In practice

$$\tilde{\gamma}_{xy}(k) = \frac{1}{\sigma_x^2\sigma_y^2}\frac{1}{N-|k|-1}\underbrace{\sum_{j=0}^{N-|k|} x_j y_{j+k}}_{\text{cross-covariance}} \tag{9.17}$$

in which we have assumed that x and y are centered.
We have seen that

$$\gamma_{xy}(t) \circ\!\!-\!\!- X(f)Y^*(f)$$

so if $Y = X$ we will have

$$\mathcal{F}(\gamma_{xx}(t)) = \mid X(f) \mid^2 = S(f).$$

S is therefore the *auto-spectrum* of X. The cross-spectrum is given by

$$\Gamma_{xy}(f) = \sum_{\tau=-\infty}^{+\infty} \gamma_{xy}(\tau) e^{-2\pi i \tau f}$$

or

$$\hat{\Gamma}_{xy}(f_n) \underset{N}{\longmapsto} \hat{\gamma}_{xy}(k) (\in \mathbb{C}).$$

We can express the cross-spectrum as

$$\Gamma_{xy}(f) = A_{xy}(f) e^{i\psi(x,y)} = A_{xy} e^{i\psi_{xy}(f)}$$

where A_{xy} and ψ_{xy} are called *amplitude* and *phase spectrum* respectively.

2. The amplitude spectrum reveals areas of common power between x and y. Most often, however, one looks at the *coherence*

$$\kappa_{xy}(f) = \frac{\mid A_{xy} \mid^2}{\mid X \mid^2 \mid Y \mid^2} = \frac{\mid \Gamma_{xy} \mid^2}{S_{xx} S_{yy}}$$

which is eerily reminiscent of the correlation coefficient

$$\rho_{xy} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\mathrm{Cov}(x,y)}{\sqrt{v_x}\sqrt{v_y}}$$

so *the coherence spectrum is the frequency-domain equivalent of the correlation coefficient*.

Properties

- We have that $0 \leq \mid \kappa_{xy} \mid \leq 1 \quad \rightarrow \quad$ and it is easy to interpret the value of $\mid \kappa_{xy} \mid$: if $\mid \kappa_{xy}(f_0) \mid = 1$ for a certain frequency f_0, then $x(t)$ and $y(t)$ are said to be *coherent* at that frequency.
- If $y(t)$ is the output of a linear filter (transfer function $H(f)$) whose input is $x(t)$, then, considering that

$$Y = [HX](f) \quad \text{and} \quad \Gamma_{xy} = X(f) Y(f)^*$$

we have

$$\kappa_{xy} = \frac{\mid \Gamma_{xy} \mid^2}{\mid X \mid^2 \mid Y \mid^2} = \frac{\mid X(f) \mid^2 \mid H(f) \mid^2 \mid X(f) \mid^2}{\mid X(f) \mid^2 \mid H(f) \mid^2 \mid X(f) \mid^2} = 1 \qquad (9.18)$$

at all f's. So, if the two series are related via a linear filter, they'll be perfectly coherent. Deviations from unity imply a non-linear response or no response at all.

3. For the phase spectrum we have

$$\psi_{xy}(f) = \begin{cases} \arctan\left(\frac{\mathrm{Im}(\Gamma_{xy})}{\mathrm{Re}(\Gamma_{xy})}\right) & \text{if both are } \neq 0 \\ 0 & \text{if } \mathrm{Im}(\Gamma_{xy}) = 0 \text{ and } \mathrm{Re}(\Gamma_{xy}) > 0 \\ \pm\pi & \text{if } \mathrm{Im}(\Gamma_{xy}) = 0 \text{ and } \mathrm{Re}(\Gamma_{xy}) < 0 \\ \pi/2 & \text{if } \mathrm{Im}(\Gamma_{xy}) > 0 \text{ and } \mathrm{Re}(\Gamma_{xy}) = 0 \\ -\pi/2 & \text{if } \mathrm{Im}(\Gamma_{xy}) < 0 \text{ and } \mathrm{Re}(\Gamma_{xy}) = 0 \end{cases}$$

The phase spectrum is usually expressed in degrees ($\times 180/\pi$). It expresses the phase lag between two signals at various frequencies. For example

$$\begin{cases} \psi_{xy}(f) \in [0, \pi] & y \textit{ lags } x \text{ at the frequency } f \\ \psi_{xy}(f) \in [-\pi, 0] & y \textit{ leads } x \text{ at the frequency } f \end{cases}$$

where "*lags*" and "*leads*" can never be interpreted causally [13]. While the coherence measures spectral bands of common power, it says nothing about the relationship between x and y. As before, correlation is not causation.

[13] e.g. does y *lead* by $T/4$ or *lag* by $3T/4$? It is impossible to tell

Chapter 10

SIGNAL PROCESSING

In this chapter we tackle two essential elements of signal processing. The first, filtering, has to do with bringing out parts of a signal that we care about and silencing others. The second, interpolation, has to do with changing time grids without altering the frequency content of a signal too much.

I Filters

A *filter* is an operation that removes only a part of a measured signal. Everyday examples include sunglasses (which filter UV and/or IR wavelengths) or the EQ on a stereo receiver or your iTunes player, where one can change the relative importance of bass, mids or trebles. Earth Science examples include:

Seismology filter out surface waves (long periods) to leave only P and S waves.

Climate filter short-term variability (weather) to focus on climate. Filter variability outside the 2-7 year periodicity band to highlight ENSO variability.

(and many more)

The uncertainty principle

Recall from Chapter 7 that localization in time is inversely related to localization in frequency; this is analogous to Heisenberg's celebrated Uncertainty Principle (Fig. 10.1).

Because of this property of all Fourier transform pairs, the fetch of the *frequency response* of a filter (*i.e.* how sharply it cuts the frequency domain) will turn out to be inversely related to its *impulse response* (*i.e.* its fetch in the time domain), so designing a perfect filter is impossible. There are hundreds of different filters designed to meet specific design

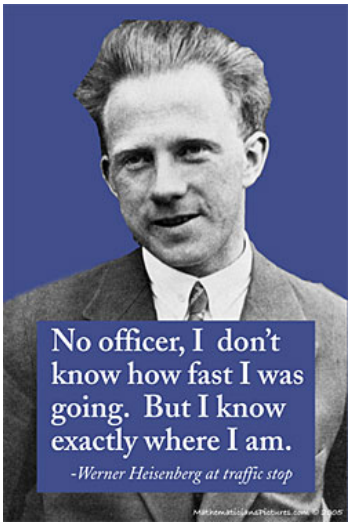


Figure 10.1: An illustration of Heisenberg's uncertainty principle, which states that Planck's constant is a fundamental limit on how accurately one might simultaneously know the location and velocity of a quantum object.

challenges; it is therefore important that you state your own desiderata before picking a filter.

Example
Lowpass filter: cut all frequencies above a certain frequency ω_c. This amounts to multiplying by a boxcar in the frequency domain, hence convolution in the time domain (Fig. 10.2). This will smear the features of the signal.

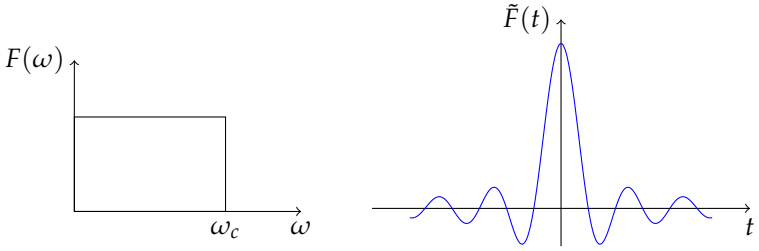


Figure 10.2: A sharp lowpass filter: boxcar function in frequency domain, cardinal sine in the time domain

Question *If the Fourier curse is so damning, why not stay in the time domain? Why not just average contiguous data points together, for instance?*

This is called a running mean. One might think that by averaging together M contiguous points, you'd filter all frequencies $f > \frac{1}{M\Delta}$. Unfortunately not: this time it amounts to convolution by a boxcar in the time domain, so smearing by a sinc function in the frequency domain: that is yucky. But there is worse: not only does this mess up the amplitude spectrum, but it destroys the phase spectrum even more. Defining the boxcar as:

$$b_k = \begin{cases} 1 & k \in [0, M-1] \\ 0 & k \geq M \end{cases} \tag{10.1}$$

Its DFT is (10.2)

$$B_n = e^{-i\pi_n \frac{M-1}{N}} \frac{\sin(\frac{n\pi M}{N})}{N \sin(\frac{n\pi}{N})} \equiv \underbrace{R_n}_{\text{amplitude}} \; e^{i \overbrace{\Phi_n}^{\text{phase}}} \tag{10.3}$$

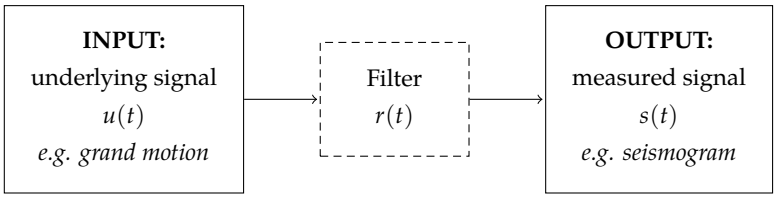
$$R_n = \left| \frac{\sin\left(\frac{n\pi M}{N}\right)}{\sin\left(\frac{n\pi}{N}\right)} \right| \qquad \text{has zeroes at } n = l\frac{N}{M}, \quad l \in \mathbb{N}^* \quad \text{and} \tag{10.4}$$

$$\Phi_n = -\frac{\pi n(M-1)}{N} - \varepsilon\pi \qquad \varepsilon = \text{sign}\left(\sin\frac{n\pi M}{N}\right) = 0 \text{ or } 1 \tag{10.5}$$

When $n = l\frac{N}{M}$, $\Phi_n = \mp\pi$: this is a 180 degree phase shift. So the running mean does cut out the frequency $n = \frac{N}{M}$ (as we hoped it would), but also all its integer multiples, and completely shifts the phase around at these points. Yikes! Such throwing the baby with the bathwater is a

terrible idea, though millions of scientists seem perfectly content with it on the grounds that it is "simple". Yes, but simple can be stupid. Let us now see how one would proceed to build better filters.

Linear Filter Theory



A *linear* filter verifies the convolution equation:

$$s(t) = r * u = \int_{-\infty}^{+\infty} r(t-\tau)u(\tau)d\tau \tag{10.6}$$

or, in the discrete world, a discrete convolution:

$$s_p = \sum_{k=0}^{p} r_k u_{n-k} \qquad \text{length } N \tag{10.7}$$

where r_k usually has length $\ll N$.

By the convolution theorem, $\tilde{s}(\omega) = \tilde{r}(\omega)\tilde{u}(\omega)$, so the output $\tilde{s}(\omega)$ contains the same frequencies as $\tilde{u}(\omega)$, but scaled by the amount $\tilde{r}(\omega)$.

Definition 13 (Transfer Function)

$$\tilde{r}(\omega) = \frac{\tilde{s}(\omega)}{\tilde{u}(\omega)} = \frac{output}{input} = \underbrace{G(\omega)}_{gain} e^{i \overbrace{\varphi(\omega)}^{phase\ shift}} \tag{10.8}$$

*where the **gain** G is the ratio of amplitudes: $\frac{|\tilde{s}(\omega)|}{|\tilde{u}(\omega)|}$ and the **phase shift** φ is the difference of phases ($phase(\tilde{s}) - phase(\tilde{u})$ at each frequency. $r(t)$ is called the* impulse response *of the filter, that is, the response to a unit pulse at $t = 0$.*

There are therefore three variables to consider for any filter. In general, one wants to specify G (e.g. lowpass) but φ and $r(t)$ will be affected as well; this trade-off is fundamentally due to the uncertainty principle). There are several classes of filters, each realizing a compromise between amplitude response (gain), phase response and impulse response:

1. *Analog* (instrument) *vs. digital* (computer)
2. *Time domain* (real time) *vs. frequency domain* (a posteriori)
3. *Infinite impulse response* (recursive) *vs. finite impulse response* (non recursive)

4. *Causal* (physically realizable put involving phase distortion *vs.* *acausal* (digital only: a zero phase shift **possible**)

Say we define a filter by a set of coefficients c_k and d_k, such that:

$$S_k = \sum_{n=0}^{n} c_n u_{k-n} + \sum_{j=i}^{N} d_j s_{k-j} \tag{10.9}$$

then:

$$\tilde{r}(f) = \frac{\sum_{k=0}^{n} c_k e^{-2\pi i f k \Delta}}{1 - \sum_{j=i}^{N} d_j e^{-2\pi i f j \Delta}} \qquad \begin{array}{l} \text{Finite Impulse Response (polynomial)} \\ \text{Infinite Impulse Response (rational)} \end{array} \tag{10.10}$$

If past values of s are allowed to enter the definition of s_k, more general filters can be designed, but their Impulse Response may not be compactly supported, and may have *poles* (singularities).

Some useful filters

Time domain

- *Shapiro filter*: $[1\ 2\ 1]/4 \quad (\sum \omega_i = 1)$
- *Shapiro* $*$ *Shapiro*: $[1\ 2\ 1] * [1\ 2\ 1] = [1\ 3\ 3\ 1]/8$
- Iterating this process n times yields a *binomial filter*, with weights proportional to binomial coefficients $\binom{n}{k}$
  Advantage: $R(f) = \cos^{2n}(\pi f)$ at order n and the phase shift is linear (easily corrected).

Frequency domain

- *Gaussian filters*: Convolution by a Gaussian leads to a certain amount of blurring, controlled by the half-width of the function. Because the Fourier transform of a Gaussian is a Gaussian with inverse scale, blurring in the time domain is inversely related to blurring in the frequency domain. See Appendix A, Sect. III. Fig. A.6
- *Butterworth filters*: $|Bu_n| = \frac{1}{1+\left(\frac{\omega}{\omega_c}\right)^{2n}}$
  Then choosing ε and r, ω_p and ω_s, allows the determination of ω_c and n:

$$n = \underbrace{\frac{\log\left(\frac{\varepsilon}{a}\right)}{\log\left(\frac{\omega_p}{\omega_s}\right)}}_{\text{governs sharpness}} \qquad \omega_c = \underbrace{\frac{\omega_p}{\varepsilon^{\frac{1}{n}}}}_{\text{governs cutoff location}} \tag{10.11}$$

  This filter's phase shift is linear as well. As $n \to \infty \to$, it tends to a boxcar. The choice of n is therefore a tradeoff between sharpness of the frequency response and the behavior in the time domain, which is quantified by the impulse response (Fig. 10.4).

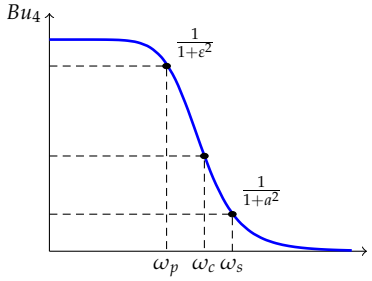


Figure 10.3: Butterworth filter

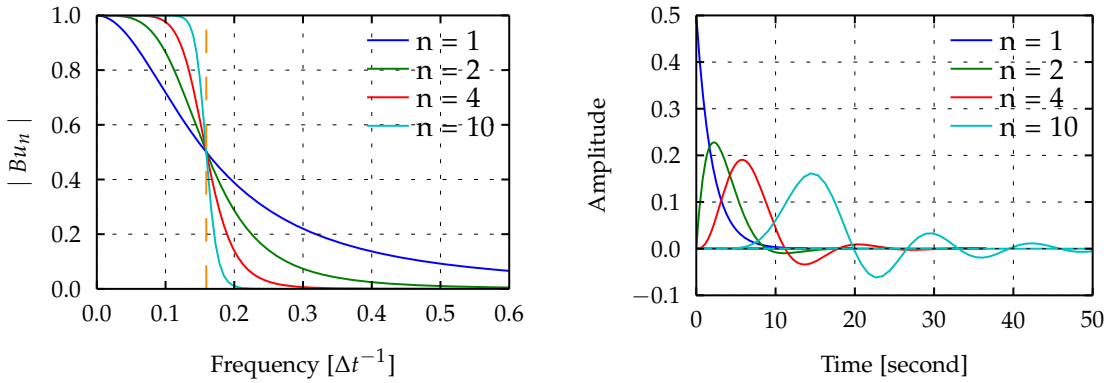


Figure 10.4: Examples of frequency (left) and impulse (right) response for Butterworth filters. The cutoff frequency is marked by a dashed red line. Notice that the larger n, the sharper the frequency response, but the longer the ringing in the time domain: there is no free lunch.

Filter nomenclature

Lowpass $L(\omega_c)$: lets through only low-frequencies ($\omega < \omega_c$)

Highpass the opposite; may be constructed as $1 - L(\omega_c)$

Bandpass lets through energy in a specific band (ω_1, ω_2). May be constructed by the difference of two lowpass filters $L(\omega_2) - L(\omega_1)$

Notch blocks the energy in a specific band (ω_1, ω_2). May be constructed as $1-$ a bandpass filter

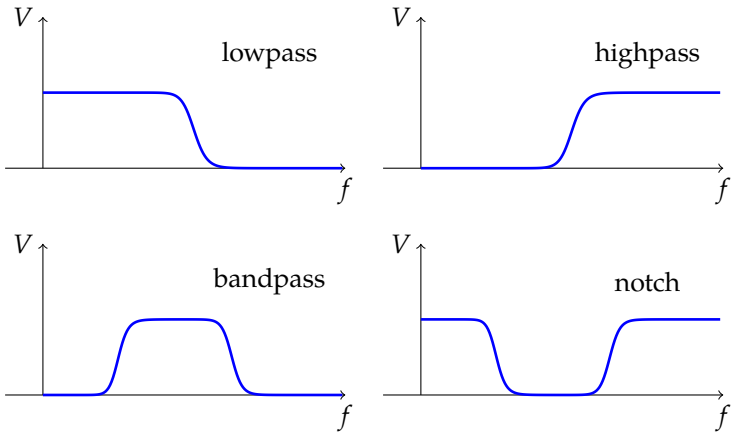


Figure 10.5: Filter nomenclature using a Butterworth filter

In practice: find coefficients (b, a) using Filter Design Toolbox. → freqz(b, a) gives the frequency response of the filter. Then apply:

1. `scipy.signal.lfilter`(b, a, x) (once) or
2. `scipy.signal.filtfilt`(b, a, x) (twice + reversal) – **zero-phase**.

This allows to construct zero-phase digital filters from filters that have a phase shift; the trick is to run it twice in opposite directions so that they phase shifts cancel each other.

Other methods

- **Wiener filter** (optimal if a known signal has been corrupted with noise of known characteristics)
- **Spline smoothing** is our personal favorite. The theory is explained in *Cook and Peters* (1981) and the code[1] uses the algorithm of *Weinert* (2009).

[1] hepta_smooth.m

Note *Bottom line:* ***never use a running mean. Everything else is ok****, but the choice depends on what you want to achieve.*

II Interpolation

General idea

The general idea can be simply seen as "connecting the dots". Given a series of datapoints taken at different t_i, we ask ourselves what happened between the t_i's [2]. The least squares method[3] could fit a curve that minimizes the squared distance to points but this is not what we want. We want to go through all the points and figure out what could plausibly have happened in between. How do we do this?

[2] Without loss of generality, this discussion also applies to spatial contexts, though in some cases more sophisticated methods (*e.g.* Kriging) must be used in multiple dimensions

[3] Interpolation can be described as "exact curve fitting". This is different from the case of "smoothed curve fitting", which would encompass least squares, spline fitting and others.

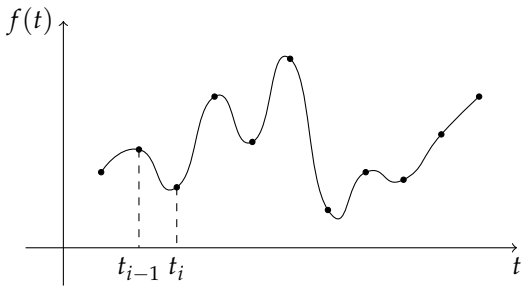


Figure 10.6: Example of interpolation of points

Linear interpolation

There are many ways to connect the dots depending on the constraints we impose. The simplest way is *linear interpolation*. This form of interpolation falls into the category of "piecewise interpolations".
On I_1 we have

$$L_1(t) = a_1 t + b_1$$

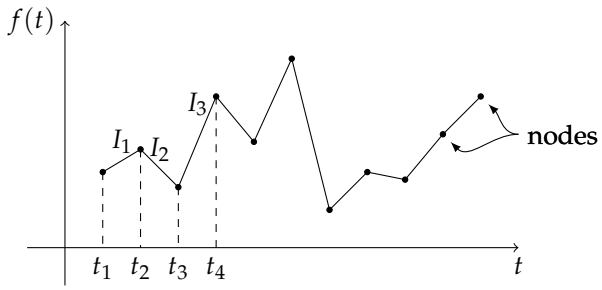


Figure 10.7: Linear interpolation

and we ask that $L_1(t_1) = f_1$ and $L_1(t_2) = f_2$ so

$$L_1(t) = f_1 + \frac{f_2 - f_1}{t_2 - t_1}(t - t_1).$$

By induction

$$L_i(t) = f_i + \frac{f_{i+1} - f_i}{t_{i+1} - t_i}(t - t_i).$$

The pros of this approach are

- is simple,
- is computationally trivial.

The cons is that the interpolating function is continuous only at the 0^{th} order, i.e. it's spikey.
Generally we want smoothness up to the second order (forces symmetry) or up to the third order.

Cubic splines

The cubic splines approach is a piecewise approach but this time each piece is a 3^{rd} degree (i.e. cubic) polynomial

$$S(x) \begin{cases} s_1(t) & t \in [t_1, t_2] \\ s_2(t) & t \in [t_2, t_3] \\ \dots & \dots \\ s_{n-1}(t) & t \in [t_{n-1}, t_n] \end{cases} \tag{10.12}$$

where

$$s_i(t) = a_i(t - t_i)^3 + b_i(t - t_i)^2 + c_i(t - t_i) + d_i$$

and we require that it's first and second derivatives are continuous in the nodes

$$s_i' = 3a_i(t - t_i)^2 + 2b_i(t - t_i) + c_i,$$
$$s_i'' = 6a_i(t - t_i) + 2b_i.$$

The cubic spline has to have four properties

1. the piecewise $s(t)$ will interpolate all points f_i,
2. $s(t) \in C^0([t_1, t_n])$,
3. $s'(t) \in C^0([t_1, t_n])$,
4. $s''(t) \in C^0([t_1, t_n])$.

So we have

1. $f_i = d_i$
2. $s_i(t_i) = s_{i-1}(t_i) \quad \forall i \in [2, n]$
   $\Rightarrow d_i = a_{i-1}(t_i - t_{i-1})^3 + b_{i-1}(t_i - t_{i-1})^2 + c_{i-1}(t_i - t_{i-1}) + d_{i-1}$ with $i \leq n-1$.
3. Let $h = t_i - t_{i-1}$. We require

$$s_i'(t_i) = s_{i-1}'(t_i)$$

   but since $s_i'(t_i) = c_i$ we have

$$c_i = 3a_{i-1}h^2 + 2b_{i-1}h + c_{i-1} \quad i \in [2, n-1]$$

4. We require

$$s_i''(t_i) = s_{i-1}''(t_i)$$

   but since $s_i''(t_i) = 2b_i$ we have

$$2b_i = 6a_{i-1}h + 2b_{i-1}.$$

Now let $M_i \equiv s_i''(t_i)$, we have $b_i = M_i/2$ and we already know that $d_i = f_i$. From (4.) we get

$$a_i = \frac{M_{i+1} - M_i}{6h}$$

while from (3.), after some tedious algebra, we get

$$c_i = \frac{f_{i+1} - f_i}{h} - \frac{(M_{i+1} - 2M_i)}{6}h.$$

Now all four coefficients are determined

$$\begin{cases} a_i = \frac{M_{i+1} - M_i}{6} \\ b_i = \frac{M_i}{2} \\ c_i = \frac{f_{i+1} - f_i}{h} - \frac{M_{i+1} + 2M_i}{6}h \\ d_i = f_i \end{cases}.$$

In order to solve this system of equations we can put the condition (3.) in matrix form

$$
\begin{aligned}
&c_{i+1} = 3a_i h^2 + 2b_i h + c_i \\
&3\left(\frac{M_{i+1} - M_i}{6}\right) h^2 + M_i h + \frac{f_{i+1} - f_i}{h} - \left(\frac{M_{i+1} + 2M_i}{6}\right) h = \\
&\frac{f_{i+2} - f_{i+1}}{h} - \left(\frac{M_{i+2} + 2M_{i+1}}{6}\right) h
\end{aligned}
$$

After some regrouping we have

$$
\forall i \in [1, n-1] \quad M_i + 4M_{i+1} - M_{i+2} = 6\left(\frac{f_i - 2f_{i+1} - f_{i+2}}{h^2}\right)
$$

leading to the matrix equation

$$
\begin{pmatrix}
1 & 4 & 1 & 0 & 0 & \dots & 0 \\
0 & 1 & 4 & 1 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & 0 & 1 & 4 & 1
\end{pmatrix}
\begin{pmatrix}
M_1 \\ M_2 \\ \dots \\ M_n
\end{pmatrix}
= \frac{6}{h^2}
\begin{pmatrix}
y_1 - 2y_2 + y_3 \\ y_2 - 2y_3 + y_4 \\ \dots \\ y_{n-2} - 2y_{n-1} + y_n
\end{pmatrix}
$$

which consists of $(n-2)$ rows and n columns. This means that the system is underdetermined so we add some *boundary conditions*. Depending on the choice of the boundary conditions we have three types of splines

- natural: $M_1 = M_n = 0$,
- parabolic runout: $M_1 = M_2$, $M_n = M_{n-1}$,
- cubic runout: $M_1 = 2M_2 - M_3$, $M_n = 2M_{n-1} - M_{n-2}$.

We can also have *periodic splines* and *clamped splines*. The most used ones are the natural cubic splines since they can extend outside the endpoints (others have crazy behaviour).
Tridiagonal system: second derivatives (curvature) is easily determined from the data $(f_i, t_{:}i)$, straightforward back-substitution.

Note *For spatial date bilinear or bicubic interpolation can be extremely useful in changing resolutions or going from an unevenly-spaced dataset to an evenly-spaced one. ESMPy*[4] *provides many routines for regridding.*

[4] https://www.earthsystemcog.org/projects/esmpy/

Lagrange interpolant

Given $(n+1)$ points, there's only one polynomial that goes through all of them. It can be written in many forms but Lagrange's is the most common and consists of a linear combination of basis polynomials

$$
L(x) = \sum_{i=0}^{n} y_i l_i(x)
$$

where $l_i(x) = \prod_{j=0, j\neq i}^{n} \left(\frac{x-x_i}{x_j-x_i}\right)$. Obviously $l_i(x_j) = \delta_{ij}$ so it goes through all the points y_i. A problem with this approach is that when the number of points gets large, so does the degree of the polynomial and in this case it is out of control. This is a global interpolant: one single function fits all the points but at the price of ginormous oscillations. For this reason this interpolant should never be used in practice. It is better to use piecewise, local interpolants: these require more coefficients (hence more computations) but they are much less sensitive to outliers since one segment is fairly insulated from remote ones.

Fourier interpolant

So far we have only considered polynomials, what if we use $\cos(\omega_i t)$, $\sin(\omega_i t)$? These can be expressed as polynomials too (remember the Euler's formula and $e^{i\omega_j k t} = \left(e^{i\omega_j t}\right)^k$). In fact Fourier analysis can be seen as a curve fitting problem with trigonometric functions. It can also be seen as an inverse problem

$$F_N = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \dots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \dots & \omega^{2(N-1)} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \omega^{2(N-1)} & \omega^{4(N-1)} & \dots & \omega^{(N-1)(N-1)} \end{pmatrix}$$

where $\omega = e^{\frac{-2\pi i}{N}}$ and F_N is a unitary matrix $F_N \bar{F}_N^\top = I$.

Part III

Living in multiple dimensions

Chapter 11

MULTIVARIATE RELATIONSHIPS

We now seek to describe dependencies among different variables. This could be variables of different nature (e.g. temperature vs pressure) or different measurements of the same quantify at different locations (*i.e.* a *field*) or in different experiments. Since the normal distribution is so omnipresent in nature, the multivariate normal will emerge as a central theme of this chapter.

I Relationships between variables

Random Vectors

Definition 14 *A random vector is a collection of random variables*

$$\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$$

A special case of a random vector is a timeseries where all observations are IID. Autoregressive models would be another example, but in this case independence is lost (neighboring values tend to resemble each other).

Joint & Marginal Distributions

A random vector is characterized by its joint distribution $f(x_1, \cdots, x_p)$. As in 1D, a multivariate PDF must verify three conditions: (1) f is continuous; (2) f is positive; (3) f has unit mass:

$$\int \cdots \int f(x_1, \cdots, x_p) dx_1 \cdots dx_p = 1$$

It gets tricky to represent in more than two dimensions, so let us focus on $p = 2$, and $f(x_1, x_2)$. An example of joint distribution between two climate variables (transient climate response and equilibrium climate sensitivity) is shown in Fig. 11.1.

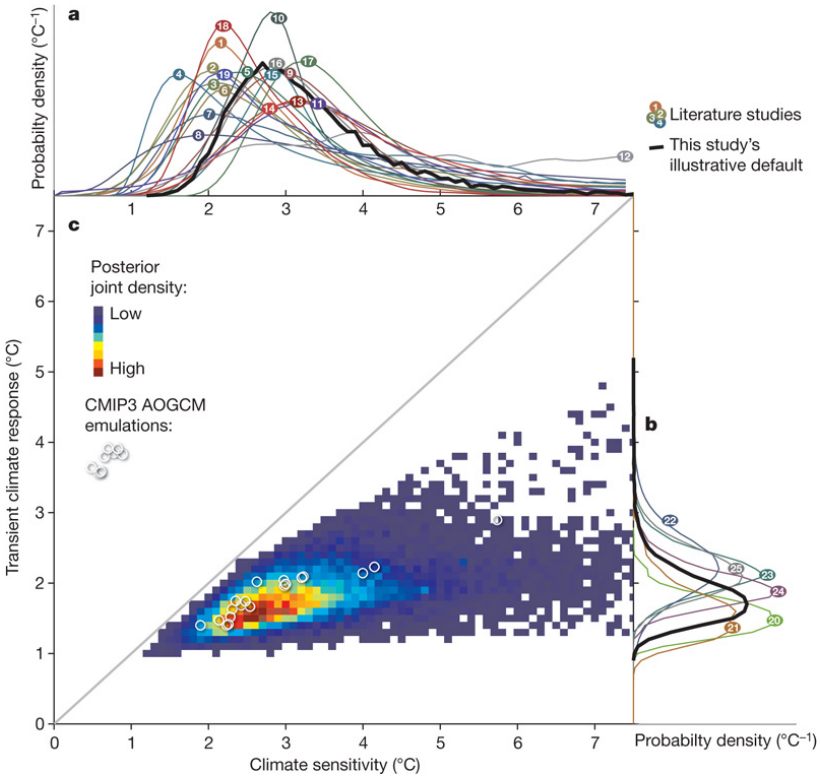


Figure 11.1: a) Marginal probability density functions (PDFs) of climate sensitivity; b), marginal PDFs of transient climate response (TCR); c), posterior joint distribution constraining model parameters to historical temperatures, ocean heat uptake and radiative forcing under representative illustrative priors. For comparison, TCR and climate sensitivities are shown in (c) for model versions that yield a close emulation of 19 CMIP3 climate models (white circles). From *Meinshausen et al.* (2009)

The joint distribution is depicted by the color field and is seen to occupy the lower right quadrant of the square of possible values. Integrating such a distribution with respect to either variable leads two *marginal distributions*. For instance:

$$f_1(x_1) = \int_{\mathbb{R}} f(x_1, x_2)\, \mathrm{d}x_2 \tag{11.1a}$$

$$f_2(x_2) = \int_{\mathbb{R}} f(x_1, x_2)\, \mathrm{d}x_1 \tag{11.1b}$$

Such marginals are shown in panels *a* and *b* of Fig. 11.1. One may think of a marginal distribution as averaging out the effect of all others variables to focus on a particular one.

Conditional Distributions

If the components of a random vector are independent, then it is easy to show that their joint distribution factorizes into marginal distributions[1]:

$$f(x_1, x_2, \cdots, x_p) = f_1(x_1) f_2(x_2) \cdots f_p(x_p) \tag{11.2}$$

[1] this is in fact the definition of independence

This is not the case in general, as shown in Fig. 11.1. In that case, it may be interesting to slice through the joint distribution at a particular value of either variable, obtaining what is known as a *conditional distribution*. In two dimensions, for instance:

$$g_2(x_1|X_2 = x_2) = f(x_1, x_2),\ x_1 \text{ variable},\ x_2 \text{ fixed} \tag{11.3a}$$
$$g_1(x_2|X_1 = x_1) = f(x_1, x_2),\ x_1 \text{ fixed},\ x_2 \text{ variable} \tag{11.3b}$$

We will return to this concept later with estimation and prediction from linear models (Chapter 15).

The covariance matrix

We saw in Chapter 3, Sect. II that covariance and correlation are useful measures of the linear association between variables. Their generalization to multiple dimensions are the *covariance matrix* and its scaled doppleganger, the *correlation matrix*. By definition, the covariance matrix associated with $\boldsymbol{X}_p$ is a $p \times p$ matrix Σ such that

$$\Sigma_{ij} = \text{Cov}(X_i, X_j) \tag{11.4}$$

Since $\text{Cov}(X, Y) = \text{Cov}(Y, X)$, the covariance matrix is *symmetric*, $\Sigma^\top = \Sigma$. Further, the diagonal elements are of the form $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$, so the matrix carries the variance of each variable along its main diagonal. For two variables, the covariance matrix would write:

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_2, X_1) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{pmatrix} \tag{11.5}$$

The correlation matrix is obtained similarly as

$$\boldsymbol{R}_{ij} = \rho(X_i, X_j) \tag{11.6}$$

Since a variable is always perfectly correlated with itself, its diagonal is made of ones.

For a random matrix to qualify as a covariance (or correlation) matrix, it must be *positive definite*[2] and symmetric. These are rather strong constraints, so this is a fairly restricted class. Positive definiteness means in particular that a covariance matrix can always be inverted[3].

[2] For any nonzero real vector $\boldsymbol{a}$, the matrix $\boldsymbol{M}$ is positive definite if and only if $\boldsymbol{a}^\top \boldsymbol{M} \boldsymbol{a} > 0$

[3] This is emphatically not true of all covariance matrices estimated from observations, especially when the number of samples n is smaller than the number of variables p. This is a common difficulty in inverse problems, which we will learn to overcome in different ways in Chapter 14

II The Multivariate Normal Distribution

Just as the normal distribution is the crown jewel of univariate distributions, the multivariate normal (MVN) is by far the most ubiquitous of multivariate distributions. This is partially because of a multivariate version of the Central Limit Theorem, but also because, just like in the univariate case, it has so many convenient properties that it is often attractive to transform multivariate data to approximate normality just so that one can use the MVN. We start with a bivariate example before generalizing to $p > 2$.

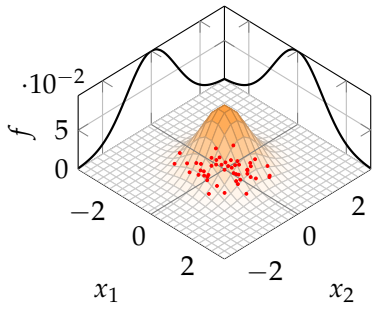


Figure 11.2: The Bivariate Normal Distribution – independent isotropic case. Black curves represent the marginal densities, while red dots represent random samples from this distribution, tending to cluster around the central bump.

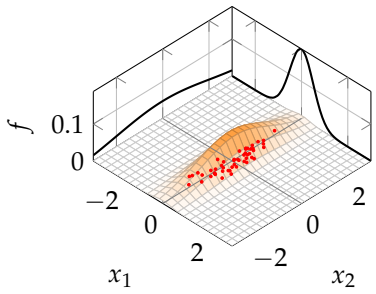


Figure 11.3: The Bivariate Normal Distribution – independent anisotropic case. As in Fig. 11.2, except that $\sigma_2 > \sigma_1$.

Easing in: The Bivariate Normal

Let (X_1, X_2) follow a *bivariate normal distribution*. Assume for now that the two variables are independent, so $f(x_1, x_2) = f_1(x_1) f_2(x_2)$. That is, if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$,

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \times \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2} \\ &= \frac{1}{2\pi}\frac{1}{\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]} \end{aligned} \tag{11.7}$$

This situation is illustrated in Fig. 11.2 & Fig. 11.3. It is, however, a rather restrictive situation: in general, X_1 and X_2 could be dependent.

General Case

Definition

To express this situation we need the tools of linear algebra. Define:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}; \quad \boldsymbol{\Sigma} = \begin{pmatrix} {\sigma_1}^2 & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & {\sigma_2}^2 & \dots & \mathrm{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \dots & {\sigma_p}^2 \end{pmatrix}$$

We say that $\boldsymbol{X_p}$ follows a p-variate normal distribution ($\boldsymbol{X_p} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$) if and only if:

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \tag{11.8}$$

where, as per standard notation, $|\Sigma|$ is the determinant of Σ (Appendix B).

Mahalanobis Distance

The quantity within the integral is $-1/2$ times the *Mahalanobis distance* between $\boldsymbol{x}$ and $\boldsymbol{\mu}$. This distance is a quadratic form in $\boldsymbol{X_p}$. It defines an *ellipsoid* with lengths $\sqrt{\Sigma_{11}}, \sqrt{\Sigma_{22}}, \cdots, \sqrt{\Sigma_{pp}}$ for the semi-major axes. You can think of it as a multivariate measure of distance (a norm) scaled by the uncertainty in each variable (Fig. 11.4). To gain intuition about the MVN, let us return to our bivariate case:

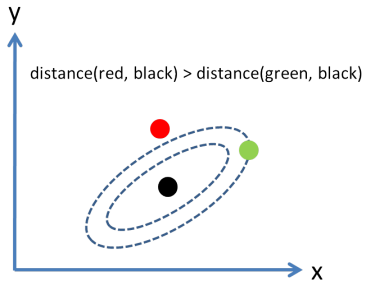


Figure 11.4: Illustration of the Mahalanobis distance. In this anisotropic case, the red dot is further than the green dot by this distance, though it is closer by the usual Euclidean distance ("as the crow flies"). Put another way, the uncertainty along the x axis is larger than the uncertainty along the y axis, so the red dot appears more distant than the green dot.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \quad \boldsymbol{\Sigma} = \begin{pmatrix} {\sigma_1}^2 & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & {\sigma_2}^2 \end{pmatrix}$$

Now, using the identity $\mathrm{Cov}(X_1, X_2) = \rho\sigma_1\sigma_2$, we get:

$$\boldsymbol{\Sigma} = \begin{pmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{pmatrix}$$

Then:

$$|\boldsymbol{\Sigma}| = {\sigma_1}^2{\sigma_2}^2(1-\rho^2)$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{|\boldsymbol{\Sigma}|}\begin{pmatrix} {\sigma_2}^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & {\sigma_1}^2 \end{pmatrix}$$

So

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]$$

Reasoning in terms of standardized variables $z_i = \frac{x_i-\mu_i}{\sigma_i}$, we get:

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \frac{{z_1}^2 - 2\rho z_1 z_2 + {z_2}^2}{1-\rho^2}$$

First note that, after all these matrix operations are said and done, we are left with a *scalar* – a single number. Second, notice that the numerator looks an awful lot like $a^2 - 2ab + b^2 = (a-b)^2$, hence the name *quadratic form* (this would hold true in higher dimensions as well). It is a normalized distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ (or between $\boldsymbol{z}$ and 0). Now, if the variables were independent[4], then $\rho = 0$ and the covariance matrix would be diagonal

[4] which we denote $X_1 \perp\!\!\!\perp X_2$

$$\Sigma = \begin{pmatrix} {\sigma_1}^2 & 0 \\ 0 & {\sigma_2}^2 \end{pmatrix} \tag{11.9}$$

In this case, it is easy to see how one would fall back on Eq. (11.7). The resulting distance measure is called a *normalized Euclidean distance*: in p dimensions,

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{p}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2} \tag{11.10}$$

Dependencies

The general case (dependent, anisotropic) of the bivariate normal distribution is shown in Fig. 11.5. This time, the axis is rotated, meaning that knowing something about X_1 tells us something about X_2 (and *vice versa*).

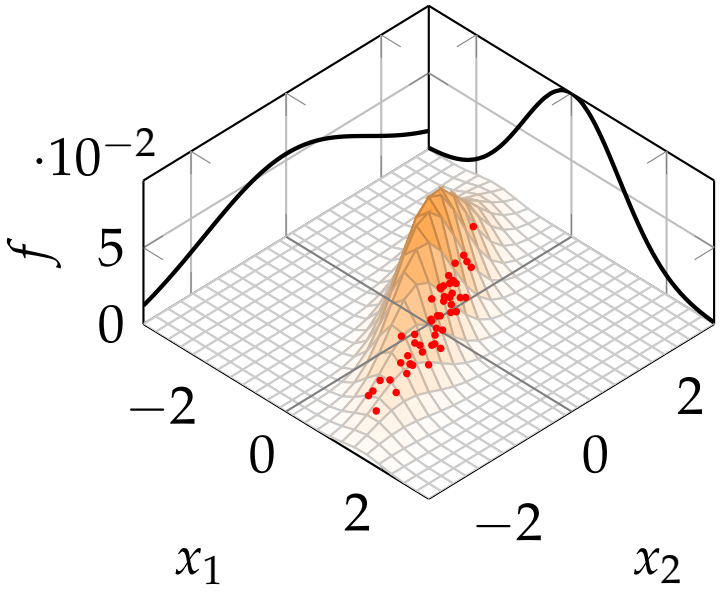


Figure 11.5: The Bivariate Normal Distribution – dependent anisotropic case. As in Fig. 11.3, now with $\rho \neq 0$.

In three dimensions, the multivariate normal distribution is a "cucumber". E.g., the case $\sigma_1 \gg \sigma_2 = \sigma_3$ is called *anisotropic independent* (Fig. 11.6), while the case in which $\sigma_{12} \neq 0$, $\sigma_{13} \neq 0$, $\sigma_{23} \neq 0$ is called *anisotropic dependent* (Fig. 11.7). In higher dimensions, the probability equisurfaces define a hyperellipsoid, which generalizes these cucumbers. The main task of principal component analysis (PCA, Chapter 12) is to identify the main axes of such ellipsoids, sometimes in very high dimensions.

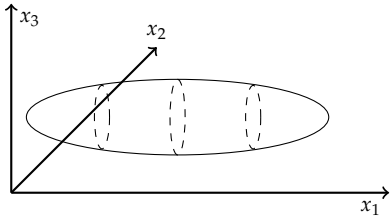


Figure 11.6: Trivariate normal density, anisotropic independent case

Properties

The MVN has four wonderful properties:

Subsets

All subsets of variables from an MVN also follow an MVN. For instance, if we split $(X_1, X_2, \ldots, X_p)$ into $\boldsymbol{X}_1 = (X_1, X_2, \ldots, X_q)$ and $\boldsymbol{X}_2 = (X_{q+1}, X_{q+2}, \ldots, X_p)$, then $\boldsymbol{X}_1$ follows a q-variate normal distribution, $\boldsymbol{X}_2$ a $(p-q)$-variate normal distribution, with parameters:

$$\boldsymbol{\mu}_1 = (\mu_1, \mu_2, \ldots, \mu_q) \tag{11.11a}$$

$$\boldsymbol{\mu}_2 = (\mu_{q+1}, \mu_{q+2}, \ldots, \mu_p) \tag{11.11b}$$

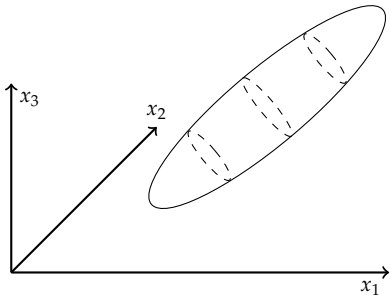


Figure 11.7: Trivariate normal density, anisotropic dependent case

and $\Sigma_{1,1}, \Sigma_{2,2}$ such that

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{1,1} & \Sigma_{1,2} \\ \hline \Sigma_{2,1} & \Sigma_{2,2} \end{array} \right) \tag{11.12}$$

Independence

You already know that independence implies zero covariance:

$$X_i \perp\!\!\!\perp X_j \Rightarrow \Sigma_{ij} = 0 \tag{11.13}$$

In the MVN world, a magical thing happens: the reverse is also true! This is because when off-diagonal elements of Σ are zero, the exponential term factorizes cleanly into distinct factors, as in Eq. (11.7). This may seem trivial, but it is a very special property that makes life incommensurately easier.

Linear combinations

Any linear combination of a multivariate normal distribution is a multivariate normal distribution as well, and the means and variances are linearly related. Specifically, if $Y = B^\top X + A \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ then $\mu_Y = B^\top \mu_X + A$ and $\Sigma_Y = B^\top \Sigma_X B$.

Conditional Distributions

Conditional distributions of a subset of the MVN, given values for the other variables, are also MVN. This means that whichever way we slice it, we always get an MVN (*e.g.* Fig. 11.8). This can be used to estimate, or predict, values of one variable given the other. For instance, using the notation of Eq. (11.11b) and Eq. (11.12), the conditional mean of $\boldsymbol{X}_1$ given that variable $\boldsymbol{X}_2$ takes the value $\boldsymbol{x}_2$ is

$$\mu_1|x_2 = \mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(x_2 - \mu_2) \tag{11.14}$$

That is, the mean is pulled in the direction of $x_2 - \mu_2$, by an intensity proportional to the projection of x_1 on x_2, scaled by the uncertainties in x_2. This is the basis for *linear regression*. The conditional covariance matrix is

$$\Sigma_{1,1}|x_2 = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1} \tag{11.15}$$

This expression does not depend on the value x_2, just on the covariance submatrices. If $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2$, $\Sigma_{1,2} = \Sigma_{2,1} = 0$ and the equations reduce to $\mu_1|x_2 = \mu_1$ and $\Sigma_{1,1}|x_2 = \Sigma_{1,1}$, which is another way of saying that no information is provided by knowledge of $\boldsymbol{X}_2$.

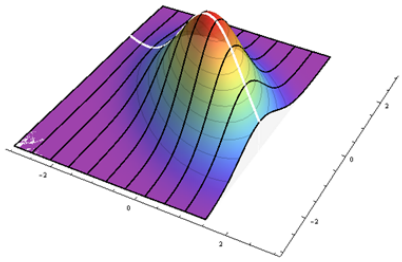


Figure 11.8: Slicing through a dependent bivariate normal, one gets a univariate normal with mean and variance dependent on value of the slicing variable.

In the bivariate case, one may rewrite these as:

$$\mu_1|x_2 = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) \tag{11.16a}$$

$$\sigma_1|x_2 = \sigma_1\sqrt{1-\rho^2} \tag{11.16b}$$

It indicates that the conditional mean is larger than the unconditional mean μ_1 if x_2 is above its mean μ_2 and $\rho > 0$, or if $x_2 < \mu_2$ and $\rho < 0$. Importantly, as long as $\rho \neq 0$, then the uncertainty about $x_1|x_2$ (measured by $\sigma_1|x_2$) is smaller than the unconditional uncertainty σ_1. In this sense, ρ^2 can be viewed as the fraction of variance in x_1 that is accounted for by x_2. These remarks also apply in the general case (Eq. (11.14) and Eq. (11.15))

Covariance Estimation

In real life, we get a random matrix $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ built out of n observations (arranged in rows) of p variables (arranged in different columns). Each column represents a Gaussian random variable. We can define the *sample mean* of each column

$$\hat{\mu}_j = \frac{1}{n}\sum_{k=1}^{n} X_{kj} = \overline{x_j} \tag{11.17}$$

and the *sample covariance matrix*

$$\widehat{\Sigma}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\Big(x_{k\,i} - \overline{x_i}\Big)\Big(x_{k\,j} - \overline{x_j}\Big). \tag{11.18}$$

We can define *centered variables* as

$$x'_{:j} = x_{:j} - \overline{x_j} \tag{11.19}$$

and rewrite the sample covariance matrix as

$$\big(\hat{\Sigma}\big)_{ij} = S_X = \frac{1}{n-1}X'^{\top}X' \tag{11.20}$$

so the sample covariance is a scaled, *inner product* of the centered data matrices X'. It is exactly analogous to the univariate case of the sample variance:

$$s_x = \frac{1}{n-1}(x-\overline{x})^{\top}(x-\overline{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2. \tag{11.21}$$

These estimators are MLE, hence optimal (consistent, efficient). In fact, since they uniquely characterize the distribution, they are called *sufficient* statistics of the MVN. However, as we shall see, things can go sour when $p > n$ – the sample covariance matrix no longer is positive definite, and will have to be regularized in order to work properly. In fact, this may even be the case for $n \gtrsim p$ (that is, for sample sizes not much larger than the number of parameters to be estimated (*i.e.* the number of random vectors))

Multivariate Central Limit Theorem

As in the univariate case, the MVN enjoys the status of gravitational attractor of the space of distributions. If $X_1, \cdots, X_p$ are IID with mean $\boldsymbol{\mu}$ and covariance Σ, then their sample mean over n observations $\overline{\mathbf{X}}_n$ converges in distribution to a MVN with mean $\boldsymbol{\mu}$ and covariance matrix $\frac{1}{n}\Sigma$.

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right) \xrightarrow{D} \mathcal{N}_p(0, \Sigma) \tag{11.22}$$

Multinormality for the sample mean implies that the sampling distribution of the Mahalanobis distance between the true and sample means asymptotically follows a χ^2 distribution with p degrees of freedom:

$$\left(\overline{\mathbf{X}}_n - \mu\right)^\top \frac{1}{n}\Sigma^{-1}\left(\overline{\mathbf{X}}_n - \mu\right) \sim \chi^2_p \tag{11.23}$$

This allows to draw confidence ellipsoids around the estimated mean, within the uncertainties of the sampled observations. When Σ is estimated from observations, these confidence regions get a little harder to compute, requiring the use of the F distribution.

Chapter 12

PRINCIPAL COMPONENT ANALYSIS

> It is not easy to arrive at a conception of a whole which is constructed from parts belonging to different dimensions
>
> Paul Klee, *On Modern Art*

Motivation: given a field $\psi(x, y, t)$ we ask ourselves if there are certain spatio-temporal patterns (or *modes*) that account for a large portion of the observed variability. If so, we can more efficiently describe variations in the dataset, and use these patterns as a new reference frame in which to view the data.

I Principal component analysis: theory

The Big Idea

Imagine that one collects observations of a multivariate normal random variable, e.g. $T(x, y, t)$, $SLP(x, y, t)$. We want to find the patterns that describe the most of the covariance found in the dataset.
One may label all locations from 1 to p so

$$T(x, y, t) \longrightarrow T(s, t). \tag{12.1}$$

If we put these observations in an $n \times p$ matrix X (n samples, p locations) and we center the columns $X' = X - \bar{X}$, then $S_X = \frac{1}{n-1} X'^{\top} X'$ is a real, symmetric matrix. *Principal component analysis* aims to find the major axes of the ellipsoid spanned by $(\mathbf{x} - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$, i.e. the directions of maximum variation.

To do this consider an eigendecomposition (Appendix D) of S_X

$$S_X = E \Lambda E^{\top} \tag{12.2}$$

which is possible because S_X is real and symmetric.

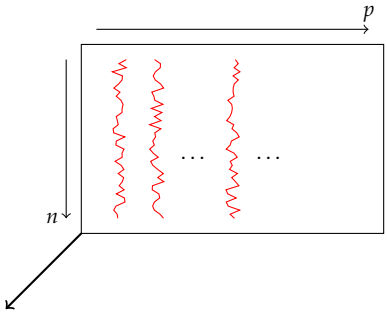


Figure 12.1: Example of data matrix

Moreover, the eigenvalues will all be non-negative. The matrix Λ has the form

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 & \dots & 0 \\ 0 & \Lambda_2 & \dots & 0 \\ \dots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \Lambda_M \end{pmatrix} \tag{12.3}$$

where $M = \min(p, n)$. Assuming that $p \leq n$, E is given by

$$E = \begin{pmatrix} e_1 & e_2 & \dots & e_p \end{pmatrix} \tag{12.4}$$

where the e_i's are the eigenvectors of S_X. The eigenvalue λ_i can be considered as the variance accounted for by the pattern defined by the corresponding eigenvector e_i. It may be charted on a *scree plot* (Fig. 12.2), along with estimates of uncertainty (12.14).

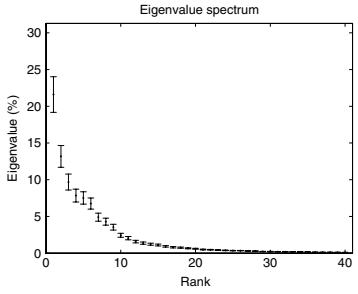


Figure 12.2: A scree plot, showing the fraction of variance accounted for by each eigenvalue. Vertical bars show approximate 95% confidence limits given by the *North et al.* (1982) rule of thumb. From *Hannachi et al.* (2007)

The temporal signature associated with mode i is given by the inner product $u_i(t) = e_i^\top X$ and defines the i^{th} *principal component*, PC_i. Note that it is only a function of the time variable (or whichever variable serves as the row index in your application).

Singular value decomposition formulation

Recall that for any matrix A, we can write $A = U\Sigma V^\top \Rightarrow A^\top A = V\Sigma^2 V^\top$ (Appendix D). Now consider the case $A = \widetilde{X} = \frac{1}{\sqrt{n-1}}(X - \overline{X})$ so that $S_X = \widetilde{X}^\top \widetilde{X}$ is the sample covariance matrix of the dataset. In this case

$$\widetilde{X} = U\,\mathrm{diag}(\sigma)\,V^\top = U\,\mathrm{diag}(\sigma)\,E^\top \Leftrightarrow \widetilde{X}^\top \widetilde{X} = E\Lambda E^\top. \tag{12.5}$$

where $\mathrm{diag}(\sigma)$ is given by

$$\mathrm{diag}(\sigma) = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \sigma_p \end{pmatrix} \tag{12.6}$$

So for each mode i

- e_i represents the spatial pattern
- $\sigma_i = \sqrt{\lambda_i} \leftrightarrow$ standard deviation associated with e_i.
- $u_i(t)$ is the temporal pattern.

The two formulations are equivalent, but SVD tends to be a more efficient algorithm than eigendecomposition (*Golub and van Loan*, 1993).

Orthogonality

The matrix U of left singular vectors therefore gathers the principal components of the field. From the definition of the singular value decomposition $E^\top E = I_p$ and

$$S_U = \mathrm{Var}(E^\top \widetilde{X}) = E^\top \mathrm{Var}(\widetilde{X}) E = E^\top S_X E = \Lambda. \tag{12.7}$$

The principal components are also orthogonal and, given an appropriate scaling $1/\sqrt{\lambda_i}$, can be made orthonormal. Hence, principal component analysis is a *bi-orthogonal* operation. From a multivariate normal dataset we can define U, $\sqrt{\lambda_i}$ and E from which we derive a decomposition on a set of biorthogonal spatio-temporal patterns:

$$X(s,t) = \sum_{i=1}^{M} \sqrt{\lambda_i}\, u_i(t) \times e_i(s). \tag{12.8}$$

We can see that space and time have been separated, each mode is weighted by the square root of its contribution to the total variance. Thanks to orthogonality, we readily obtain the *fraction of variance* associated with each pattern:

$$F_i = \frac{\lambda_i}{\sum_{i=1}^{M} \lambda_i} = \frac{\sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2} \tag{12.9}$$

We usually sort the eigenvalues in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and all of them are positive, so F_i increases with i.

Interpretations

Directions Principal component analysis finds the main directions of variability which can be identified with the main axes of an ellipsoid. In the case $p = 2$ we have an ellipse (Fig. 12.3),

$$H^2 = \left(\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_2}{\sigma_2}\right)^2 \tag{12.10}$$

When $p = 3$ the surface defines an ellipsoid (Fig. 12.4), while in the case $p > 3$ we have a *hyper-ellipsoid* (which our perceptual system cannot fathom without projection).

Decomposition on orthogonal functions

these functions are like the sines and cosines found in the Fourier transform but are dictated by the data, so they are *empirical* orthogonal functions[1]. The following terminology is used in the atmospheric/oceanic sciences:

$$\begin{aligned} U_i(t) &= PC_i(t) \\ E_i(s) &= EOF_i(s) \end{aligned} \tag{12.11}$$

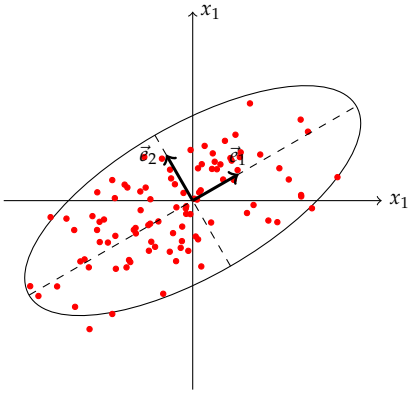


Figure 12.3: PCA aims to identify the main axes of an ellipse from noisy data

[1] The jargon of PCA is unnecessarily murky, due in no small part to climatologists taking the work of *Lorenz* (1956) too literally. Lorenz pioneered the use of PCA in climate research, under the name empirical orthogonal functions, or EOF, analysis. The name has stuck, but it makes conversations with a statistician needlessly complicated, because to them the PCs are the right singular vectors, not the left singular vectors. Know your audience!

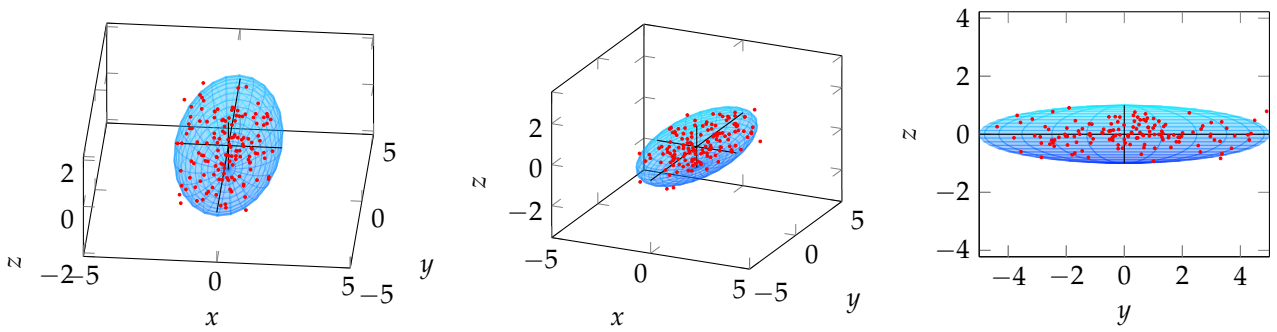


Figure 12.4: A three dimensional ellipsoid representing the probability space occupied by the data points in red. The major axes, aka principal components, (black) delineate the main directions of variability. Three viewing angles are offered here to illustrate the difficulty of visualizing a three-dimensional object on a two dimensional page.

Reformulation

Principal components are linear combinations of the data that efficiently *describe* their variations; by themselves, they do not *explain* anything, but they may highlight important modes, some of which may have a scientific interpretation (*e.g.* Sect. III).

PC compression

Because PCA is fundamentally a singular value decomposition, it allows to re-express the field as a sum of rank-one matrices[2].

$$\widetilde{X}_K = \sum_{i=1}^{K} \sigma_i u_i e_i^\top \tag{12.12}$$

If $K = M$ then we have a complete recovery (analysis $\Leftrightarrow$ synthesis); all we have done is reformulating the original observations in a different reference frame, but no information was created or lost. If $K \ll M$ but, for example, $\sum_{i=1}^{K} F_i \approx 90\%$, then we can describe 90% of the variance with just a few modes. It is so because most eigenvalue spectra decay relatively fast (*e.g.* Fig. 12.2).

[2] This is a special case of the Eckhardt-Young-Mirsky theorem (Appendix D)

II Principal component analysis in practice

Pre-processing

- Centering: $X' = X - \overline{X}$. Failing to do so will result in nonsensical results.
- Scaling (normalization): $X'' = S_X^{-1/2} X'$ amounts to working with the correlation rather than the covariance matrix.
- Area weighting: it is common to multiply the field by $\sqrt{\cos\phi}$ because the area of a surface element on the sphere is $dA = R^2 \cos\phi \; d\phi \; d\lambda$ [3]. On a uniform grid, this transformation accounts for the area of each grid box, thus preserving the energy (variance) of the field on the sphere.

[3] On a sphere, the integral of some field $\psi(\lambda, \phi)$ over a domain $\mathcal{D}$ is $\iint_{\mathcal{D}} \psi(\lambda, \phi) R^2 \cos\phi \;\; d\phi$, where λ is the longitude, ϕ the latitude, and R the radius of the sphere.

Dimensionality

If $n < p$ then S_X is singular, so its inverse is undefined. This situation may be overcome via *sparse PCA* (e.g. *Zou et al.*, 2006; *Johnstone and Lu*, 2009).

Truncation

The choice of truncation determines the amount of compression achieved. Heuristically, we may choose the first K patterns that account for a sizable portion of the variance, *e.g.*

$$\tilde{r}_K = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \simeq 0.7 \text{ or } 0.9 \tag{12.13}$$

Other criteria exist, in particular Kaiser's (retain all modes that account for more variance than the average eigenvalue) or more elaborate rules that hunt for breaks in the eigenvalue spectrum using decision theoretic criteria (*Wax and Kailath*, 1985; *Nadakuditi and Edelman*, 2008).

Separability

The *North et al.* (1982) rule of thumb assesses the separation between two successive eigenvalues as

$$\delta\lambda_i = \lambda_i \sqrt{\frac{2}{n-1}} \tag{12.14}$$

which is related to the uncertainty about λ_i. Indeed, a 95% confidence interval for λ_i under the null hypothesis of an uncorrelated random field (*i.e.* white noise) is

$$\lambda_i \left(1 \pm \sqrt{\frac{2}{n-1}} \right) \tag{12.15}$$

Hence, the smaller λ_i, the more difficult it is to separate it from its neighbors, and from noise.

Significance

Not all modes are statistically significant. Fewer still are physically meaningful. To find the meaningful ones we can use "Preisendorfer's Rule N" (*Overland and Preisendorfer*, 1982), which tests against the hypothesis that the eigenvalues λ_i arise from a random Gaussian process; any $\lambda_i \geq \lambda_{threshold}$ is retained. A more sophisticated test is against multivariate red noise, as implemented by *Dommenget* (2007).

If the sample size is small, one true physical "mode" may be spread over several of the statistical modes that we call principal components. To some extent rotation can correct for this, but it brings challenges of its own (Lab 9).

III Geoscientific uses of PCA

The use of principal component analysis in many fields of applied research, including the geosciences, runs deep. Here we give but three examples of its use in climate research, but examples abound in paleobiology, mineralogy, and others. Following the usual jargon of the atmospheric sciences, EOF and PCA will be used interchangeably.

ENSO dynamics

In a recent study, *Takahashi et al.* (2011) used PCA to identify a state space in which to describe the evolution of the El Niño-Southern Oscillation phenomenon. They performed PCA on monthly sea-surface temperature (SST) data and found that the first two PCs combined account for most of the variance (68% and 14%, respectively) in the domain. The associated spatial patterns (EOFs) are shown in Fig. 12.5.

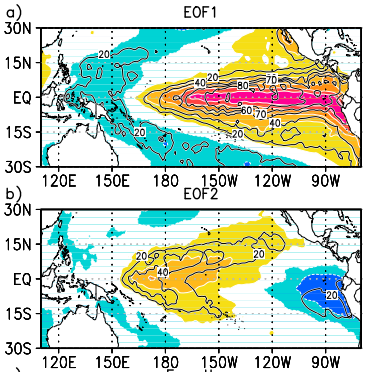


Figure 12.5: EOF patterns (°C, shading) between the 1870–2010 HadISST sea surface temperature anomalies associated with (a) PC1, (b) PC2. The percentage of explained variance is contoured (the interval is 20% (10%) below (above) 60%).From *Takahashi et al.* (2011).

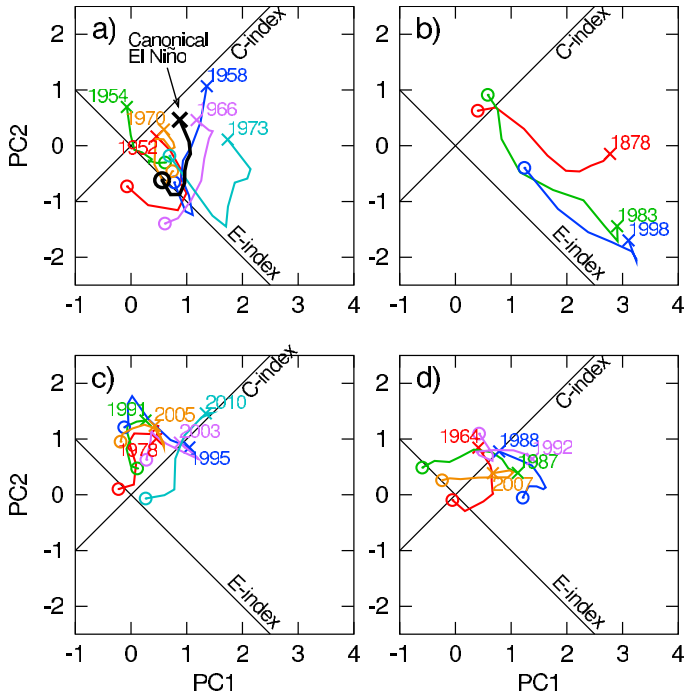


Figure 12.6: Evolution of PC1 and PC2 from May (indicated with circles) to the following January (crosses, corresponding year shown) El Niño events: (a) events considered by *Rasmussen and Carpenter* (1982) (their composite is shown thicker); (b) extraordinary events; (c) central Pacific events (*Kug et al.*, 2009) including the recent 2009–10 event (*Lee and McPhaden*, 2010); (d) other moderate events since 1950 according to NOAA, corresponding to DJF Oceanic Nino Index ≥ 1°C (see `http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml` for details). From *Takahashi et al.* (2011).

They used this as a new coordinate system, enabling the identification of two ENSO regimes: the regime of extraordinary warm events (the 1982–83, 1997–98 and 1877–78 events) and the regime that includes the cold, neutral, and moderately warm years (Fig. 12.6). This PCA-based decomposition has durably shifted the characterization of ENSO events, and exemplifies how PCA can be used to generate dynamical insight.

Patterns in age-uncertain coral data

PCA may also be applied to data matrices that are not associated with uniformly spaced data. For instance, *Emile-Geay and Eshleman* (2013) used it to identify the main mode of variability in a network of 25 coral $\delta^{18}O$ records from the tropical Indo-Pacific ocean (Fig. 12.7).

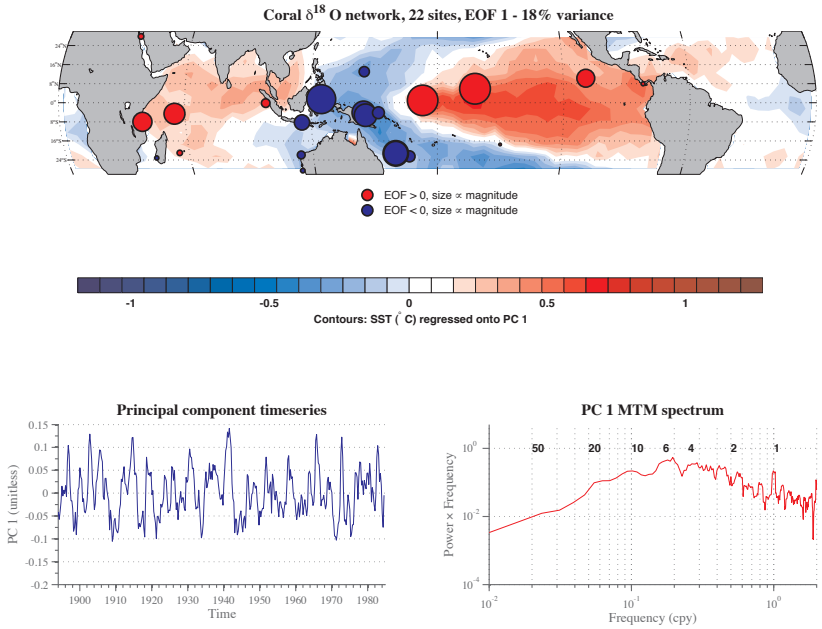


Figure 12.7: First EOF mode of the network 25 coral $\delta^{18}O$ records. (top) EOF coefficients (blue < 0, red > 0) overlain on a map of HadSST2 DJF temperature regressed onto the first principal component. The center each dot is collocated with a coral reef, and their area is proportional to the EOF loading. $\delta^{18}O$ values were multiplied by -1 so that positive excursions correspond to warming temperature; (bottom left) timeseries of the first principal component; (bottom right) Multi-taper spectrum (*Thomson*, 1982) of the first principal component; note that the y-axis is the product of the power by the frequency so that the relative area under the spectrum is preserved in this logarithmic scale. Numbers refers to the period of oscillation in years. From *Emile-Geay and Eshleman* (2013)

As expected, the spatial pattern of sea-surface temperature (SST) associated with this mode bears a strong resemblance to El Niño-Southern Oscillation (ENSO), as confirmed by its temporal expression (PC1), which displays maxima and minima coincident with known ENSO events. The MTM spectrum of the mode reveals a relatively strong annual component and dominant interannual variability, consistent with what is independently known about ENSO (e.g. *Sarachik and Cane*, 2010). The relative area covered by each circle illustrates the magnitude of the EOF coefficients ("loadings"), while the color refers to their sign (after multiplying $\delta^{18}O$ values by -1 so that negative excursions in oxygen isotopes correspond to positive temperature anomalies). Their signs are consistent with the ENSO thermal signal expected in each isotopic record, though the disparate amplitudes reflect the varying influence of other factors (climatic or not). A question arising from this work is the extent to which the EOF modes are affected by dating uncertainties in these layer-counted records. *Comboul et al.* (2014) investigated this question with a perfectly known "pseudocoral" network. Age perturbations were introduced according to a Poisson probability model, with a 5%

chance of miscounting each year (*i.e.* every 100 years, one expects $\pm 5y$ offsets due to age errors). They performed a "Monte Carlo EOF analysis" (*Anchukaitis and Tierney*, 2013) on this time-uncertain ensemble and used it to compute the relevant statistics.

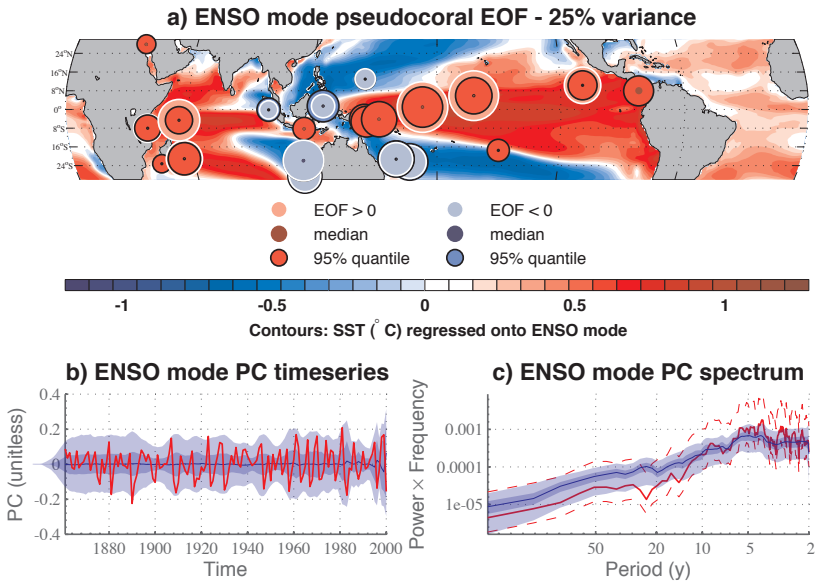


Figure 12.8: Spatiotemporal uncertainty quantification on a pseudocoral network. (a) EOF loadings (circles) corresponding to the ENSO mode of an ensemble of age-perturbed pseudocoral records with miscounting rate $\theta = 0.05$. EOF loadings for error-free data are shown in light colors circled in white, while the median and 95% quantile are shown by dark disks and black-circled disks, respectively. Contours depict the SST field associated with the mode's principal component PC (panel b), whose power spectrum is shown in (c). Results for the time-uncertain ensemble are shown in blue: median (solid line), 95% confidence interval (light-filled area) and interquartile range [25%-75%] (dark-filled area). Results for the original (error-free) dataset are depicted by solid red lines. Dashed red lines denote χ^2 error estimates for the MTM spectrum of the error-free dataset. From *Comboul et al.* (2014)

The results are shown in Fig. 12.8, and show that time uncertainty may greatly alter the spatial expression of interannual variability. They also result in a transfer of power from high frequencies to low frequencies, which suggests that age uncertainties are a plausible cause of the observed enhanced decadal variability in coral networks (*Ault et al.*, 2009).

Patterns of ocean-atmosphere covariability

Maximum Covariance Analysis is a close cousin of PCA's. As the name implies, it seeks to maximize the covariance between different fields or variables. Its mathematics are reviewed in *Bretherton et al.* (1992).

To get a feel for the method, let us apply it to the diagnosis of present day teleconnections, the study of climate relationships over wide distances. Let us use the ERSST dataset for sea-surface temperature (*Reynolds and Smith*, 1994) and NCEP/NCAR Reanalysis dataset (*Kalnay*, 1996) for the upper atmosphere geopotential height (a measure of atmospheric circulation). The SST data was limited to the tropical domain [30S,30N], while the geopotential height data includes only the Northern Hemisphere. The season considered is Northern Hemisphere winter (DJF). The variables were suitably transformed into anomalies and multiplied by the square-root of the cosine of latitude, in order for variance integrals to be representative of geometric areas (*North et al.*, 1982).

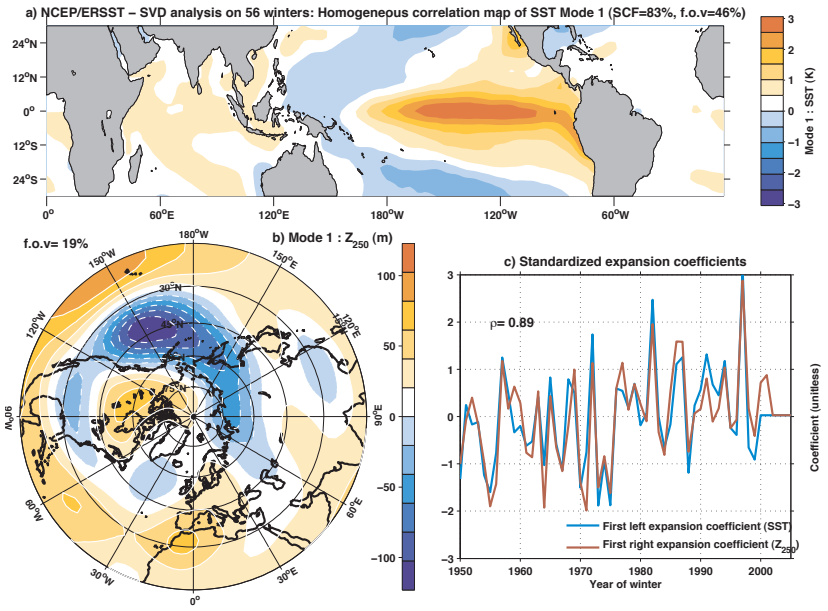


Figure 12.9: Maximum Covariance Analysis of Present Day Teleconnections from the tropical Oceans. Mode 1 a) SST pattern (left singular vector) ; b) Geopotential height field (Z250) (right singular vector) c) Expansion coefficients (normalized). From *Emile-Geay* (2006, Chap 5)

The first mode is displayed in Fig. 12.9. It accounts for an overwhelming fraction of the covariance (83%). Panel a) shows its very distinctive El Niño sea-surface temperature (SST) pattern in accounting for about half of all SST variability in the domain. The associated heterogeneous correlation map for Z_{250} (Panel B) displays the familiar Pacific North American (PNA) pattern (*Wallace and Gutzler*, 1981) that has long been recognized as the main ENSO teleconnection pattern. The pattern explains only 19% of the total variability in Z_{250}, meaning that more than 81% is due to other factors (namely, atmospheric dynamics not simply related to SST). In other words, although the pattern accounts for 83% of the covariance between the two variables, it only accounts for a limited amount of the total atmospheric variability.

This type of analysis can cleanly identify patterns of covariability between datasets. A close cousin of it is Canonical Correlation Analysis (CCA), which works on maximizing correlation rather than covariance.

Further reading

For a more thorough introduction to PCA and EOF analysis, the reader is referred to *Hannachi et al.* (2007) and *Wilks* (2011, chap 12). For an introduction to maximum covariance analysis and canonical correlation analysis, read *Bretherton et al.* (1992) and *Wilks* (2011, chap 13).

Chapter 13

LEAST SQUARES

Motivation: find relationships among noisy data.
For example, consider a free falling body whose position is given by

$$z(t) = \frac{1}{2}gt^2 + v_0 t + z_0. \tag{13.1}$$

Measuring (z_i, t_i) many times should give us good estimates of (g, v_0, z_0). To simplify the problem, assume that initial speed and position are 0: $(v_0, z_0) = (0, 0)$. So we would have

$$\begin{cases} z_1 = \frac{1}{2}gt_1^2 \\ z_2 = \frac{1}{2}gt_2^2 \\ \dots \\ z_n = \frac{1}{2}gt_n^2 \end{cases} \tag{13.2}$$

How accurately can we estimate the gravitational acceleration g?

Very often one casts this type of problem (parameter estimation) as fitting a line through a cloud of points. It's often easier to reparametarize the problem so that lines are straight (in this case, working with the variable t^2 as opposed to t). The goal of the famed *least squares* method is to find the straight line that best fits the data.

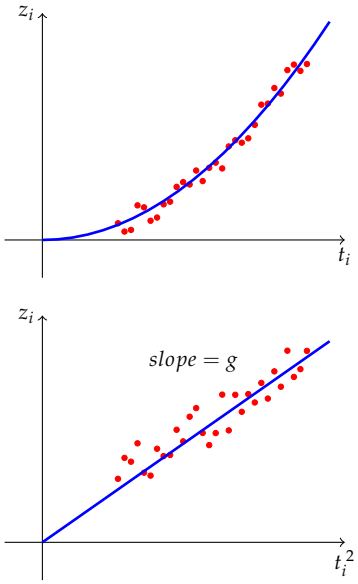


Figure 13.1: Example of data for a free falling body

I Ordinary Least Squares

Straight line fit

Assume the data $\boldsymbol{y}$ are related to some independent variable $\boldsymbol{x}$ via the model $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ such that $y_i = \beta_0 + \beta_1 x_i$.
In practice, the measurements have errors, so

$$y_i = \beta_0 + \beta_1 x_i + e_i = \hat{y}_i + e_i \tag{13.3}$$

and this expression can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{13.4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ and

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{13.5}$$

is the design matrix[1]. It is a $n \times 2$ matrix, so it's not invertible. How do we find $\boldsymbol{\beta}$? We seek the solution that *minimizes the mean squared error*:

[1] The design matrix is often called G in inverse theory. The following notations are equivalent:

$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (statistics)
$\boldsymbol{d} = G\boldsymbol{m} + \boldsymbol{e}$ (geophysics).

$$\text{MSE} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}E^2 \tag{13.6}$$

$$e_i = y_i - (\beta_0 + \beta_1 x_i) = y_i - \hat{y}_i \tag{13.7}$$

$$e_i = y_i - \sum_{j=1}^{2} X_{ij}\beta_{j-1}. \tag{13.8}$$

The total error E^2 is the sum of the individual errors:

$$E^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{2} X_{ij}\beta_{j-1}\right)^2 \tag{13.9}$$

that must be minimized. For the case of p parameters β_k, with $k = (0, 1, \ldots, p-1)$, we would set

$$\begin{aligned} &\frac{\partial E^2}{\partial \beta_k} = 0 \quad \Leftrightarrow 2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} X_{ij}\beta_{j-1}\right)(-X_{ik}) = 0 \Leftrightarrow \\ &\sum_{i=1}^{n}\underbrace{X_{ik}}_{X_{ki}^\top} y_i = \sum_{i=1}^{n}\sum_{j=1}^{p} X_{ij}X_{ik}\beta_{j-1} \Leftrightarrow \sum_{i=1}^{n} X_{ik}y_i = \sum_{j=1}^{p}\underbrace{\left(\sum_{i=1}^{n} X_{ij}X_{ik}\right)}_{X^\top X}\beta_{j-1} \end{aligned} \tag{13.10}$$

Thus,

$$X^\top \boldsymbol{y} = (X^\top X)\boldsymbol{\beta} \tag{13.11}$$

This is called a normal equation. The matrix $(X^\top X)$ is square, real and symmetric, therefore it is positive semi-definite. That is, provided $n > p$, $(X^\top X)$ always has an inverse (it has no zero eigenvalues)[2]. Therefore the solution is:

[2] What if its eigenvalues are close to, but not exactly equal to, zero, you ask? We will get to that in a bit.

$$\boxed{\boldsymbol{\beta}_{\text{OLS}} = (X^\top X)^{-1}X^\top \boldsymbol{y}} \tag{13.12}$$

The case $n = p$ is a special case: X is square and $\boldsymbol{\beta} = X^{-1}\boldsymbol{y}$.

Estimation of coefficients

In the previous case where the design matrix only involves x_i's

We have

$$X^\top X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i {x_i}^2 \end{pmatrix}.$$

Using the usual rules (Appendix B), its inverse is:

$$\begin{aligned}(X^\top X)^{-1} =& \frac{1}{n\sum_i {x_i}^2 - (\sum_i x_i)^2} \begin{pmatrix} \sum_i {x_i}^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} \\ =& \frac{1}{n^2 {S_X}^2} \begin{pmatrix} \sum_i {x_i}^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}\end{aligned}$$

Where ${S_x}^2$ is the sample variance of dataset $\boldsymbol{x}$. Further,

$$X^\top \boldsymbol{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}.$$

The slope β_1 is given by

$$\begin{aligned}\beta_1 &= \frac{1}{n^2 {S_x}^2} \left(n \sum_i x_i y_i - \sum_i x_i \sum_j y_j \right) = \frac{\frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_i (x_i - \overline{x})^2} \\ &= \frac{\widehat{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}}{s_x^2} = \hat{\rho}_{xy} \frac{s_y}{s_x}. \end{aligned} \tag{13.13}$$

So in this simple case, the least squares slope is proportional to the sample linear correlation coefficient, scaled by the ratio of sample standard deviations. Since the correlation coefficient is unitless, this factor ensures that the units of y get correctly mapped to the units of x[3].

For the intercept we have

$$\beta_0 = \overline{y} - \beta_1 \overline{x}. \tag{13.14}$$

Which depends on the slope. If any outliers or noise bias the estimate of the slope, this will affect the estimate of the intercept too.

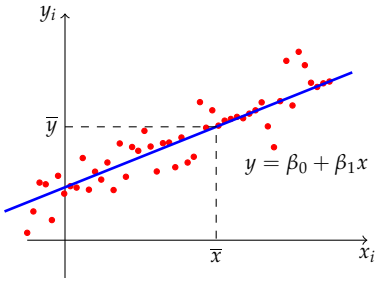


Figure 13.2: Example of linear fit

[3] One useful cross-check that you did the math right, thus, is that β_1 should be in units of y divided by units of x

II Geometric interpretation

We just obtained the OLS solution by straightforward algebra after formulating a minimization principle. There is more to it. What we were really doing is finding an approximate solution to the problem

$$Ax \approx b$$

where in this case $A = \boldsymbol{X}, x = \boldsymbol{\beta}, b = \boldsymbol{y}$. More precisely, we seek a vector b_p that is the projection of b into the range of matrix A – the vector space spanned by the columns of A, vectors $\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_p}$ (Appendix B, section VI). In other words, we seek the solution to

$$Ax = b_p + e$$

, where e is a residual that must be (1) as small as possible; (2) orthogonal to b_p. Why the latter condition? If it weren't orthogonal, then it would project onto the columns of A, thus contribute to b_p. If we put these words into equations, it means $A^\top e = \mathbf{0}$, which implies:

$$A^\top (b - Ax) = \mathbf{0} \tag{13.15}$$

Thus:

$$\boxed{A^\top A x = A^\top b}$$

If you substitute A, x and b, for their definitions, you get back the OLS normal equation (Eq. (13.11)). This intuition is purely geometric: it's about finding the best approximation to the data vector y in the space spanned by the design matrix X.

III Statistical interpretation

As in most things, there is also a statistical interpretation.

Maximum Likelihood Estimator

Assume $e_i \sim \mathcal{N}(0, {\sigma_i}^2)$ are IID errors and further assume that $\forall i, {\sigma_i}^2 = \sigma^2$. Then the most likely model $\boldsymbol{\beta}$ is the one that maximizes the likelihood

$$L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} = e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \tag{13.16}$$

so we define

$$-\log L = \frac{1}{2}\sum_{i=1}^{n} {e_i}^2 = \chi^2$$

which is called *misfit function*. Minimizing the errors (Eq. (13.9)) is equivalent, in the Gaussian[4] context, to maximizing the likelihood. Given the

[4] Note that only the errors $e_i = \hat{y}_i - y_i$ need be normal; x_i and y_i can have whatever distribution they please.

data, these are the most likely parameters to fit a straight line through the cloud of points. Further more, the OLS estimate is now imbued with all the privileges that come with ML status: it's consistent, and it's got the lowest MSE.

Uncertainties in the parameters

$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ follows the distribution $\mathcal{N}(\mu_0, {\sigma_0}^2)$ with

$$\mu_0 = \beta_0$$
$$\sigma_0 = s_e \sqrt{\frac{\sum_{i=1}^n {x_i}^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}} \quad \text{with } s_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^n {e_i}^2} \text{ (MSE).}$$

$\hat{\beta}_1$ follows the distribution $\mathcal{N}(\mu_1, {\sigma_1}^2)$ with

$$\mu_1 = \beta_1$$
$$\sigma_1 = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

Estimates are unbiased and precision depends on the MSE, as well as the variance of x. The more variable x, the better. So, the experiment should be designed in order to cover a broad dynamic range (Fig. 13.3).

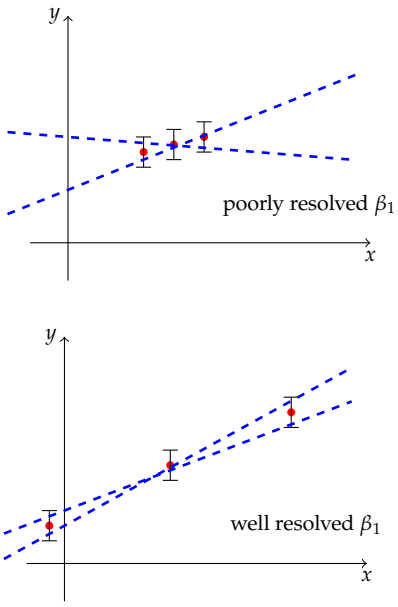


Figure 13.3: Example of linear fit for different spread in x

Relationship

$$r_{\beta_0,\beta_1} = \frac{-\overline{x}}{\frac{1}{n}\sqrt{\sum_{i=1}^n {x_i}^2}} \tag{13.17}$$

The two estimates are anti-correlated for $\overline{x} > 0$, uncorrelated for $\overline{x} = 0$, and correlated for $\overline{x} < 0$.

Trivial case

Imagine that we get no measurements x_i, so $X = \mathbb{1}_n$:

$$X\boldsymbol{\beta} = \boldsymbol{y} \Leftrightarrow \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{13.18}$$

The least squares estimate should yield the same parameter n times. Indeed, $X^\top X = n$ and $(X^\top X)^{-1} = 1/n$ so

$$(X^\top X)^{-1} X^\top \boldsymbol{y} = \frac{1}{n} \sum_i y_i = \overline{y} \tag{13.19}$$

The model's covariance matrix writes:

$$C_\beta = {\sigma_y}^2 (X^\top X)^{-1} = \frac{1}{n} {\sigma_y}^2 \quad \Rightarrow \quad \sigma_\beta = \frac{\sigma_y}{\sqrt{n}} \tag{13.20}$$

which is essentially a paraphrase of the Central Limit Theorem. The least squares estimate is both intuitive and consistent with probability theory, as it should be.

IV Exotic Least Squares

Weighted Least Squares

What if the errors are not IID? Specifically, what if some measurements are more precise than others, so $\sigma_1 \neq \sigma_2 \cdots \neq \sigma_n$. In general, for any matrix H, one may write:

$$\boldsymbol{x}' = H\boldsymbol{x} \quad \Rightarrow \quad C_{\boldsymbol{x}'} = HC_{\boldsymbol{x}}H^\top.$$

with $C_{\boldsymbol{x}} = V\Lambda V^\top$. A special case is when

$$\boldsymbol{x}' = \Lambda^{-1/2}V^\top\boldsymbol{x}$$

then

$$C_{\boldsymbol{x}'} = I_p$$

that is we have *univariant data*.
This is a good idea if we have different units or instrument precisions. So

$$E^2 = e^\top V\Lambda^{-1/2}\Lambda^{-1/2}V^\top e = \underline{e}^\top {C_e}^{-1}\underline{e}$$

where $\underline{e} = \boldsymbol{y} - X\boldsymbol{\beta}$. Thus, the error term is now a true Mahalanobis distance, a quadratic form using measurement precision as a weight. Expanding this term:

$$\begin{aligned} E^2 =& (\boldsymbol{y} - X\boldsymbol{\beta})^\top {C_e}^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^\top {C_e}^{-1}\boldsymbol{y} - 2\boldsymbol{y}^\top {C_e}^{-1}X\boldsymbol{\beta} + \\ & \boldsymbol{\beta}^\top X^\top {C_e}^{-1}X\boldsymbol{\beta} \end{aligned}$$

and imposing $\partial E^2/\partial\boldsymbol{\beta} = 0$ we get

$$(X^\top {C_e}^{-1}X)\boldsymbol{\beta} = X^\top {C_e}^{-1}\boldsymbol{y}. \tag{13.21}$$

If

$${C_e}^{-1} = \begin{pmatrix} 1/{\sigma_1}^2 & 0 & \cdots & 0 \\ 0 & 1/{\sigma_2}^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & \cdots & 0 & 1/{\sigma_n}^2 \end{pmatrix}$$

each datum is independent of the others and gets weighted by $w_i = 1/{\sigma_i}^2$. This will make high-precision measurements count more than poor precision ones, confirming the experimental knowledge that one good measurement can be worth 10 or 100 bad ones.

General Least Squares

So far we have considered the simple model:

$$y_i = \beta_0 + \beta_1 x_i$$

which is linear in $\boldsymbol{\beta}$ and $\boldsymbol{x}$. This is not always the case. In general we have

$$\boldsymbol{y} = X\widetilde{\boldsymbol{\beta}}$$

where X can be given by, for example, a cubic polynomial

$$X_{ij} = \begin{pmatrix} 1 & x_i & x_i^2 & x_i^3 \end{pmatrix}.$$

We can also have $\widetilde{\boldsymbol{\beta}} = f(\beta_i)$, e.g. $\log(\beta_i)$, β_i^{α} (non linear functions). It is possible to fit very general classes of models this way. Examples are:

- Spherical harmonic coefficients,
- Fourier series,
- more complex functions (non-linear least squares).

In summary, least squares are a powerful tool for curve-fitting, and we've only just scratched the surface. In the next Chapter we shall see how to deal with ill-posed problems where even least squares need to be amended (regularized) to yield reasonable estimates of geophysical model parameters.

Chapter 14

DISCRETE INVERSE THEORY

Estimating parameters from observations is almost always an inverse problem.

I Classes of inverse problems

Consider the seismic tomography problem

$$\delta t_i = \int_{path} \delta s(\boldsymbol{x}) d\boldsymbol{l} \tag{14.1}$$

where s represents the "slowness" $s = 1/v = dt/dl$.

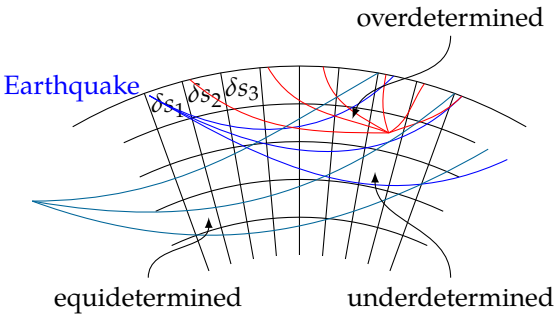


Figure 14.1: Schematic representation of the seismic tomography problem: from seismic measurements of travel times between different disturbances, we try to invert for fundamental properties of Earth materials (the elastic moduli), which are related to the propagation speed of elastic waves. Some boxes are illuminated by several ray paths, others by one, and still others are in the dark.

Continuous problems must always be solved numerically if no analytical solution exists (as is the case here):

$$\underbrace{\delta t_i}_{d_i} \approx \sum_{j=1}^{p} \underbrace{P_{ij}}_{G_{ij}} \underbrace{\delta s_j}_{m_j} \quad \Leftrightarrow \quad G\boldsymbol{m} = \boldsymbol{d} \quad \text{matrix inversion problem.}$$

With n observations and p model parameters, there are 3 possibilities:

1. $n = p$: **equidetermined problem**, unique solution $\Leftrightarrow \det(G) \neq 0$,
2. $n > p$: **overdetermined problem**, $G^\top G$ is $(p \times p)$;
3. $n < p$: **underdetermined problem**, $GG^\top$ is $(n \times n)$.

Overdetermined problems typically have no solution, and underdetermined problems potentially an infinite number of solutions. Something (regularization) has to be done in order to obtain a unique solution. The seismic tomography problem is a *mixed-determined* problem, some boxes are "illuminated" by many rays, some by none and the corresponding entries in G are zero. This means that we can find some non-null vectors $\boldsymbol{m}^0$ such that

$$G\boldsymbol{m}^0 = \mathbf{0}$$

so the null space of G can be large. If $\boldsymbol{m}^s$ is a solution, i.e.

$$G\boldsymbol{m}^s = \boldsymbol{d}$$

then $\boldsymbol{m}^{(s)} + c\boldsymbol{m}^{(0)}$ (c a real constant) is also a solution

$$G(\boldsymbol{m}^{(s)} + c\boldsymbol{m}^{(0)}) = \boldsymbol{d} \quad \text{(non-uniqueness)}.$$

So which solution should we pick?

II Reduced rank (TSVD) solution

SVD solution

Remember the singular value decomposition (SVD) for a "fat" matrix. Given an $n \times p$ matrix, G we can express it as

$$G = U\Sigma V^\top$$

where U is an $n \times n$ matrix and $U^\top U = I_n$, V is a $p \times p$ matrix, $V^\top V = I_p$ and Σ is an $n \times p$ matrix with the following structure

$$\begin{pmatrix} \sigma_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n & \underbrace{\dots \quad 0}_{p-n \text{ columns}} & \end{pmatrix}.$$

We want to solve the following equation

$$\boldsymbol{m} = G^{-1}\boldsymbol{d}.$$

It can be shown that the Moore-Penrose pseudoinverse of a matrix always exists:

$$G^\dagger = V\Sigma^\dagger U^\top$$

where $\Sigma^\dagger$ is given by

$$\begin{pmatrix} 1/\sigma_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1/\sigma_2 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 1/\sigma_n & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

The solution of rank k assigns $(\Sigma^\dagger)_{jj} = 0$ for $j > k \rightarrow$ truncation of singular values. For an overdetermined problem, the truncated SVD (TSVD) solution is the OLS solution. For an underdetermined problem, it is the *minimum norm solution*. One may write

$$\boldsymbol{m}^{(k)}_{\text{TSVD}} = \sum_{i=1}^{k} \left(\frac{\boldsymbol{u}_{(i)} \cdot \boldsymbol{d}}{\sigma_i} \right) \boldsymbol{v}_{(i)}.$$

The solution is a linear combination of the right singular vectors, weighted by a data dot-product and $1/\sigma_i$. The $\boldsymbol{v}_{(i)}$ are orthonormal and span the axes of the error ellipsoid for $\boldsymbol{m}$:

$$\text{Cov}(m_i, m_k) = \sum_{j=1}^{p} \frac{V_{ij} V_{kj}}{{\sigma_i}^2} \tag{14.2}$$

(cf principal components and EOFs).
Now, if G is singular then $\sigma_i \approx 0$ for i bigger than some index L from which we get an infinite weight. In this case $1/\sigma_i$ should be set to zero in $\Sigma^\dagger$, implying that one adds zero weight to the corresponding eigenvectors.

If $\sigma_i \approx 0$, the solutions will be dominated by noise. The singular value spectrum must be inspected for a break. Truncating at k will increase $\chi^2 = \|G\boldsymbol{m} - \boldsymbol{d}\|^2$ only slightly, while greatly decreasing the model complexity.

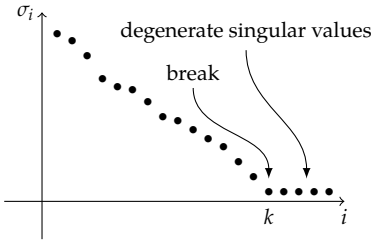


Figure 14.2: Searching for a break in the singular value spectrum. In this case, k is an obvious truncation

Model stability

$$\frac{\left\| \boldsymbol{m}_{true} - \boldsymbol{m}^{(k)} \right\|}{\|\boldsymbol{m}_{true}\|} \leq \kappa(G) \frac{\left\| \boldsymbol{d} - \boldsymbol{d}^{(k)} \right\|}{\|\boldsymbol{d}\|} \tag{14.3}$$

where $\boldsymbol{d}^{(k)} = G\boldsymbol{m}^{(k)}$, $\kappa(G) = \sigma_{\max}/\sigma_{\min}$ is the condition number of matrix G[1] and $k = \min(n, p)$. The larger $\kappa(G)$, the larger the distance from the true model.

[1] The condition number measures the feasibility of inverting G. It is such that a value close to 1 implies a well-conditioned inversion problem, while a large value implies a numerically unstable inverse. Singular matrices are characterized by $\sigma_{\min} = 0$, therefore $\kappa = \infty$.

Model Resolution

Consider $G\boldsymbol{m} = \boldsymbol{d}$. In order to evaluate how well G *resolves* $\boldsymbol{m}$, one could input a synthetic (i.e. made-up) model $\boldsymbol{m}'$, yielding the synthetic observations $\boldsymbol{d}' = G\boldsymbol{m}'$. Solve the inverse problem

$$\tilde{\boldsymbol{m}} = G^{\dagger}\boldsymbol{d}' = \left(G^{\top}G\right)^{-1} G^{\top}G\boldsymbol{m}$$

where $\left(G^{\top}G\right)^{-1} G^{\top}G$ is called *resolution matrix* R. For a perfect resolution $R = I$, in reality

$$R = G^{\dagger}G = \begin{pmatrix} & & 0 \\ & \iddots & \\ 0 & & \end{pmatrix}$$

That is, we have a "blurry diagonal" matrix. This tells which models are well resolved by the observations, and which are not. One can also test the inversion with a spike function[2] which is given by

$$\boldsymbol{m}_{spike} = \begin{pmatrix} 0 & \dots & 0 & \underset{i}{1} & 0 & \dots & 0 \end{pmatrix}$$

for the model i. This method is known as the Backus-Gilbert method (*Backus and Gilbert*, 1968).

[2] The numerical equivalent of a Dirac delta function

Data resolution

How well does a model fit the data?

$$\hat{\boldsymbol{m}} = G^{\dagger}\boldsymbol{d} \Rightarrow \hat{\boldsymbol{d}} = G\hat{\boldsymbol{m}} \Rightarrow \hat{\boldsymbol{d}} = G\left(G^{\dagger}\boldsymbol{d}\right) = \underbrace{GG^{\dagger}}_{N}\boldsymbol{d}.$$

N is called *data resolution matrix* and we have $N = U_k U_k^{\top}$ from TSVD. For a perfect model $N = I_p$ but it is typically "blurry", like R.

Note *Both R and N are independent of the data $\boldsymbol{d}$ and only depend on G. This can be taken into account for the experimental design and modeling assumptions, so as to meet scientific goals.*

III Tikhonov regularization

Motivation

A disadvantage of TSVD is that singular vectors are either *on* or *off*: either a singular vector $\boldsymbol{v}_i$ participates in the solution or it doesn't. It is a form of *regularization*. One may desire a smoother transition between "valid" and "invalid" eigenvectors $\boldsymbol{v}_i$. The overarching goal is still the same: to find the *least complex model* that fits the noisy data *without overfitting*.

To do so, we now seek $\boldsymbol{m}$ such that $\|\boldsymbol{d} - G\boldsymbol{m}\| \leq \delta$, and $\|\boldsymbol{m}\|$ isn't too large. Equivalently, we seek $\boldsymbol{m}$ such that the cost function

$$J_\alpha(\boldsymbol{m}) = \|\boldsymbol{d} - G\boldsymbol{m}\|^2 + \alpha^2\|\boldsymbol{m}\|^2$$

with $\alpha \geq 0$, is minimized. α is a *Lagrange multiplier* or *regularization parameter*. It represents the price to pay for having a more complex model: α discourages a large $\|\boldsymbol{m}\|$, because in the limit of $\alpha \to \infty$, $\|\boldsymbol{m}\| \to 0$.

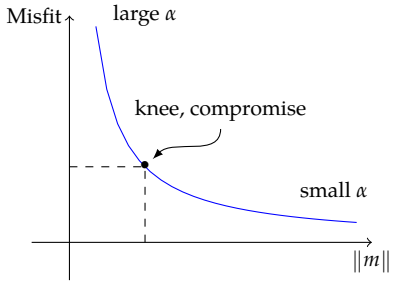


Figure 14.3: Typical Regularization tradeoff: a small regularization parameter means a good fit to the data but potentially a very wobbly model (large norm), while a large regularization parameter will encourage a smooth model at the expense of some model misfit. One again, there is no free lunch in a world of finite information.

Formulation

This approach may be recast as

$$\tilde{G}\boldsymbol{m} = \begin{bmatrix} G \\ \alpha I \end{bmatrix} \boldsymbol{m} = \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{0} \end{bmatrix} = \tilde{\boldsymbol{d}}$$

with I an $n \times p$ "identity" matrix. This can be rewritten

$$\tilde{G}^\top \tilde{G}\boldsymbol{m} = \tilde{G}^\top \boldsymbol{d} \Rightarrow \left(G^\top G + \alpha^2 I\right) \boldsymbol{m} = G^\top \boldsymbol{d}$$

The Tikhonov solution is given by

$$\boldsymbol{m}_{tikh}^{(\alpha)} = \left(G^\top G + \alpha^2 I\right)^{-1} G^\top \boldsymbol{d}.$$

The TSVD solution is given by

$$\boldsymbol{m}_{TSVD}^{(k)} = \sum_{i=1}^{p} f_i^{(k)} \left(\frac{\boldsymbol{u}_{(i)} \cdot \boldsymbol{d}}{\sigma_i}\right) \boldsymbol{v}_{(i)}$$

where the f_i's are called *binary filter factors*

$$f_i^{(k)} = \begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}.$$

The Tikhonov solution may be expressed similarly using the SVD formulation

$$\boldsymbol{m}_{\text{tikh}}^{(\alpha)} = \sum_{i=1}^{p} f_i^{(\alpha)} \left(\frac{\boldsymbol{u}_{(i)} \cdot \boldsymbol{d}}{\sigma_i}\right) \boldsymbol{v}_{(i)}$$

but now the filter factors are given by

$$f_i^{(\alpha)} = \frac{{\sigma_i}^2}{{\sigma_i}^2 + \alpha^2} \begin{cases} f_i \to 1 & \text{for } \alpha \ll {\sigma_i}^2 \\ f_i \to 0 & \text{for } \alpha \gg {\sigma_i}^2 \end{cases}$$

This solution keeps the large singular values almost intact while it damps the small ones to zero. Tikhonov solutions push all eigenvalues α^2 away from zero, which removes the matrix singularity. This is at the cost of damping some of the modes with $\sigma_i \gg 0$, so it represents a trade-off between misfit and complexity. The solution is always smoother than the true model (a common feature of "ℓ^2 methods" – those that seek to minimize the ℓ^2 norm). ℓ_1 methods are generally free of such problems, but those are mathematically more complex: the optimization problem can no longer be solved analytically, even in simple cases. Nowadays, there exist well-established convex optimization algorithms (e.g. simplex) that can efficiently solve them, but they are still less prevalent than ℓ^2 methods. This could be because people actually like smooth solutions, though Nature is rarely that smooth.

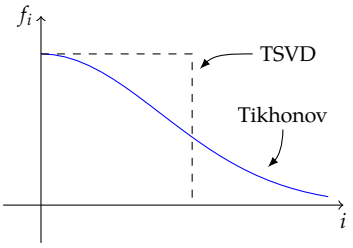


Figure 14.4: TSVD and Tikhonov filter factors

Choice of regularization parameter

Given this tradeoff, how should one go about picking the regularization parameter (α or k)? There are many methods:

- Picking the knee of the L-curve,
- picking the largest slope of the L-curve,
- visual inspection of smoothness,
- minimize the generalized cross validation function (GCV)[3]

[3] The GCV consists in finding the value of α that makes the best prediction of all $d_{(i)}$ for all inversions where the datapoint d_i is withheld (*Wahba*, 1990)

IV Recipe for underdetermined problems

1. Establish the theoretical null space (often hard or impossible in practice)
2. SVD(G) $\rightarrow$ inspect singular values
3. Regularize by using TSVD, Tikhonov or something else (guided by physics of the problem, not by observations)
4. Compute resolution matrices

   Model resolution $R = G^\dagger G$ (hat matrix)
   Data resolution $N = GG^\dagger$

   Tikhonov $G_\alpha^\dagger = \left(G^\top G + \alpha^2 I\right)^{-1} G^\top$.
5. figure out if this is what you want, else go back to step 3.

Chapter 15

LINEAR REGRESSION

In the previous chapter we saw how to fit curves through data using various techniques, all involving an exploration of the null space of a matrix. Linear regression is closely related, but the emphasis is slightly different. For instance, the goal may not solely be to find a best fit, but to make predictions and quantify uncertainties about these predictions (statistical forecasting). Thus, while the mathematics are very similar (least squares will pop up again), the spirit is much more akin to Part I, rooted in probability theory.

I Regression basics

Least Squares Solution

Regression seeks to estimate a variable Y from a predictor X, given a deterministic link function f. In general, one writes

$$Y = f(X) + \epsilon \tag{15.1}$$

where ϵ is some random quantity (errors). $f(X)$ may be understood as the conditional expectation of Y given X, so we have

$$Y = E(Y|X) + \epsilon \tag{15.2}$$

Linear models make the further assumption that

$$E(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{15.3}$$

hence $E(Y|X)$ is expressed as a linear combination of the predictor variables. It is most common to treat ϵ as a random sample from a normal random variable with zero mean:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{15.4}$$

II Simple linear regression

The 1D case ($p = 1$) is just the ordinary least squares we saw in Eq. (13.4) and Eq. (13.12). The idea is to minimize the residual sum of squares

$$\text{RSS}(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_i e_i^2 = \boldsymbol{e}^\top \boldsymbol{e}. \tag{15.5}$$

Yielding the familiar solution

$$\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{15.6}$$

which gives the slope and the intercept of the line fitting trough a cloud of points whose scatter is described by σ (Fig. 15.1)

$$\beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}. \tag{15.7}$$

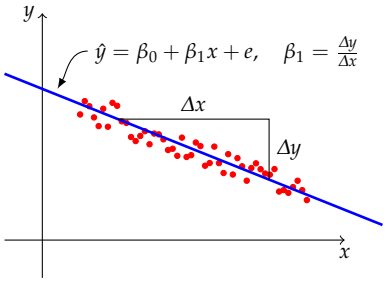


Figure 15.1: Example of simple linear regression

Analysis of Variance

The error associated with the linear estimate $\hat{y}_i = \beta_0 + \beta_1 x_i$ is $e_i = y_i - \hat{y}(x_i)$[1]. So the total error is given by the mean squared error (MSE):

[1] also known as "residual"

$$S_e^2 = \frac{1}{n-2} \sum_i e_i^2$$

in which there are only $n - 2$ degrees of freedom because two parameters (β_0 and β_1) had to be estimated. We can rewrite this term as follows:

$$S_e^2 = \frac{1}{n-2} \sum_i [y_i - \hat{y}(x_i)]^2 .$$

and we can insert a $(\bar{y} - \bar{y})$ in every term of the sum

$$S_e^2 = \frac{1}{n-2} \sum_i [(y_i - \bar{y}) + (\bar{y} - \hat{y}(x_i))]^2$$

and define the *sum of squares total*

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = (n-1) S_y^2$$

with S_y the sample standard deviation in y, and the *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2 = {\beta_1}^2 \sum_i (x_i - \bar{x})^2 = (n-1) {\beta_1}^2 S_x^2.$$

with S_x the sample standard deviation in x. One can show that:

$$S_e^2 = \frac{1}{n-2} \{SST - SSR\} = \frac{1}{n-2} \{SSE\}$$

where we have introduced $SSE = \sum_i e_i^2$. We then have

$$
\begin{aligned}
SST &= SSR + SSE \qquad (15.8)\\
SST &= (n-1) \times \text{ total variance in } y\\
SSR &= (n-1)\beta_1^2 \times \text{ total variance in } x\\
SSE &= \text{sum of squared regression residuals.}
\end{aligned}
$$

This is the most primitive form of an analysis of variance (ANOVA) and can help ascertain to what extent the linear fit accounts for the observed variations in Y.

Prediction intervals

Very often we fit models through data so we can make predictions from them. It is common to call x the *predictor* and y the *predictand*[2], and the idea is to use the relation at a point x_{new} that has not been used in fitting the regression model $\boldsymbol{y} = \beta_0 \mathbb{1}_n + \beta_1 \boldsymbol{x}$, and use it to predict a new value of the predictand, y_{new}. What error would we make?

[2] "that which we want to predict"

To see this, note again that **linear regression expresses the conditional distribution of y given x**, as:

$$y \mid x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, S_e^2) \qquad (15.9)$$

So what we want is $\mathbb{P}(y \mid x_{new})$. This is illustrated in Fig. 15.2.

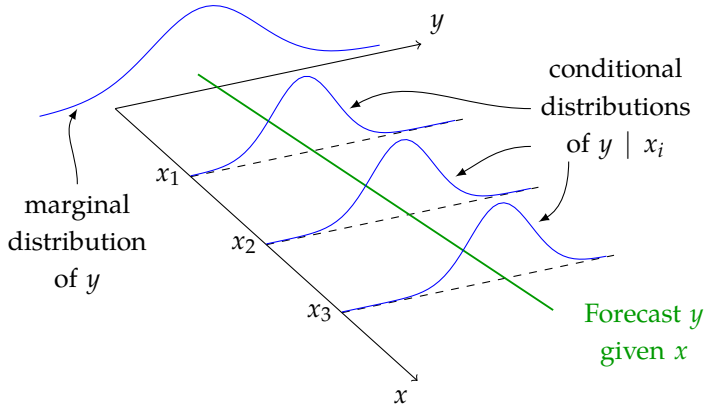


Figure 15.2: Conditional and marginal distributions for normal random variables Y and X. Note that for linear regression to work, only ϵ need be a normal RV, but this example nicely illustrates what is going on. At every point x, the regression model allows to predict y as a Gaussian with mean $\beta_0 + \beta_1 x$. Note how the marginal distribution of y is much wider than any of the conditional distributions, illustrating the fact that x reduces the uncertainty about y by virtue of their approximately linear relation

Because of the normality of the errors about the regression line, it turns out that $\hat{y} \pm 2S_e$ would be a good 95% prediction interval for y. Now, if x has not yet been observed, it makes sense that we would have more uncertainty about $y \mid x_{new}$. So a 95% prediction interval for y_{new} would be $\beta_0 + \beta_1 x_{new} \pm 2S_y$, with

$$S^2_{y|x_{new}} = S_e^2 \left[1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right] \qquad (15.10)$$

The expression Eq. (15.10) is proportional to the MSE but is inflated by two factors:

- the second term in the bracket comes from estimating μ as $\bar{x}$ over a finite sample; it goes to zero for $n \to \infty$,
- the third term comes from the uncertainty in the estimate of the slope and it grows as we move away from $(\bar{x}, \bar{y})$, the centroid of the dataset. Designing an experiment with a large range in x will mitigate that.

Conditional prediction interval of the mean

Now if we only try to estimate the mean of the dataset given new observations, its variance:

$$S^2_{\bar{y}|x_{new}} = S_e^2 \left[\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$

is smaller than (15.10) by an amount S_e^2.

Note that in the case of autocorrelated data, once again the IID assumption would be violated, and we would need to inflate the variance, hence the expected errors, by a factor $(1 + \phi)/(1 - \phi)$, with ϕ the lag-1 autocorrelation.

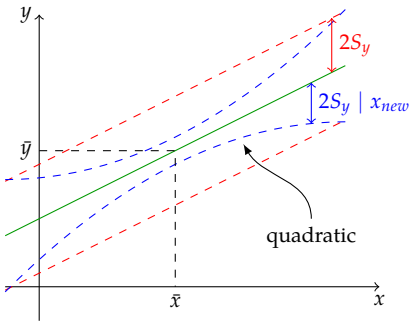


Figure 15.3: Prediction intervals. Note the parabolic shape of the prediction intervals, which widen away from the centroid of the dataset $(\bar{x}, \bar{y})$

III Model checking

Thanks to the wonders of modern programming, fitting a regression model is now so unbelievably trivial[3] that a monkey could do it. However, the result is of little value unless one can confirm that a few basic assumptions are met. The most important one is that residuals e_i need to be IID normal, since this was the basis for applying least squares (maximum likelihood estimation under a normal model) in the first place. The onus is on you to convince your readers that your statistical model actually fits the data. If not, there are plenty of more complex modeling options to go by.

[3] Matlab:fitlm(); Python:scikit-learn(); or statsmodels; R:lm()

Regression residuals

Inspecting residuals is absolutely critical. Important things to check for:

Normality Residuals should be normally distributed, which can be graphically checked with a simple histogram, a Q-Q plot, or more formally via a Lilliefors or Jarque-Bera test. They may only be approximately normal, which is in general no cause for alarm.

Homogeneity Residuals should not exhibit structure (*e.g.* Fig. 15.4).

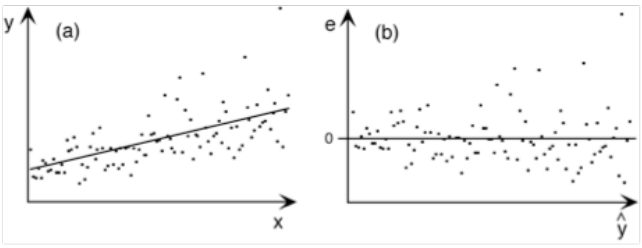


Figure 15.4: Heteroskedastic residuals. The left panel shows an attempt at fitting data whose variance increase over time, a property called heteroskedasticity. When such is the case, the errors inherit the same property (right panel), which violates the regression assumptions.

Independence Residuals should be non-persistent.

If x is serially correlated (i.e. autocorrelated), then regressing on x_i and not x_{i-1} might carry some memory, so the e_i's will exhibit autocorrelation: consecutive values will no longer be independent. A useful diagnostic is the *Durbin-Watson statistic* d:

e.g. `statsmodels.stats.stattools.durbin_watson()`

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}.$$

d can vary in the range $0 \leq d \leq 4$, and is 2 for uncorrelated residuals. A value substantially less than 2 indicates that neighboring residuals tend to be very similar, which is usually cause for alarm, though it must be established quantitatively via a (tabulated) distribution for the statistic. A value substantially larger than 2 indicates anti-correlated residuals, also a cause for alarm.

ANOVA table

Another important aspect of model checking is to parse the various components of variance, via an ANOVA table. There are three main measures of the quality of the fit:

Mean Squared Error $MSE = \frac{1}{n-2}\sum_i e_i^2$, it should be as small as possible.

Coefficient of determination R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

A perfect regression has no scatter about the regression line $\hat{y}$ so, in this case, we would have $MSE = 0$, $R^2 = 1$. $R^2 = 0$ indicates a useless regression. More generally, R^2 represents the fraction of variance in y accounted for by the variance in x.

F ratio $F = MSR/MSE$ follows an $F_{1,n-2}$ distribution for Gaussian residuals (rarely useful in practice).

They can be summarized in the form of an ANOVA table (Table III)

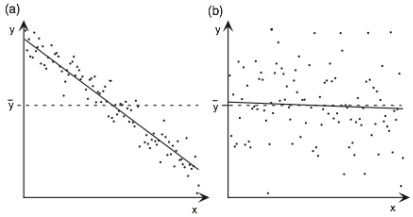


Figure 15.5: Regression fit and variance "explained". On the left, the slope of the regression is very strong (a confidence interval for β_1 would exclude zero with high confidence), so SSR dominates SST. On the right, the slope is near zero and SSE dominates the total variation SST

Source	df	SS	MS	F
Total	$n-1$	SST		
Regression	1	SSR	$MSR = SSR/1$	MSR/MSE
Residual	$n-2$	SSE	$MSE = S_e^2$	

IV Multiple Linear Regression

Now Y depends on multiple inputs $(X_1, X_2, \ldots, X_p)$

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon \tag{15.11}$$

so there are $(p+1)$ regression coefficients to estimate.

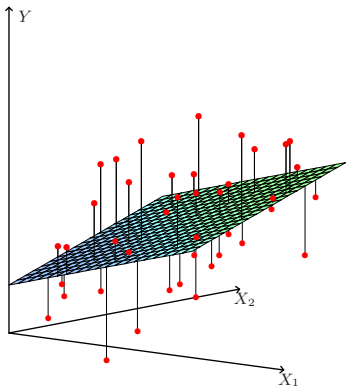


Figure 15.6: Least squares solution for $p = 2$, minimizing the sum of squared residuals from Y [from *Hastie et al.* (2008, Chap. 3)]

Least squares solution

It is customary to gather all variables into the design matrix X, of dimensions $n \times (p+1)$.

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{pmatrix}$$

And the solution is the familiar ordinary least squares (OLS) solution :

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X^\top \boldsymbol{y}.$$

which consists of fitting a hyperplane through the data (Fig. 15.6). It can be shown that

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}, (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma^2) \tag{15.12}$$

which tells us everything we want to know about the solution. The noise variance σ^2 must usually be estimated from the data, often via the residual sum of squares: $\hat{\sigma}^2 = \frac{1}{n-p-1}\text{RSS}$ (no bias), which has distribution

$$(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1} \tag{15.13}$$

How significant is the role of each predictor X_j? $|\beta_j| \gg 0$ indicates a large influence of x_j on y[4]. More formally, one tests the hypothesis that a particular coefficient $\beta_j = 0$, by forming the standardized coefficient or Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \tag{15.14}$$

[4] this comparison is only fair when all the input variables have similar magnitudes

where v_j is the j^{th} diagonal element of $(X^\top X)^{-1}$. Under the null hypothesis that $\beta_j = 0$, z_j is distributed as t_{N-p-1} (a t distribution with $N - p - 1$ degrees of freedom), and hence a large (absolute) value of z_j will lead to rejection of this null hypothesis. If $\hat{\sigma}$ were replaced by a known value σ, then z_j would have a standard normal distribution. The difference between the tail quantiles of a t-distribution and a standard normal become negligible as the sample size increases, and so normal quantiles are often quite a good approximation (*Hastie et al.*, 2008).

Another way to see this is to report an approximate 95% CI for β_j, $[\hat{\beta}_j \pm 2 \cdot s.e.(\hat{\beta}_j)]$. If this interval excludes zero, then the effect is significant, provided the predictors $X_1, \cdots X_p$ are independent. When they are not, one must be far more careful, and control for how one variable may speak through another (see "Regularized regression").

The Multivariate ANOVA is

Source	df	SS	MS	F
Total	$n-1$	SST		
Regression	k	SSR	$MSR = SSR/k$	MSR/MSE
Residual	$n-k-1$	SSE	$MSE = SSE/(n-k-1)$	

An example: fitting the Keeling curve

Let us try to fit the famous Keeling Curve, sketched out in Fig. 15.7, using several predictors. First note that the curve displays some regular oscillations superimposed on a roughly exponential trend. At first, let's see how well we would do with a simple linear fit, and then add complexity. Our primary variable is time t, expressed in months since Jan 1, 1958.

Linear fit $[CO_2] = \beta_0 + \beta_1 t$; the least squares solution yields $\beta_0 = 308.6$, $\beta_1 = 0.12$, and $R^2 = 0.977$. Despite this glorious statistic, this is obviously a poor fit; in particular, we are missing the curvature, which is a first order feature of the dataset. Prediction intervals are of order 6.6ppm.

Quadratic fit $[CO_2] = \beta_0 + \beta_1 t + \beta_2 t^2$. Adding a quadratic term captures the exponential curvature to a very large extent, so additional terms (t^3, t^4, etc) are no warranted. The prediction interval width is on the order of 4.4ppm, which is better than the linear fit. However, an inspection of the residuals (not shown) reveals a non-random scatter. The Durbin-Watson statistic ($\simeq 0.135$) is also very low, suggesting highly autocorrelated residuals. Clearly, this model ignores the well-known seasonal cycle in carbon capture and release by the biosphere, so we need to add sinusoidal terms.

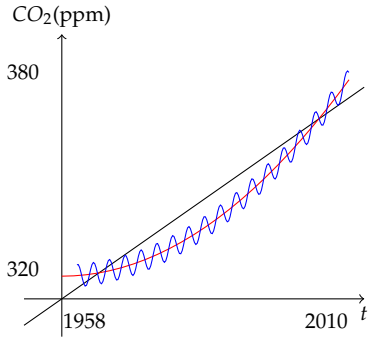


Figure 15.7: Keeling Curve: CO_2 measurements at Mauna Loa observatory, Hawaii, from March 1958 to May 2010.

Quadratic + Harmonic fit

$$[CO_2] = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \beta_4 \sin\left(\frac{2\pi t}{12}\right)$$

This last fit has an R^2 of 99.83%, prediction intervals $\pm 2\sqrt{MSE}$ are only 1.8 ppm so it is now a much better fit, nearly 73% more precise than the linear fit.

One lesson here is that we had to include not only t but also non-linear functions of t; these we call "derived predictor variables", since they are transformed versions of the original predictor. The result is non-linear in t, but linear in β so we still have a linear regression. Also, note that predictors themselves need not be Gaussian as long as the residuals are Gaussian.

Overfitting

One peculiar feature of multiple linear regression models is that one can always add variables to a regression model, e.g.

$$\begin{aligned}[CO_2] =& \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \beta_4 \sin\left(\frac{2\pi t}{12}\right) + \\ & \beta_5 \times \text{IBM stock} + \beta_6 \times (\text{Monkey population in Bhutan}) + \\ & \beta_7 \times (\text{SAT scores at USC}).\end{aligned}$$

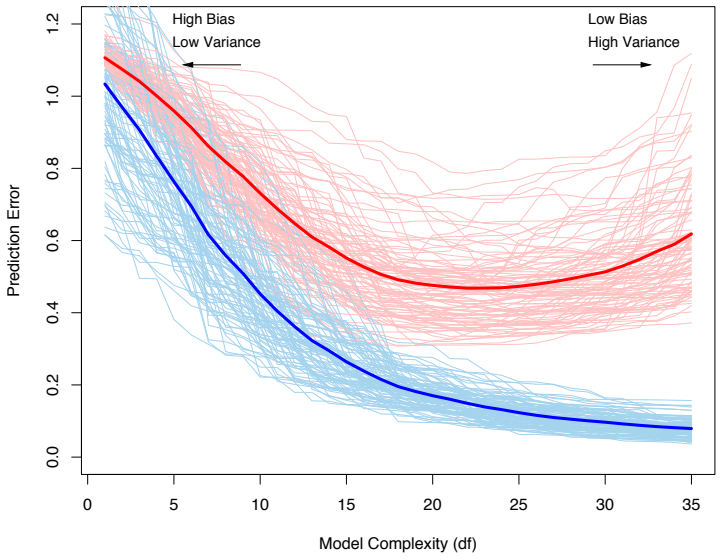


Figure 15.8: In-sample (blue) vs out-of-sample (red) prediction error. Each curve is the result of a given random sample from the data, whose average is shown by the thick line. This figure makes plain that a more complex model will *always* decrease in-sample prediction error, while out-of-sample prediction error will bottom out for some intermediate value, and then increase away from this minimum. This is the essence of overfitting: a more complex model will do better at predicting the training data, but may be completely off outside of it. Since we most often use regression models for prediction, it is absolutely essential to be aware of this fact. [figure from *Hastie et al.* (2008, Chap. 7)]

Why would we do that? It turns out that such "nonsense predictors" might lower the MSE on the *training sample*[5], so if we fish for potential predictors in this myopic way, we would be tempted to use them for prediction. However, these predictors may completely ruin *out-of-sample predictions*; this is called *overfitting* (Fig. 15.8). To guard against this phenomenon, one can do several things:

[5] the n observations used to fit the regression

- include only variables relevant to your problem,
- screen them for significant relations with y
- use some "IC", like AIC or BIC, to select the appropriate predictors.
- cross-validation: reserve part of the training set as validation set.

The best model is the one minimizing out-of-sample prediction error to guard against overfitting[6]. AIC or BIC attempt to do this, but depending on the situation may either overfit or underfit. A more reliable method to estimate out-of-sample prediction error is k-fold cross validation (KCV). It consists of chopping the training sample in k parts ("folds"), training on $k-1$ folds and using the remaining fold to compute prediction error. The expected prediction error is then estimated as the average prediction error on each fold. The choice of k is a balance between information amount and computational feasibility. KCV can be expensive but is usually worth it, since the EPE would take a U-shape very much like the red curve on Fig. 15.8 (though usually not as smooth).

[6] Note that all these comments apply *a fortiori* when Y is multivariate, so several models are being fitted together.

Regularized regression

Often predictor variables are colinear (e.g. t and t^2); one variable says something about another. In that case $\boldsymbol{X}^\top \boldsymbol{X}$ will be singular: some of its eigenvalues will be zero, or numerically very close to zero. As a result, the central quantity $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ may be unbounded. To overcome this problem, the covariance matrix needs to be *regularized*, which amounts to filtering out small eigenvalues. There are several ways to do it:

PC regression The idea here is to "orthogonalize" predictors by using principal component analysis. By definition, PCA yields orthogonal coordinates that maximize the variance of a dataset, so we can use them as a new basis for the regression. This is exactly like the TSVD solution of Chap 14, Sect. II. One problem is to decide objectively how many PCs (SVD modes) to retain; there are some rules to do this *a priori* (*Wax and Kailath*, 1985; *Nadakuditi and Edelman*, 2008), but cross-validation usually is a safer bet when prediction is involved.

Ridge regression (aka Tikhonov regularization) bumps diagonal entries of the covariance matrix, resulting in a smooth filtering of its eigenvalue spectrum; see Chap 14, Sect. III. Ridge regression may be seen

as adding a penalty to the ℓ^2 penalty to the MLE of β to ensure that the solution favors small coefficients. It turns out that this lowers the effective number of parameters to estimate. Because it keeps coefficients in check, ridge regression is variously called "biased regression", or a "shrinkage method" [7]. "Bias" is usually something to avoid, but in this case we trade a little bias for a lot of variance, so even though coefficients are biased low, the total MSE may be substantially lowered. The optimal ridge parameter may be identified via generalized cross-validation (*Golub et al.*, 1979), with some complications (*Wahba and Wang*, 1995).

[7] for large values of the ridge parameter, β shrinks towards zero

Least Angles Regression (LASSO) The lasso is a shrinkage method like ridge regression, with subtle but important differences. The main one is that the ℓ^2 penalty is replaced by an ℓ^1 penalty on $\boldsymbol{\beta}$. Big deal, you say? Well the ℓ^1 norm encourages coefficients to go to zero *exactly*, so the LASSO can actually eliminate variables altogether. In contrast, ℓ^2 methods will tend to shrink all coefficients, but will never tell you: get rid of this guy. The LASSO can do that, so it is much more of a *model selection* tool than a regularization tool[8]. In the end, starting from a large pool of potential predictors, it reduces the number of parameters that must be estimated. It has been used very successfully in genetics, where it is used to identify genotypes that have predictive power over phenotypes.

[8] "*If a traveler comes to a fork in the road, the ℓ^1 norm tells him to turn either left or right, which the ℓ^2 norm tells him to head straight for the bushes.*" – A. Tarantola

V What would a Bayesian do?

As in every field of data analysis, there is a Bayesian view. And as in every field, for a suitable choice of priors, the Bayesian solution yields an identical solution as the traditional method, but often with stronger insights. A Bayesian would first write the model

$$y \mid \{\beta, \sigma^2, X\} \sim \mathcal{N}(X\beta, \sigma^2 I_n) \tag{15.15}$$

Starting from a non-informative prior on the model variables:

$$p(\beta, \sigma^2 \mid X) \propto \sigma^{-2} \tag{15.16}$$

application of Baye's theorem yields the posterior distribution of model parameters:

$$\beta \mid \sigma^2, y \sim \mathcal{N}(\hat{\beta}, V_\beta \sigma^2) \tag{15.17}$$

where, as before:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y \tag{15.18a}$$

$$V_\beta = (X^\top X)^{-1} \tag{15.18b}$$

which is just the ordinary least squares solution. With different priors one gets different results, of course. The ridge regression solution may

be obtained via an inverse Wishart prior on the covariance matrix; the LASSO solution via a Laplace prior on the covariance matrix.

The advantages of the Bayesian viewpoint are as usual: the inference is model-based, everything has a well-defined distribution, the assumptions are transparent, and the model can be made as complex as one wants and still follow the basic desiderata of probability theory as logic (Chap 2). Thus, while in elementary cases (and given suitable choices) the Bayesian solution may yield the good old OLS solution, the framework offers much more flexibility when complex (e.g. hierarchical) models are warranted (see, *Gelman et al.*, 2013, Chap. 14, for an in-depth treatment), or when less friendly distributions are encountered.

Chapter 16

CONCLUSION & OUTLOOK

Like many of the students who take this class, you may have started this book with little mathematical background and a thirst to better understand the Earth. Or you may have wandered here through the intertubes, trying to get a little understanding of how to analyze data beyond the shallow teachings of the "Data Science" fad.

We hope that the journey so far has lifted the curtain on what lies behind data analysis, and that you are now a more sophisticated consumer of data analysis methods.

In future editions of this book, we would like to bring a few improvements:

- Fix all the pesky typos lacing this book (the more we fix, the more we find)
- Fix figures so that they appear correctly in all PDF readers. [in progress]
- include an additional chapter on age modeling and related techniques (leveraging the advances of the GeoChronR project)
- include a new chapter on climate field reconstructions of the Common Era, deconstructing a peer-reviewed article and isolating all the data-analytic components for pedagogical purposes.

Part IV

Mathematical Foundations

Appendix A

CALCULUS REVIEW

I Differential Calculus

Derivative

In calculus, we are interested in the *change* or *dependence* of some quantity, *e.g.* u, on small changes in some variable $t \in \mathbb{R}$. If u has value u_0 at t_0 and changes to $u_0 + \delta u$ when t changes to $t_0 + \delta t$, the incremental change can be written as

$$\delta u = \frac{\delta u}{\delta t}|_{t_0} \delta t. \tag{A.1}$$

The δ here means that this is a small, but finite quantity. If we let δt get asymptotically smaller around t_0, we arrive at the *derivative*, which we denote with $u'(t)$:

$$\lim_{\delta t \to 0} \frac{\delta u}{\delta t}|_{t_0} = \frac{du}{dt}. \tag{A.2}$$

The limit in Eq. (A.2) will work as long as u doesn't do any funny stuff as a function of t, like jump around abruptly. When you think of $u(t)$ as a function (some line on a plot) that depends on t, $u'(t)$ is the slope of this line that can be obtained by measuring the change δu over some interval δt, and then making the interval progressively smaller.

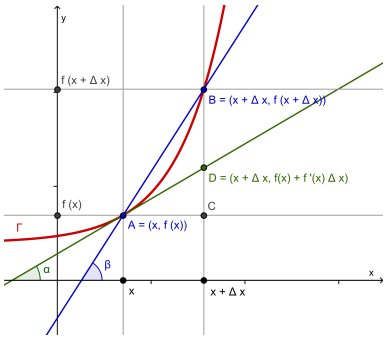


Figure A.1: The derivative as a tangent slope (Source:Wikimedia Commons). Note that the first order approximation $f(x) + f'(x)\Delta x$ captures the qualitative behavior of $f(x)$, but is still very far off. To do better, one should include more terms – this is the point of a Taylor expansion (Sect. III)

Interpretation

$u'(t_0)$ has a clear geometric interpretations as the slope of the tangent to the curve $\{t; u(t)\}$ at point $(t_0, u(t_0))$ (Fig. A.1). Its physical interpretation is that it reflects the instantaneous rate of change of the function u (for instance, if $u(t)$ is the velocity, then $u'(t)$ is the acceleration).

Note that in mechanics, $u'(t)$ is often denoted $\dot{u}(t)$, though more a rigorous notation is that of Leibniz $\frac{du}{dt}$. Leibniz's "differential" notation is clear in three respects:

- it makes clear which variation we are considering, because u could depend on other variables (e.g. x, y, z).

- it makes clear that this is ratio of differentials, so the notation is faithful to the definition.

- it expresses differentiation as an *operator* acting on a function, which allows differential operators to be defined easily, and makes multiple derivatives and the chain rule very easily understood.

That being said, it is more cumbersome than the prime, so most lazy people use $u'(t)$ to denote the derivative.

Properties

If you need to take derivatives of combinations of two or more functions, here called f, g, and h, there are four important rules (with a and b being real constants):

Linearity

$$(af + bg)' = af' + bg' \tag{A.3}$$

Product rule:

$$(fg)' = f'g + fg' \tag{A.4}$$

Quotient rule:

$$\text{If} \quad f(x) = \frac{g(x)}{h(x)} \tag{A.5}$$

$$\text{Then } f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2} \tag{A.6}$$

Chain rule (inner and outer derivative):

$$\text{If} \quad f(x) = h(g(x)) \tag{A.7}$$

$$\text{Then } f'(x) = \frac{df}{dx} = \frac{dh}{dg}\frac{dg}{dx} = h'(g(x))g'(x) \tag{A.8}$$

i.e. derivative of nested functions are given by the outer times the inner derivative.

Common Derivatives:

Here are some of the most common derivatives of a few functions:

function $f(x)$	derivative $f'(x)$	comment
x^p	px^{p-1}	special case: $f(x) = c = cx^0 \rightarrow f'(x) = 0$ where c, p are constants
$\exp(x) = e^x$	e^x	that's what makes e so special
$\ln(x)$	$1/x$	
$\sin(x)$	$\cos(x)$	
$\cos(x)$	$-\sin(x)$	
$\tan(x)$	$\sec^2(x) = 1/\cos^2(x)$	

Higher Order Derivatives

If you need higher order derivatives, those are obtained by successively computing derivatives, *e.g.* the third derivative of $f(x)$ is

$$\frac{d^3 f}{dx^3} = \frac{d}{dx}\left(\frac{d}{dx}\left(\frac{df(x)}{dx}\right)\right).$$

Say, $f(x) = x^3$, then

$$\frac{d^3 x^3}{dx^3} = \frac{d}{dx}\left(\frac{d}{dx}\left(\frac{dx^3}{dx}\right)\right) = \frac{d}{dx}\left(\frac{d}{dx}3x^2\right) = \frac{d}{dx}6x = 6.$$

In general, the n^{th} derivative is denoted by $f^{(n)}$. This will become useful in Sect. III.

Differentiability

We should not take a function's niceness for granted. Let us introduce the $\mathcal{C}^n$ notation, which formalizes this notion. A function is called $\mathcal{C}^n$ if it can be differentiated n times, and the n^{th} derivative is still continuous. Continuity may be loosely described as the property that a function's graph can be drawn with a pencil without ever lifting it from the page[1]. We call $\mathcal{C}^n(I)$ the set of functions that are $\mathcal{C}^n$ over some interval I. Obviously, a function that is n times differentiable is $n-1$ times differentiable, so we have the following Russian doll relationship[2], valid for all I:

$$\mathcal{C}^0(I) \in \mathcal{C}^1(I) \in \cdots \in \mathcal{C}^\infty(I) \tag{A.9}$$

[1] A more formal definition is that, if for every x in a *neighborhood* of x_0, $f(x)$ is in a neighborhood of $f(x_0)$, then f is said to be continuous at x_0. This definition is valid over the whole real line, including points at infinity.

[2] which mathematicians call an inclusion

Yes, that was an ∞ symbol: some functions can be continuously differentiated an infinite number of times. A trivial example is the function $x \to f(x) = C$ (a constant), but more interesting examples are the usual functions exponential, sine, cosine, log, etc.

Partial Derivation

In physics, it is very common for a function to depend on more than one variable. For instance u could depend on space and time, which we write $u = u(x, y, z, t)$. In such cases one must account for variations along each of the coordinates, which are themselves described by *partial derivatives*:

$$\lim_{\delta x \to 0} \frac{\delta u}{\delta x} = \frac{\partial u}{\partial x}. \tag{A.10}$$

and similarly along other coordinates. The full differential, accounting for all variations, then writes:

$$du = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy + \frac{\partial u}{\partial z}dz + \frac{\partial u}{\partial t}dt \tag{A.11}$$

Which states that the total variation in u is the sum of all its (partial) variations along each coordinates, multiplied by the variation in each coordinate. Partial derivatives and the equations derived from them underlie much of modern physics, and we can't do them justice here. In this class they shall only be used for optimization, such as likelihood maximization (Sect. III).

II Integral Calculus

Inverse Derivatives

Taking an integral

$$F(x) = \int f(x)dx,$$

in a general (indefinite) sense, is the inverse of taking the derivative of a $\mathcal{C}^0$ function f, *i.e.* $F'(x) = f(x)$. Another way to put this in the fundamental theorem of calculus, valid for any $\mathcal{C}^1$ function:

Theorem 2 *Every continuously differentiable function f verifies*

$$\int_a^b f'(x)dx = f(b) - f(a)$$

Which states that integration and differentiation are inverse operations ; that is, if we integrate the derivative we get back where we started (with this subtlety that a definite integral from a to b will equal the difference $f(b) - f(a)$, not just $f(x)$, say).

Interpretation

Graphically, the definite (with bounds) integral over $f(x)$

$$\int_a^b f(x)dx = F(b) - F(a)$$

along x, adding up the value of $f(x)$ over little chunks of dx, from the left $x = a$ to the right $x = b$, corresponds to the *area under the curve* $f(x)$. This area can be computed by subtracting the analytical form of the integral at b from that at a, $F(b) - F(a)$. If no bounds a and b are given, the function $F(x)$ is only determined up to an integration constant c, because the derivative of a constant is zero. In physics, initial or boundary conditions are often used to determine the value of such a constant, which can be very important in practice.

If $f(x) = c$ (c a constant), then:

$$\begin{aligned} F(x) &= cx + d && \text{(A.12)} \\ F(b) &= cb + d && \text{(A.13)} \\ F(a) &= ca + d && \text{(A.14)} \\ F(b) - F(a) &= c(b - a), && \text{(A.15)} \end{aligned}$$

the area of the box $(b - a) \times c$.

Properties

A few conventions and rules for integration:

Notation:

Everything after the $\int$ sign is usually meant to be integrated over up to the dx, or the next major mathematical operator if the dx is placed next to the $\int$ if the context allows. Also:

$$\int dx\, f(x) = \int f(x)dx \qquad \text{(A.16)}$$

Linearity:

$$\int_a^b (cf(x) + dg(x))\, dx = c\int_a^b f(x)dx + d\int_a^b g(x)dx \qquad \text{(A.17)}$$

Reversal:

$$\int_a^b f(x)dx = -\int_b^a f(x)dx \qquad \text{(A.18)}$$

Zero length:

$$\int_a^a f(x)dx = 0 \tag{A.19}$$

Additivity:

$$\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx \tag{A.20}$$

Product rules:

$$\int f'(x)f(x)dx = \frac{1}{2}\left(f(x)\right)^2 + C \tag{A.21}$$

$$\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx \tag{A.22}$$

Quotient rule:

$$\int \frac{f'(x)}{f(x)}dx = \ln|f(x)| + C \tag{A.23}$$

Symmetry:

$$\int_{-a}^a f(x)dx = \begin{cases} 2\int_0^a f(x)dx & \text{if } f \text{ is even} \\ 0 & \text{if } f \text{ is odd} \end{cases} \tag{A.24}$$

Common Integrals

Here are the integrals of a few common functions, all only determined up to an integration constant C

function $f(x)$	integral $F(x)$	comment
x^p	$\frac{x^{p+1}}{p+1} + C$	special case: $f(x) = c = cx^0 \rightarrow F(x) = cx + C$
e^x	e^x +C	
$1/x$	$\ln(\|x\|) + C$	
$\sin(x)$	$-\cos(x) + C$	
$\cos(x)$	$\sin(x) + C$	

In general, the integral of an arbitrary function may be difficult to find, unless one recognizes one of the forms shown above. This is in

contrast to the derivative of a differentiable function, which can always be found. Two main methods exist to find more complicated integrals: integration by parts, and variable substitution. When no analytical solution exists, one must resort to numerical quadrature.

Numerical Quadrature

The term quadrature invokes squares, and it is not used here coincidentally. Indeed, one of the early applications of integrals was to compute areas, usually by diving domains into rectangles whose area was easy to compute. The basic idea behind the following methods to numerically evaluate the integral of a function f over $[a;b]$ is to chop up its the interval into smaller, contiguous segments, collectively known as a *subdivision* of $[a;b]$. There are many ways to do this, but the simplest is to divide it into n equal slices of width $\Delta = \frac{b-a}{n}$:

$$\begin{aligned}[a;b[&= \underset{k=n}{[a;a+\Delta[\cup [a+\Delta;a+2\Delta[\cup \cdots \cup [a+(n-1)\Delta;b[} \\ &= \bigcup_{k=1} [a+(k-1)\Delta;a+k\Delta[\end{aligned}$$

(whether b gets included in the calculation or not makes no difference, as a point has zero width. This of course, is only true if f isn't doing anything fishy at $x = b$, so we require that f be $\mathcal{C}^0$).

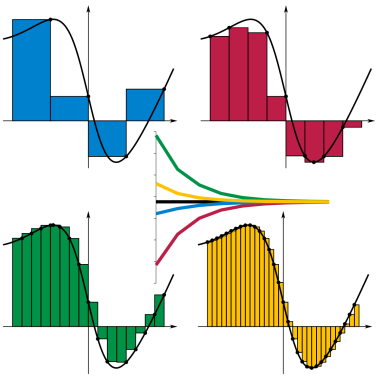


Figure A.2: Riemann summation in action. Right and left methods make the approximation using the right and left endpoints of each subinterval, respectively. Maximum (green) and minimum (yellow) methods make the approximation using the largest and smallest endpoint values of each subinterval, respectively. The values of the sums converge as the subintervals halve from top-left to bottom-right. The central panel describes the error as a function of number of bins n. Wikimedia Commons)

Rectangular Rule

The rectangular rule (aka Riemann sums) assumes that the function is constant over each subdivision. If we pick the end of each interval, we get:

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} f(a+k\Delta) \equiv A_n \tag{A.25}$$

As always in the numerical world, the approximation improves with n. In fact, one can show that

$$\lim_{n\to\infty} A_n = \int_a^b f(x)dx \tag{A.26}$$

That is, with infinitely fine subdivisions, one recovers the area under the curve *exactly*. This is something that a teacher of mine calls the "Stupid Limit", because one never has that much luxury. The goal of numerical analysis is to obtain an estimate that is as accurate as possible given computational constraints (which means that $n \ll \infty$, for starters). The rectangle rule is illustrated in Fig. A.2, exploring the impact of different quadrature choices.

Trapezoidal Rule

The rectangular rule is a pretty dumb one: in general f is far from constant, and the Stupid Limit is the only way to get away with assuming that it is. We surely can do better. The trapezoidal rule takes this one step further, and assumes that f is *piecewise linear* over each interval. This is akin to a first-order Taylor expansion (Sect. III). As you can see in Fig. A.3, now we can get away with a very crude subdivision and still espouse the true area rather closely. In Matlab, this method can be called via `trapz.m`.

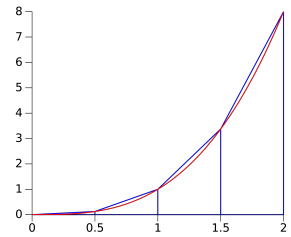


Figure A.3: Illustration of the trapezoidal rule. This case is particularly favorable to it as the function is extremely smooth, and varies monotonically, so even a rather broad discretization interval $\Delta = 0.5$ is enough to produce an excellent approximation of the full integral. Wikimedia Commons)

Simpson's rule

What in the world is better than a first-order Taylor expansion? A second-order Taylor expansion! That is the essence of Simpson's rule[3]. The basic principle is illustrated in Fig. A.4, though as in the previous methods, the approximations would improve with finer subdivisions. Over each interval $[x_i; x_{i+1}]$, of length Δ,

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \frac{\Delta}{6}\left[f(x_i) + 4f\left(\frac{x_i + x_{i+1}}{2}\right) + f(x_{i+1})\right] \qquad \text{(A.27)}$$

How good to these approximations go? Amazingly, the error is at least fourth order accurate[4], which means that it integrates cubics *exactly*, and its numerical efficiency makes it a darling of data analysts and engineers. In Matlab it may be called via `quad.m`, which is especially nifty because it adaptively chooses the subdivisions to focus on the parts of the function that are most variable (hence, most difficult to approximate by a parabola).

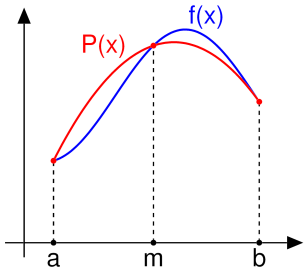


Figure A.4: Illustration of Simpson's rule. This figure is somewhat lame since it considers only a subdivision with $n = 1$. Wikimedia Commons)

[3] first used by the famous astronomer Johannes Kepler, who used it about 100 years before Thomas Simpson (1710-1761), whose name has stuck to it

[4] The error is $\frac{1}{90}\left(\frac{\Delta}{2}\right)^5 f(\xi)$, for some $\xi \in [a; b]$

Improper Integrals

So far we have only considered intervals of the form $[a; b]$, where a and b are both finite. What would happen if we let one or both of those bounds to reach infinity? It turns out that functions that decay fast enough toward either end of the real line may be integrated over $\mathbb{R}$. As an example consider the function $f(x) = \lambda e^{-\lambda x}$. It is trivial to show that:

$$\int f(x)dx = -e^{-\lambda x} + C = F(x)$$

so

$$\int_0^{\zeta} f(x)dx = F(\zeta) - F(0) = 1 - e^{-\lambda\zeta}$$

Now, as ζ approaches $+\infty$, the second term decreases exponentially fast, and the limit is well defined:

$$\lim_{\zeta\to+\infty}\int_0^{\zeta} f(x)dx \equiv \int_0^{+\infty} f(x)dx = 1 \qquad \text{(A.28)}$$

Such integrals are called *improper*, but they are legitimate as long as this limit exists. This particular function describes the *exponential distribution*, which is very useful in the study of waiting times and memoriless processes (Chapter 3).

Consider now the integral $\int_{-\infty}^{+\infty} \sin x dx$. Does such a thing exist? Given the symmetry property Eq. (A.24), one can show that for every a,

$$\int_{-a}^{+a} \sin x dx = 0$$

since sin is an odd function (symmetric about $(0,0)$). It is tempting to take the limit $a \to \infty$ and declare that the improper integral exists and equals zero. This reasoning, however, is 100% incorrect. For an improper integral with two infinite bounds to be defined, one has to look separately at the limits at $\pm\infty$. In the case above, neither limit is defined, so the integral doesn't exist.

Now, it is a fact of life that many of the most useful functions have no closed-form anti derivatives, a case in point being $\int e^{-\frac{x}{2}^2} dx$. Not only does its integral exist, but most remarkably

$$\int_{-\infty}^{+\infty} e^{-\frac{x}{2}^2} dx = \sqrt{2\pi} \tag{A.29}$$

However, you would not find that out by application of the two standard tools to find integrals (integration by parts or variable substitution); you have to either invoke Fubini's theorem or use a contour integral in the complex plane. In general, improper integrals are best tackled using measure theory, which is beyond our scope. For our purposes, a standard math textbook (*e.g.* *Abramowitz and Stegun*, 1965), table of integrals, the Mathematica software, or Matlab's Symbolic Math Toolbox will be of help with such integrals, as well as more complicated forms.

III Earth Science Applications

Measuring the Earth

Measuring mass and probability

One direct application of integrals is that if ρ is the density of some substance, its integral over the medium gives the total mass of that substance. Consider for instance atmospheric density (the mass per unit volume) as a function of height z:

$$\rho(z) = \rho_0 e^{-z/H}$$

where H is the so-called "scale height" (about 7km in Earth's troposphere).

The total mass of the atmosphere per unit area is:

$$\int_0^\infty \rho(z)dz = \rho_0 H$$

Similarly, we define in Chapter 3 the notion of *probability density*, which is a measure of the likelihood for a given variable to lie in a certain range. Any probability density $f(x)$ must have, by definition, unit mass over the real line[5], which writes:

$$\int_{-\infty}^{+\infty} f(x)dx \equiv \int_{\mathbb{R}} f(x)dx = 1 \tag{A.30}$$

[5] we are 100% sure that the variable cannot be smaller than $-\infty$ or greater than $+\infty$

Indeed this is what we found for the exponential distribution above. If we defined $\varphi(x)$ such that

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}^2} \tag{A.31}$$

Then, by virtue of Eq. (A.29) and the linearity of the integration operator, we'd find that

$$\int_{-\infty}^{+\infty} \varphi(x)dx = 1$$

which is a restatement of Eq. (A.30). $\varphi(x)$, the infamous "bell curve" is a centerpiece of the Laplace-Gauss distribution (Chapter 4), which is to probability theory what the Sun is to the solar system.

Measuring energy

Another application of integrals is to measure the energy of a signal, say $X(t)$, which could be either deep-sea temperature, atmospheric ground velocity at a point, the ground motion measured by a seismometer, or anything you'd like:

$$E = \int_{-\infty}^{+\infty} |X(t)|^2 dt \tag{A.32}$$

This, coupled with Parseval's theorem, is the foundation of spectral analysis (Chapters 7, 8 & 9).

Projection

Another application of integrals is their definition of an *inner product* (*cf.* Appendix B) between two functions f and g over some interval I:

$$\langle f, g \rangle = \int_I f(x)g(x)dx \tag{A.33}$$

which is how we measure the similarity between functions, and can project one onto the other.

Convolution

The convolution product, or simply *convolution* of two functions f and g is defined as:

$$(g * f)(x) = \int_{-\infty}^{+\infty} f(u)g(x-u)du = \int_{-\infty}^{+\infty} f(x-u)g(u)du \quad \text{(A.34)}$$

This operation is essential to describing the action of a linear system. For instance, let $x(t)$ be some input signal into a linear system (say, a filter) characterized by an impulse function $L(t)$, then the response of the system is

$$y(t) = \int_{-\infty}^{+\infty} x(u)L(t-u)du = \int_{-\infty}^{+\infty} L(u)x(t-u)du \quad \text{(A.35)}$$

If L describes the behavior of a *filter* acting on $x(t)$, then the convolution expresses the output of the filter once x has run through it. An 1D example is provided in Fig. A.6, where two filters (Gaussian and boxcar) are compared. A 2D example is shown in Fig. A.5, using bivariate Gaussian windows with half-widths set to 3 and 10 pixels, respectively.

The inverse operation, that of retrieving X given Y and some knowledge of L, is called *deconvolution*. Image enhancing (going from the bottom to the top of Fig. A.5) is one way to think of it. It is an inverse problem that can become rather thorny in the presence of noise, and usually needs some regularization (Chapter 15).

Figure A.5: Gaussian blurring, an example of discrete 2D convolution (Source:Wikimedia Commons).

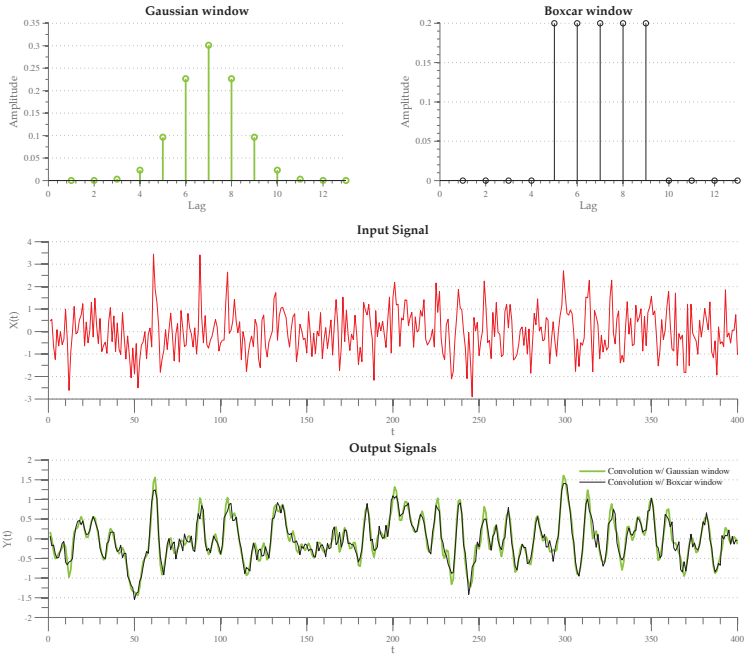


Figure A.6: Filtering timeseries, an example of 1D convolution. The output is visibly smoother than the inout, because convolution with either window amounts to averaging nearby points together, which cancels out high-frequency fluctuations and emphasizes the long-term behavior of the series. This is an example of *low pass* filter (Chapter 10). Note that the boxcar filter is blurrier and noisier than the Gaussian filter, which is one reason one running means are a bad idea.

We now show some applications of differential calculus.

Special points

Derivatives allow to characterize several *special points* around which the graph of $f(x)$ is organized:

Roots correspond to the points where $f(x) = 0$. In the case of the function shown in Fig. A.7 (a cubic polynomial), the roots can be found by various methods. For arbitrary functions, one must resort to iterative techniques like Newton-Raphson.

Extrema Extrema are either maxima or minima of a function, verifying $f'(x) = 0$ (a flat rate of change, locally). The derivative itself does not tell us whether the point is a maximum or a minimum, however. Higher-order derivatives are required in order to find this out. If $f''(x) > 0$ then the function is convex, and the extremum is a minimum ; if $f''(x) < 0$ the extremum is a maximum – see Fig. A.7. Care must be taken if $f''(x) = 0$.

Inflexion points are points where the second derivative $f''(x)$ passes through zero while changing sign. This is a turning point for f, as its curve changes from being concave upwards (positive curvature) to concave downwards (negative curvature), or vice versa.

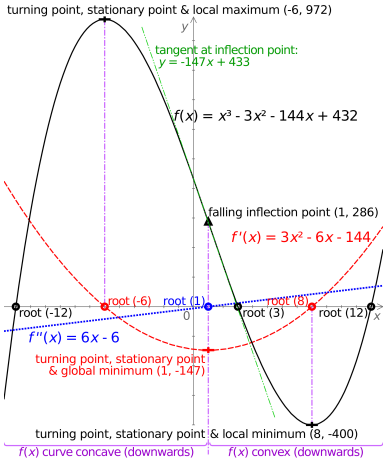


Figure A.7: Special Points of a Function $f(x) = x^3 - 3x^2 - 144x + 432 = (x - 3)(x - 12)(x + 12)$ (Source:Wikimedia Commons).

Optimization

A generic problem in applied mathematics is to find some sort of optimal solution to a problem: for example, one wants to find a curve that minimizes tension between points, a surface that minimizes the misfit to a set of measurements, fitting a line through a cloud of points, or finding the most likely value of a parameter given some observational constraints and prior knowledge. In many of these cases, one can define an objective function $S(x)$ whose maximum or minimum yields the desired solution. The equation $S'(x) = 0$ is used to find extrema and the sign of $S''(x) = 0$ (the curvature of S) is used to determine whether they are maxima or minima. In this class, we use optimization in Chapter 5 to estimate the parameters of a probability distribution (maximum likelihood method) and in Chapter 13 to fit curves to observations (it turns out that the two are deeply related).

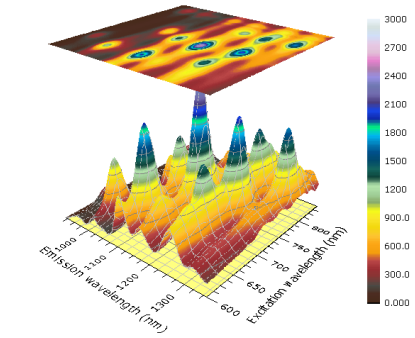


Figure A.8: Optimizing a multivalued function amounts to finding the peaks in a graph such as this one (Source:Imran Nazar).

Root finding

A related problem is that of finding the roots (solutions) of an equation, say $f(x) = 0$. We begin with a first guess x_0 for a root. Provided the function is $\mathcal{C}^1$, a better approximation x_1 is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{A.36}$$

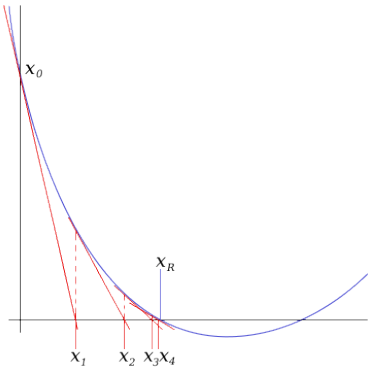


Figure A.9: Finding the roots of an equation via the method of Newton-Raphson

Geometrically, $(x_1, 0)$ is the intersection with the x-axis of the tangent to the graph of f at the point $(x_0, f(x_0))$.

The process is repeated as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{A.37}$$

until a sufficiently accurate value is reached (Fig. A.9). This is the method of Newton-Raphson. One can show that the process always converges unless the derivatives have too many zeroes, but its rate is only quadratic, so it can be very slow. Also, it depends critically on the initial guess, so one must use another method (e.g. eyeballing the graph) to find a good one.

Functional Representations

For many purposes it is sometimes useful to express an arbitrary function $f(x)$ in terms of simpler components. This indeed, is one meaning of the word *analysis*. A generic representation uses a linear combination of *basis functions*

$$f(x) = \sum_{k=0}^{\infty} a_k \phi_k(x) \tag{A.38}$$

As a general rule, the complexity of $\phi_k(x)$ tends to increase with k. Two examples are mentioned here.

Polynomial representation: Taylor Expansions

What if we pick $\phi_k(x) = x^k$? That is, what if seek a representation of f in terms of monomials of increasing complexity, up to degree $n > 1$? It turns out that we can do this for every well-behaved function, and that the value of the a_k coefficients only involves derivatives of f. This is called a Taylor approximation or Taylor series, and is valid for any $\mathcal{C}^n$ function. It is generally done in a neighborhood of a point x_0, where $f(x)$ can be expressed as follows:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2!} + \ldots \tag{A.39}$$

$$f(x) = \sum_{k=0}^{n} f^{(k)}(x_0)\frac{(x - x_0)^k}{k!} + R_n(x) \tag{A.40}$$

Here, $f^{(k)}(x_0)$ is the k^{th} derivative. $n!$ denotes the factorial, i.e.

$$n! = 1 \times 2 \times 3 \times \ldots n. \tag{A.41}$$

$R_n(x)$ is the *remainder* of order n , whose absolute value measures the goodness of this approximation.

There are a few formulations of the remainder, which determine the name of the formula. For instance, the Taylor-Lagrange formula states that there exists a number θ between x and x_0 such that:

$$R_n(x) = f^{(n+1)}(\theta)\frac{(x-x_0)^{n+1}}{(n+1)!} \tag{A.42}$$

So, as long as f is well-behaved (none of the derivatives get too large)[6] and $x - x_0$ is small, each successive term is smaller than the next, and the approximation becomes better and better: $\lim_{n\to\infty} R_n(x_0) = 0$. However, this ceases to be true as we get away from x_0, so this approximation is only useful locally.

Fig. A.10 illustrates this in the case of $f(x) = e^x$ an $x_0 = 0$, in which case $\forall k, f^{(k)}(x_0) = 1$, thus $a_k = 1/k!$. Note that the approximation is quite good around 0 even for $n = 2$, but gets worse as one gets away from the origin. If one wanted a useful approximation at $x = 3$, say, one could either set $x_0 = 3$, or keep $x_0 = 0$ but push the expansion to a higher order. Indeed, by $n = 6$ or 7, the true curve and its polynomial approximations are virtually indistinguishable on Fig. A.10. Which solution is the smarter choice?

Now, because $\exp(x) \in \mathcal{C}^\infty(\mathbb{R})$, one could push n to infinity, leading to the *exponential series*:

$$\forall \lambda \in \mathbb{R}, \quad e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{A.43}$$

This forms the basis for the Poisson distribution (Chapter 3, Sect. III). Functions that are equal to their Taylor series on their domain of definition are called *analytic*; this is a remarkable property, which we won't use too much in this course, but you should know how very extraordinary it is.

Another very important series is the *geometric series*:

$$\forall x \in]-1, 1[, \quad \frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \tag{A.44}$$

(that is, $a_k = 1$ for all k). With this, one can derive expansions of a host of other functions, like $\frac{1}{1+x}$, $\frac{1}{1+x^2}$, arctan, arcsin, arccos, etc. It comes in handy to approximate almost any ratio.

Example

What is a back-of envelope estimate of $q = 1/0.93$?
One recognizes the form above, with $x = 0.07$. So to first order, $q = 1 + 0.07 + \mathrm{O}(0.07^2) \simeq 1.07$. To second order, $q = 1 + 0.07 + 0.07^2 + \mathrm{O}(0.07^3) \simeq 1.07 + 0.0049 = 1.0749$. Here we used the "Big O" notation, which means "terms of this order, or higher".

This is in fact how calculators work (behind the scenes)! You can see how the higher order terms add up to increasingly small contributions,

[6] more precisely, $\mathcal{C}^{n+1}$ in an interval that encompasses x_0

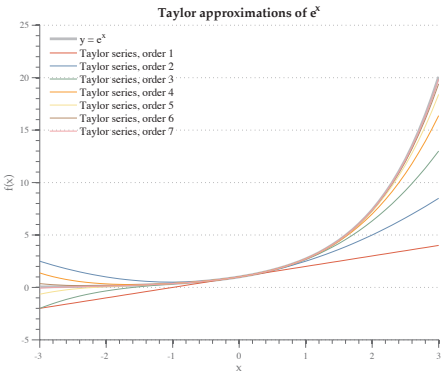


Figure A.10: Approximations of the exponential function by polynomials: the truncated Taylor series.

because $x^m < x^n$ for any $0 < m < n$ *as long as* $|x| < 1$. For larger $|x|$, these geometric approximations quickly become gruesome.

The geometric series is also useful in probability theory, giving its name to the geometric distribution (Chapter 4).

Trigonometric Representation: Fourier series

Another common form of functional representation takes the form of *trigonometric series*, and applies to T-periodic functions. This time, $\phi_k(x) = Z^k$, where $Z = e^{i2\pi}$ and $i = \sqrt{-1}$ (*cf.* Appendix C), so we are dealing with complex polynomials, which is not the most intuitive. After some rearrangements, it can be shown that this representation is equal to a sum of sines and cosines:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} [a_k \cos(k\omega x) + b_k \sin(k\omega x)] \qquad \omega = \frac{2\pi}{T} \tag{A.45}$$

This is called a *Fourier series* and is omnipresent throughout Chapter 7. The coefficients $\{a_k, b_k, k \in \mathbb{N}\}$ can be found by *projecting* f onto sines and cosines of different frequencies, using Eq. (A.33).

Transformation

Another major application of calculus that is useful for this course is the concept of integral transforms. In particular Laplace & Fourier transforms are omnipresent in signal processing and probability theory; we will mostly encounter them in Chapter 7.

Appendix B

LINEAR ALGEBRA

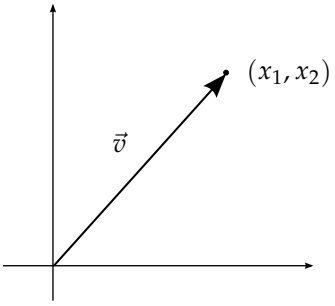
I Vectors

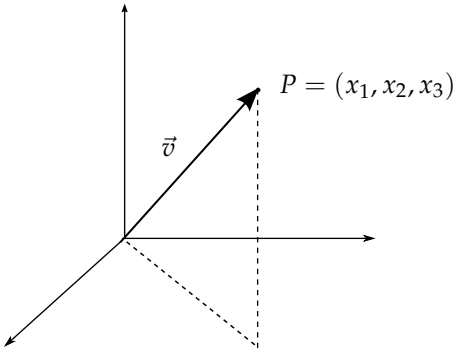
Definition 15 *A vector is a quantity having three properties: a magnitude, a direction and a sense.*

Example

- *Velocity vector indicating the movement of an object (e.g. planet)*
- *An n-tuple in the n-dimensional real space* $\mathbb{R}^n$*:* $(x_1, x_2, \ldots, x_n)$ *where* $x_i \in \mathbb{R}, i \in \{1, \ldots n\}$
- *in* $\mathbb{R}^3, \boldsymbol{v} = (1, 0, -1)$[1]

[1] Numpy: `v= [1, 0, -1]`





Different Ways to Measure the Length (Norms):

- Euclidean norm (or 2-norm): $\|\mathbf{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$.
- p-norm: $\|\mathbf{v}\|_p = (|v_1|^p + \ldots + |v_n|^p)^{\frac{1}{p}}$, $p \geq 1$[2].

  $p = 1$ $\quad \|\mathbf{v}\|_1 = |v_1| + \ldots + |v_n|$.

  $p = 2$ $\quad$ Euclidean norm.[3]

[2] `numpy.linalg.norm(x,p)`

[3] `numpy.linalg.norm(x,2)`

Limiting case: $p = \infty$ $\|\mathbf{v}\|_\infty = \max_{1 \le i \le n} |v_i|$.[4]

One can show that $\lim_{p\to\infty} \|\mathbf{v}\|_p = \|\mathbf{v}\|_\infty$

[4] `numpy.linalg.norm(x,Inf)`

If $\|\mathbf{v}\| = 1$, we say that $\mathbf{v}$ is a unit vector or equivalently, that $\mathbf{v}$ is normalized.

Scalar Multiplication

$$\alpha\mathbf{v} = \alpha\,(v_1, \ldots, v_n) = (\alpha v_1, \ldots, \alpha v_n)\,, \qquad \alpha \in \mathbb{R}\,.$$

II Matrices

Definition

A matrix is a quantity with two indices (a 2D array of numbers), which can be seen as a linear operator (a function) on vectors.

Linear means:

$$f\,(\alpha\mathbf{v}) = \alpha f(\mathbf{v})\,, \qquad \alpha \in \mathbb{R}\,, \tag{B.1a}$$

$$f\,(\mathbf{v} + \mathbf{w}) = f\,(\mathbf{v}) + f\,(\mathbf{w})\,. \tag{B.1b}$$

Example

$$\underbrace{\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}}_{A} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 2v_1 + 3v_2 \\ v_1 + 4v_2 \end{pmatrix}$$

$$\textit{Matrix} \times \textit{vector} = \textit{vector}$$

$$\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} \alpha v_1 \\ \alpha v_2 \end{pmatrix} = \begin{pmatrix} 2\alpha v_1 + 3\alpha v_2 \\ \alpha v_1 + 4\alpha v_2 \end{pmatrix} = \alpha \begin{pmatrix} 2v_1 + 3v_2 \\ v_1 + 4v_2 \end{pmatrix} = \alpha\,(A\mathbf{v})$$

Similarly $A\,(\mathbf{v} + \mathbf{w}) = A\mathbf{v} + A\mathbf{w}$.

Remark:

Canonical basis:

- $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix}$
- $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}$

$\Longrightarrow Ae_i = i$ column of A.

Hence, every linear operator on $\mathbb{R}^r$ may be represented as an $r \times r$ matrix. Conversely, every $r \times r$ matrix may be seen as representing the action of this operator.

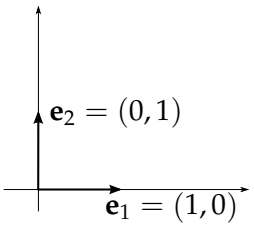
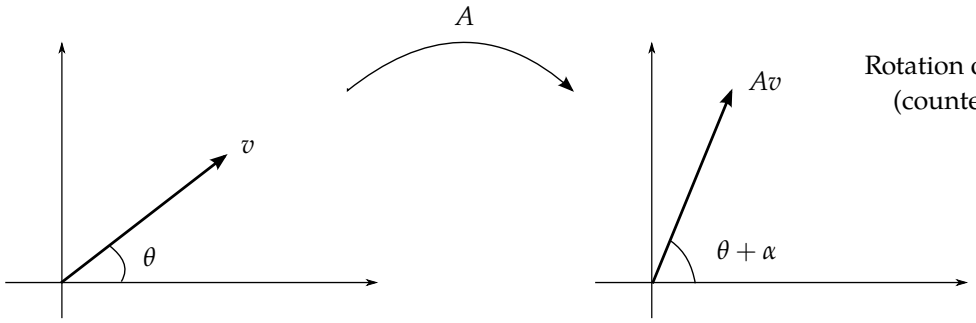


Figure B.1: The canonical basis of $\mathbb{R}^2$

An Example: Rotation matrices

Consider the linear operation of rotating a vector v by an angle α.

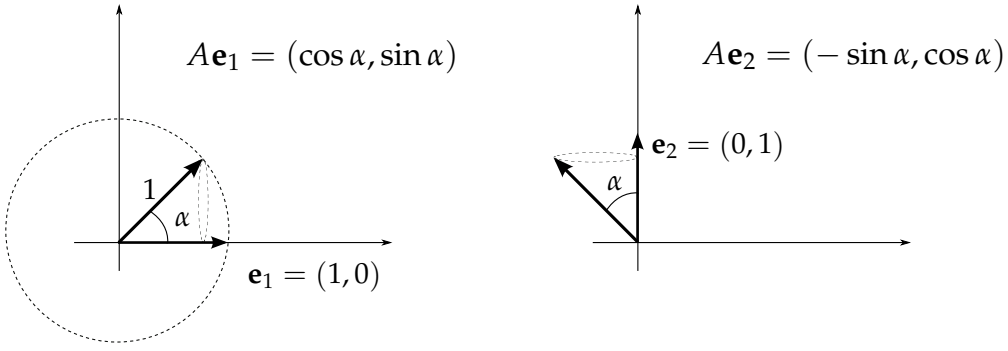


Rotation of an angle α (counterclockwise)

What formulation of A would represent this transformation? To figure this out, consider the action of the operator on the canonical basis vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$:

$$\begin{pmatrix} | & | \\ Ae_1 & Ae_2 \\ | & | \end{pmatrix}$$

It turns out:

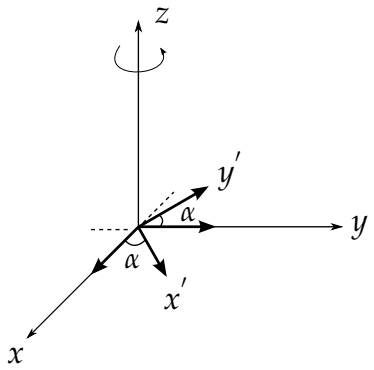


Therefore: $A = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$

In 3D

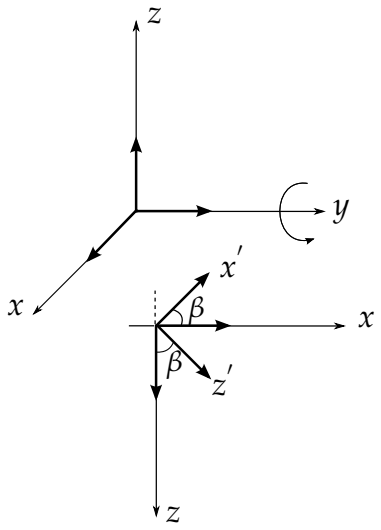
- Rotation around z of an angle α:

$$R_z^{(\alpha)} = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



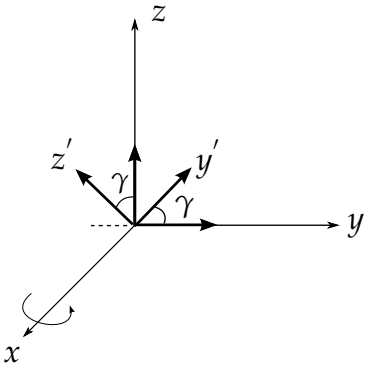
- Rotation around y of an angle β:

$$R_y^{(\beta)} = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}$$



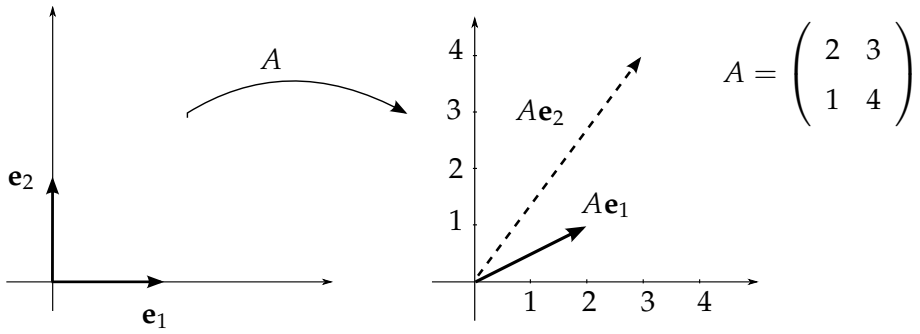
- Rotation around x of an angle γ:

$$R_x^{(\gamma)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{pmatrix}$$

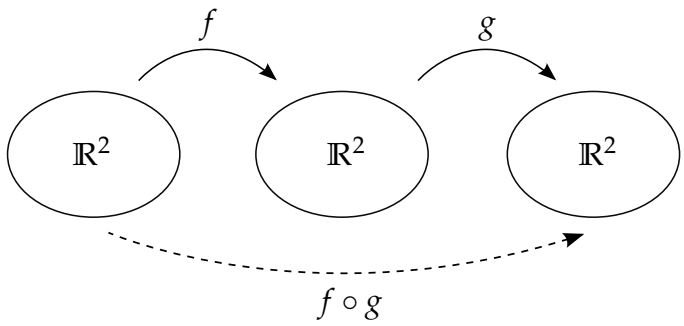


General rotation $R(\alpha, \beta, \gamma) = R_x(\gamma) R_y(\beta) R_z(\alpha)$. That is, one can express any rotation as successive applications of such rotation operators.

Matrix Multiplication



Matrix multiplication = "composition of functions"



A, B matrices:

$$AB = "A \circ B", \qquad (AB)\,\mathbf{v} = A\,(Bv)\ . \tag{B.2}$$

Matrix product ≠ Elementwise product

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix} \tag{B.3}$$

Formally, $(AB)_{ij} = \sum_k A_{ik} B_{kj}$. [5]

[5] In Numpy, `A*B` denotes elementwise product; For matrix multiplication, use `np.matmul(A,B)`

Transposition

The transpose of A, denoted $A^\top$ is obtained by permuting its rows and columns. If A is square, this is equivalent to flipping A along its main diagonal:

$$\left(A^\top\right)_{ij} = A_{ji} \tag{B.4}$$

Special Matrices

$$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \text{ identity} \qquad \begin{pmatrix} & & \\ & & \\ 0 & & \end{pmatrix} \text{ upper triangle}$$

$$\begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix} \text{ diagonal} \qquad \begin{pmatrix} & & 0 \\ & & \\ & & \end{pmatrix} \text{ lower triangle}$$

III Matrices and Linear System of Equations

$$\begin{cases} ax + by = c\,, \\ dx + ey = f\,, \end{cases} \tag{B.5}$$

with a, b, c, d, e and f given. Solve for x and y.

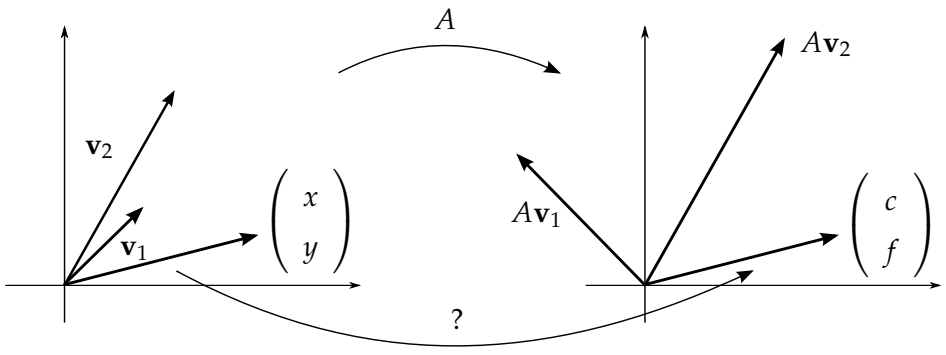
Matrix Form

$$\overset{A}{\begin{pmatrix} a & b \\ d & e \end{pmatrix}} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ f \end{pmatrix}\,. \tag{B.6}$$

matrix unknown vector known vector

The matrix $\begin{pmatrix} a & b \\ d & e \end{pmatrix}$ maps vectors to vectors (linear operation).

Geometric interpretation



3 possibilities:

- The system has a **unique** solution.
- The system has **multiple** solutions.
- The system has **no** solution.

We may know this via inspection of the **determinant** of A.

Determinants

Motivation

For square matrices $n \times m$, where $n = m$ with n the number of rows and m the number of columns, solve for $\mathbf{x}$.

$$\left(\quad \mathrm{A} \quad \right) \left(\quad \mathbf{x} \quad \right) = \left(\quad \mathbf{b} \quad \right), \tag{B.7}$$

$$\underset{matrix}{A} \cdot \underset{vector}{\mathbf{x}} = \underset{vector}{\mathbf{b}}.$$

When does this system of linear equations admit a **unique solution**? The following theorem links this notion to the determinant $\det A$ or $|A|$, which is a scalar that packs as much information about the matrix as possible.

Theorem 3 *The system $A\mathbf{x} = \mathbf{b}$, with given A, has a unique solution $\mathbf{x}$ for any given vector $\mathbf{b}$* ***if and only if*** $\det A \neq 0$.

Computing the Determinant of a Matrix

For a 2×2 matrix, this is easy.

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc\,. \tag{B.8}$$

It gets a little uglier for a 3×3 matrix, but it is still manageable:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a \underbrace{(ei - fh)}_{\det \begin{pmatrix} e & f \\ h & i \end{pmatrix}} - d \underbrace{(bi - ch)}_{\det \begin{pmatrix} b & c \\ h & i \end{pmatrix}} + g \underbrace{(bf - ce)}_{\det \begin{pmatrix} b & c \\ e & f \end{pmatrix}} \tag{B.9}$$

(note the minus sign in front of d).

Important properties:

- $\det(AB) = \det(A)\det(B)$.
- $\det A^\top = \det A$ (transpose).
- $\det \alpha A = \alpha^n \det A$.

Theorem 4 (Cramer's rule) *Suppose* $\det A \neq 0$. *Then the solution* $\mathbf{x} = (x_1, \ldots x_n)$ *of* $A\mathbf{x} = \mathbf{b}$ *is given by:*

$$x_i = \frac{\det A_i}{\det A},$$

where A_i *is the matrix formed by replacing the* i^{th} *column of* A *by* b.

→ Very interesting in theory;

→ Not very useful in practice (computationally intensive).

Matrix Inverse

Definition

To solve for x when $A\mathbf{x} = \mathbf{b}$ (if $\det A \neq 0$), we introduce the inverse:

$$\mathbf{x} = A^{-1}\mathbf{b} \tag{B.10}$$

A^{-1} is a matrix such that:

$$A \cdot A^{-1} = A^{-1} \cdot A = I, \tag{B.11a}$$
$$\left(A^{-1}\right)^{-1} = A, \tag{B.11b}$$
$$\left(A^\top\right)^{-1} = \left(A^{-1}\right)^\top, \tag{B.11c}$$
$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}. \tag{B.11d}$$

We already know that this matrix exists if and only if $\det A \neq 0$.

The 2×2 *case*

For a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

whose determinant $|A| = ad - bc \neq 0$, we have the simple formula:

$$A^{-1} = \frac{1}{|A|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

As a simple application, consider the case of the bivariate normal density. We saw in Chapter 11 that the multivariate normal density Eq. (11.8) involves an exponential argument of the form $(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$, which we called the Mahalanobis distance. The reader is referred to Chapter 11, Sect. II for a full account. The bivariate independent case simplifies to something surprisingly lean Eq. (11.7), all thanks to the formula above.

For larger matrices this trick becomes impractical, so a common method is Gauss-Jordan elimination, which is simple, but takes quite a bit of practice. This is not where you will learn it: either take a proper linear algebra course, or try one of many excellent MOOCs on the topic.

Although there exist general, analytical solutions to compute matrix inverses, it turns out that for n larger than 3 they get quite ugly, and computationally just as intensive as computing the inverse by hand, so let's just tell a computer to do that and be done with it:

Numerical Solutions

1. Python (NumPy) has a ready-made function to invert matrices [6], but it is usually a terrible idea. One reason is that if the purpose of finding A^{-1} is only to multiply it by $\boldsymbol{b}$ to get the solution $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, then one is better served by solving for x directly[7].

[6] `np.linalg.inv()`

[7] `np.linalg.solve(A,b)`

2. One can also play with LU factorization (write $A = LU$[8], where L is the lower triangular and U is the upper triangular).

[8] `scipy.linalg.lu()`

$$LU\mathbf{x} = \mathbf{b}\,,$$
$$L\,(U\mathbf{x}) = \mathbf{b}\,.$$

Then solve $L\mathbf{y} = \mathbf{b}$, for $\mathbf{y} = U\mathbf{x}$.

There are many kinds of other matrix factorizations (Cholesky, QR, etc) which can become advantageous when A exhibits certain properties (e.g. symmetry). SciPy incorporates the whole LAPACK machinery to do so.

IV Linear Independence. Bases

Two vectors $\mathbf{v}$ and $\mathbf{w}$ are linearly independent if one cannot be expressed as a scaled version of the other. i.e. $\mathbf{v} \neq \alpha\boldsymbol{w}$, for any $\alpha \in \mathbb{R}$. More generally, $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are linearly independent if each $\mathbf{v}_i$ cannot be written as a linear combination of the other vectors *i.e.*

$$\mathbf{v}_1 \neq \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 + \ldots + \alpha_k\mathbf{v}_k\,, \tag{B.12a}$$
$$\mathbf{v}_2 \neq \alpha_1\mathbf{v}_1 + \alpha_3\mathbf{v}_3 + \ldots + \alpha_k\mathbf{v}_k\,, \tag{B.12b}$$
$$\text{etc}\ldots$$

Equivalently,

$$\alpha_1 \mathbf{v}_1 + \ldots + \alpha_k \mathbf{v}_k = 0 \Longleftrightarrow \alpha_1, \ldots, \alpha_k = 0\,.$$

(the right arrow is the more difficult one to prove; the left one is trivial).

That is, linearly dependent vectors are redundant: at least one can be expressed as a combination of the others, so there are fewer than k degrees of freedom in this system.

A **basis** of $\mathbb{R}^n$ is a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ such that:

- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are linearly independent
- Any vector $\mathbf{v}$ can be written as a linear combination of $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$

$$\mathbf{v} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \ldots + \lambda_k \mathbf{v}_k \qquad \text{(these vectors span the whole space)} \tag{B.13}$$

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a basis, then the decomposition Eq. (B.13) is unique.

Example

$\mathbf{e}_1 = (1,0,0)$, $\mathbf{e}_2 = (0,1,0)$ and $\mathbf{e}_3 = (0,0,1)$ form a basis of $\mathbb{R}^3$. Because these vectors are so intuitive, it is further called a **canonical**[9] basis.

[9] None of them got canonized, however

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \Longrightarrow \mathbf{v} = 1.\mathbf{e}_1 + 2.\mathbf{e}_2 + 3.\mathbf{e}_3$$

A basis of $\mathbb{R}^n$ has always **exactly** n vectors.

It is trivial to verify that with this choice of vectors, the only possible way for $\lambda_1 \mathbf{e}_1 + \lambda_2 \mathbf{e}_2 + \lambda_3 \mathbf{e}_3$ to be zero is if $\lambda_1 = \lambda_2 = \lambda_3 = 0$ (Linear independence). Since the space is 3-dimensional, these 3 linearly-independent vectors form a basis of it. QED

V Inner Products and Orthonormality

An **inner product** on $\mathbb{R}^n$ is a function $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, such that:

Bilinear $$\langle \alpha \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \; \langle \mathbf{x}, \gamma \mathbf{z} + \mathbf{w} \rangle = \gamma \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{x}, \mathbf{w} \rangle \tag{B.14a}$$

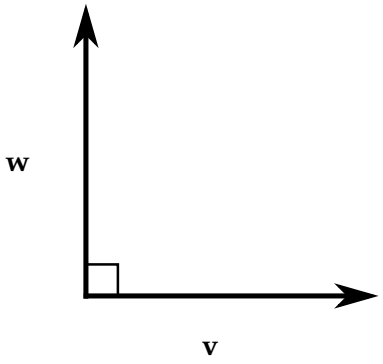
Symmetric $$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \tag{B.14b}$$

Positive definite $$\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \,. \text{ If } \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = 0 \tag{B.14c}$$

Example (Canonical Dot product in $\mathbb{R}^n$)

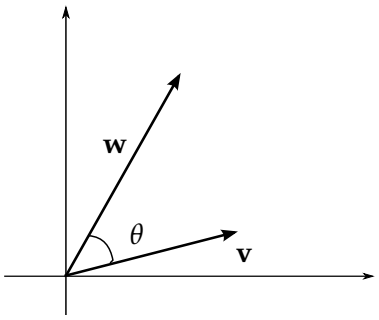
$$\begin{aligned} \langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v} \cdot \mathbf{w} &= v_1 w_1 + v_2 w_2 + \ldots + v_n w_n \\ &= \sum_i v_i w_i \,. \end{aligned}$$

Definition 16 (Orthogonality) *Two vectors* $\mathbf{v}$ *and* $\mathbf{w}$ *are said to be orthogonal (with respect to a given inner product* $\langle \cdot, \cdot \rangle$*) if* $\langle \mathbf{v}, \mathbf{w} \rangle = 0$.



A Classic Case: the Dot Product

$\mathbf{v} \cdot \mathbf{w} = \|\mathbf{v}\|_2 . \|\mathbf{w}\|_2 . \cos\theta$,where θ =smallest angle between $\mathbf{v}$ and $\mathbf{w}$, for $\mathbf{v} \cdot \mathbf{w} \neq 0$.



So $\mathbf{v} \cdot \mathbf{w} = 0$ if and only if $\cos\theta = 0$ *i.e.* $\mathbf{v}$ and $\mathbf{w}$ are perpendicular.

A basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is:

- **orthogonal** if
$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \,, \forall i \neq j \,.$$
- **orthonormal** if
$$\begin{aligned} &\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \,, \forall i \neq j \,, \\ &\qquad \& \\ &\langle \mathbf{v}_i, \mathbf{v}_i \rangle = \|\mathbf{v}_i\|^2 = 1 \,, \end{aligned}$$

where this norm is induced by the inner product. Clearly, orthonormality is a stronger constraint (orthonormality implies orthogonality).

Advantages of an Orthonormal Basis

$\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal basis. If $\mathbf{v}$ is a given vector, it may be written as a linear combination of basis vectors:

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \ldots + \alpha_n \mathbf{v}_n$$

Now that's true in any basis, so what we really want to know is how to find the coordinates (α_i's), and quickly. Let us project v onto v_i, $i \in \{1, \cdots, n\}$:

$$\langle \mathbf{v}, \mathbf{v}_i \rangle = \alpha_1 \langle \mathbf{v}_1, \mathbf{v}_i \rangle + \ldots + \alpha_n \langle \mathbf{v}_n, \mathbf{v}_i \rangle$$

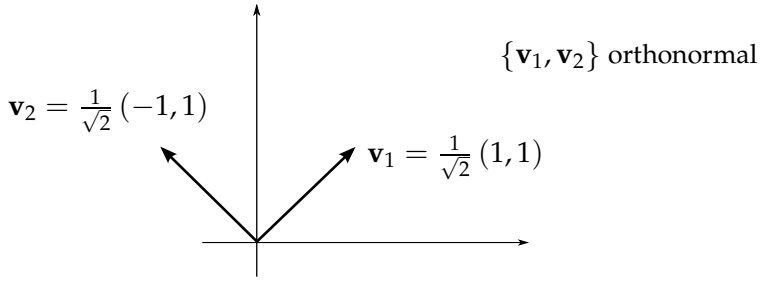
Of course, because the $\mathbf{v}_j$ are orthogonal, the cross terms in the sum drop out, so we are left with $\alpha_i = \langle \mathbf{v}, \mathbf{v}_i \rangle$.

That is, one can find the coordinates α_i by simple projection onto the basis vectors. This property is so intuitive that you probably take it for granted – but it is remarkable, and entirely due to orthonormality. The result is that we may write:

$$\mathbf{v} = \sum_i^n \langle \mathbf{v}, \mathbf{v}_i \rangle \mathbf{v}_i \tag{B.15}$$

As extra gravy, this formula enables to compute the norm of $\mathbf{v}$: $\|\mathbf{v}\| = \sum \langle \mathbf{v}, \mathbf{v}_i \rangle^2$, which is nothing more than Pythagoras' theorem in n dimensions.

Example ($\mathbb{R}^2$; dot product)



Question *Find the coordinates of* $\mathbf{v} = (1, 0)$ *with respect to the basis* $\{\mathbf{v}_1, \mathbf{v}_2\}$ *i.e. find* α_1 *and* α_2 *such that* $\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2$.

Solution *Since* $\{\mathbf{v}_1, \mathbf{v}_2\}$ *is an orthonormal basis,*

$$\mathbf{v} = \langle \mathbf{v}, \mathbf{v}_1 \rangle \mathbf{v}_1 + \langle \mathbf{v}, \mathbf{v}_2 \rangle \mathbf{v}_2 \,,$$

where

$$\mathbf{v} \cdot \mathbf{v}_1 = 1.\frac{1}{\sqrt{2}} + 0.\frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}} \,,$$
$$\mathbf{v} \cdot \mathbf{v}_2 = 1. - \frac{1}{\sqrt{2}} + 0.\frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}} = -\frac{1}{\sqrt{2}} \,.$$

Thus:

$$\mathbf{v} = \frac{1}{\sqrt{2}}\mathbf{v}_1 - \frac{1}{\sqrt{2}}\mathbf{v}_2 \,.$$

Let's check:

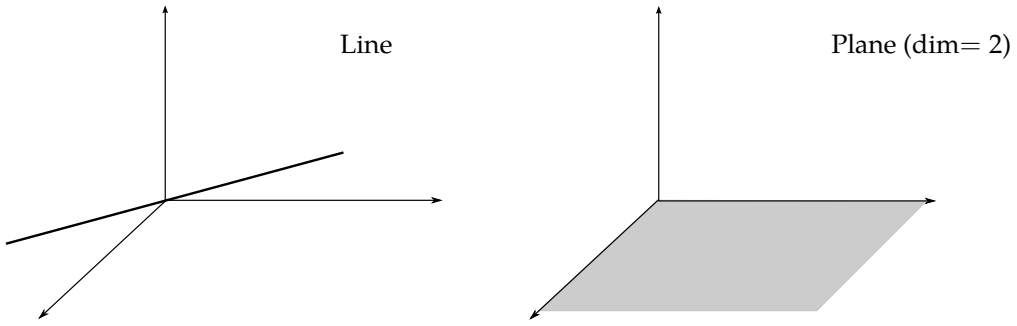
$$\frac{1}{\sqrt{2}}\mathbf{v}_1 - \frac{1}{\sqrt{2}}\mathbf{v}_2 = \frac{1}{2}\,(1,1) - \frac{1}{2}\,(-1,1) = (1,0) \;.$$

VI Projections

E is a **subspace** of $\mathbb{R}^n$ if:

- $\forall \mathbf{v} \in E, \qquad \alpha\mathbf{v} \in E$. (Invariant under scaling)
- $\forall \mathbf{v}, \mathbf{w} \in E, \mathbf{v} + \mathbf{w} \in E$. (Invariant under addition)

Example (in $\mathbb{R}^3$)

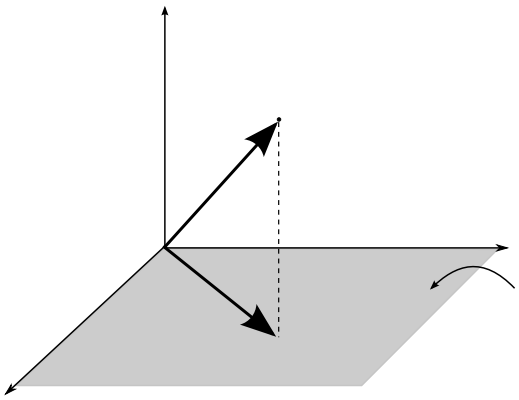


Projecting onto a Subspace

Idea: E a subspace and $\mathbf{v} \notin E$. Find a vector in E closest to $\mathbf{v}$ *i.e.* best approximation of $\mathbf{v}$ by a vector of E.

$$\mathrm{Im}A = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\} \;. \tag{B.16}$$

Example



Definition 17 *E is subspace of $\mathbb{R}^n$. The **projection of $\mathbf{v}$ on** E, denoted by $P_E(\mathbf{v})$, is the unique vector $P_E(\mathbf{v}) \in E$ such that:*

$$\|P_E(\mathbf{v}) - \mathbf{v}\| = \min_{\mathbf{w} \in E} \|\mathbf{v} - \mathbf{w}\| . \tag{B.17}$$

A Very Important Application: Least Squares

Problem: Solve $A\mathbf{x} = \mathbf{b}$.

- $\det A \neq 0 \Rightarrow$ Unique solution.
- $\det A = 0 \Rightarrow$ No solution or multiple solutions. What to do?

Image of $A = \{A\mathbf{v} : \mathbf{v} \in \mathbb{R}^n\} \equiv \mathrm{Im}(A)$

The image of A is a subspace of $\mathbb{R}^n$! For a matrix the image is also known as the **column space**. A solution exists if and only if $b \in \mathrm{Im}(A)$.

Least Squares:

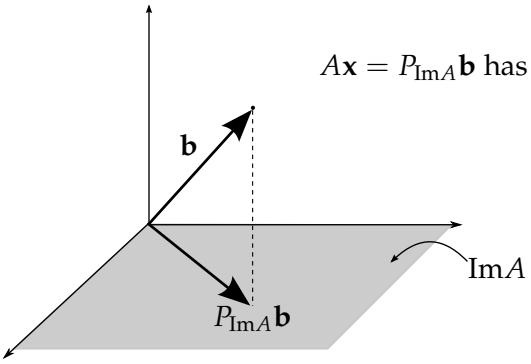
$$\min_x \|A\mathbf{x} - \mathbf{b}\| \Longrightarrow \qquad \text{Best solution} . \tag{B.18}$$

Equivalent to:

$$\min_{\mathbf{y} = A\mathbf{x}} \|\mathbf{y} - \mathbf{b}\| \tag{B.19}$$

$$\Leftrightarrow \min_{\mathbf{y} \in \mathrm{Im}(A)} \|\mathbf{y} - \mathbf{b}\| \tag{B.20}$$

So, $\mathbf{y} = P_{\mathrm{Im}(A)}\mathbf{b}$ is the solution of (B.18).



How can we compute a projection?

E subspace of $\mathbb{R}^n$.

$$\underbrace{\underbrace{\mathbf{v}_1, \ldots, \mathbf{v}_k}_{\text{orthonormal basis of E}}, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n}_{\text{orthonormal basis of } \mathbb{R}^n}$$

So, $\mathbf{v}$ can be decomposed in a basis of $\mathbb{R}^n$ as:

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \ldots + \alpha_n \mathbf{v}_n \,.$$

If $\mathbf{w} \in E$ *i.e.* $\mathbf{w} = \beta_1 \mathbf{v}_1 + \ldots + \beta_k \mathbf{v}_k$:

$$\min_{\mathbf{w} \in E} \|\mathbf{v} - \mathbf{w}\| = ?$$

Given that:

$$\mathbf{v} - \mathbf{w} = (\alpha_1 - \beta_1)\, \mathbf{v}_1 + \ldots + (\alpha_k - \beta_k)\, \mathbf{v}_k + \alpha_{k+1}\mathbf{v}_{k+1} + \ldots + \alpha_n \mathbf{v}_n \,,$$

we have:

$$\|\mathbf{v} - \mathbf{w}\|^2 = (\alpha_1 - \beta_1)^2 + \ldots + (\alpha_k - \beta_k)^2 + \alpha_{k+1}^2 + \ldots + \alpha_n^2 \,.$$

The minimum is obtained when $\beta_1 = \alpha_1$, $\beta_2 = \alpha_2$, ... and $\beta_k = \alpha_k$ *i.e.* when

$$\mathbf{w} = \alpha_1 \mathbf{v}_1 + \ldots + \alpha_k \mathbf{v}_k \,.$$

Therefore:

$$P_E(\mathbf{v}) = \alpha_1 \mathbf{v}_1 + \ldots + \alpha_k \mathbf{v}_k \,. \tag{B.21}$$

and

$$\mathbf{v} = \underbrace{\alpha_1 \mathbf{v}_1 + \ldots + \alpha_k \mathbf{v}_k}_{P_E(\mathbf{v})} + \alpha_{k+1}\mathbf{v}_{k+1} + \ldots + \alpha_n \mathbf{v}_n \tag{B.22}$$

[10] In Python, `np.linalg.lstsq(A, b)` solves the least square problem

The solution to the least square problem $A\mathbf{x}_{ls} = P_{\mathrm{Im}A}.\mathbf{b}$ satisfies[10]:

$$A^\top A \mathbf{x}_{ls} = A^\top \mathbf{b} \,.$$

So,

$$\mathbf{x}_{ls} = \left(A^\top A\right)^{-1} A^\top \mathbf{b} \tag{B.23}$$

Solving $A\mathbf{x} = \mathbf{b}$

A solution exists if and only if $\mathbf{b} \in \text{Im}A$, with A a $m \times n$ matrix.

Typically:

1. Unique solution $\Rightarrow\ m = n,\ \det A \neq 0.$
2. No solution $\Rightarrow\ m > n$ More equations than unknowns.
   $\rightarrow$ Least squares
3. Multiple solutions $\Rightarrow\ m < n$ Fewer equations than unknowns.

What to do? We need to make an assumption to obtain a unique solution. For example, we can choose to seek the solution with the smallest Euclidean (L^2) norm, which is the *least squares* solution.

VII Vector Spaces

Generalities

A vector space is a set V, where two operations:

1. Addition,
2. Multiplication by a scalar ($\mathbb{R}$ or $\mathbb{C}$),

are defined and satisfy the following properties:

Addition

P1a) $\mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3) = (\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3\,.$ (Associativity)

P1b) $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1\,.$ (Commutativity)

P1c) There exists $0 \in V$, such that $0 + \mathbf{v} = \mathbf{v} + 0\,.$ (Identity element)

P1d) $\forall \mathbf{v} \in V$, there exists $-\mathbf{v} \in V$, such that $\mathbf{v} + (-\mathbf{v}) = 0\,.$ (Inverse element)

Scalar multiplication

P2a) $\lambda\,(\mathbf{v}_1 + \mathbf{v}_2) = \lambda\mathbf{v}_1 + \lambda\mathbf{v}_2,\ \forall\,\lambda \in \mathbb{R}$ or $\mathbb{C}$ and $\forall\,\mathbf{v}_1, \mathbf{v}_2 \in V\,.$ (Distributivity w.r.t. vector addition)

P2b) $(\lambda_1 + \lambda_2)\,\mathbf{v} = \lambda_1\mathbf{v} + \lambda_2\mathbf{v},\ \lambda_1, \lambda_2 \in \mathbb{R}$ and $\mathbf{v} \in V\,.$ (Distributivity w.r.t. scalar addition)

P2c) $\lambda_1\,(\lambda_2)\,\mathbf{v} = (\lambda_1\lambda_2)\,\mathbf{v}\,.$ (Compatibility of vector and scalar mult.)

P2d) $\mathbb{1} \cdot \mathbf{v} = \mathbf{v},\ \forall\,\mathbf{v} \in V\,.$ (Identity element)

Example

1. $\mathbb{R}^n$.
2. $\mathbb{C}^n$.
3. $\mathcal{C}([a,b]) =$ *the set of continuous functions on the interval* $[a,b]$

   $(f+g)(x) = f(x) + g(x)$.
   $(\lambda f)(x) = \lambda f(x)$.
4. $\mathcal{L}^2(\mathbb{R})$, *the set of square integrable functions on* $\mathbb{R}$. *That is, all functions* f *such that:*
   $$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty .$$
5. $\mathcal{M}_n(\mathbb{R}) =$ *the space of* $n \times n$ *matrices with real coefficients.*

Theorem 5 *Every vector space has a basis.*

If # of basis $< \infty$ $\Rightarrow$ *finite dimensional* *Examples:* 1, 2 *and* 5.

If # of basis $\to \infty$ $\Rightarrow$ *infinite dimensional* *Examples:* 3 *and* 4.

Application to Fourier Series

Consider the space (*i.e.* E) of periodic functions on $[0, 2\pi]$ which satisfy:

$$\int_0^{2\pi} |f(\theta)|^2 d\theta < \infty , \tag{B.24}$$

An inner product on that space is given by:

$$\langle f, g \rangle = \int_0^{2\pi} f(\theta)\, g(\theta)\, d\theta . \tag{B.25}$$

Indeed it satisfies the properties of the inner product (B.14)

1. Linearity (B.14a)

$$\begin{aligned} \langle \alpha f + \beta g, h \rangle &= \int_0^{2\pi} (\alpha f + \beta g)\, h d\theta \\ &= \alpha \int_0^{2\pi} f h d\theta + \beta \int_0^{2\pi} g h d\theta \\ &= \alpha \langle f, h \rangle + \beta \langle g, h \rangle . \end{aligned} \tag{B.26a}$$

2. Symmetry (B.14b)

$$\langle f, g \rangle = \int_0^{2\pi} f g d\theta = \int_0^{2\pi} g f d\theta = \langle g, f \rangle .$$

3. Positive definiteness (B.14c)

$$\langle f, f\rangle = \int_0^{2\pi} |f(\theta)|^2 d\theta \geq 0. \tag{B.26b}$$
$$\langle f, f\rangle = 0 \Leftrightarrow \int_0^{2\pi} |f(\theta)|^2 d\theta = 0 \Rightarrow f = 0\,.$$

The norm induced by the inner product is

$$\|f\| = (\langle f, f\rangle)^{\frac{1}{2}} = \left(\int_0^{2\pi} |f(\theta)|^2 d\theta\right)^{\frac{1}{2}}. \tag{B.27}$$

Now consider the following set of functions:

$$\{1\} \cup \{\cos(n\theta)\}_{n=1}^{\infty} \cup \{\sin(n\theta)\}_{n=1}^{\infty}\,, \tag{B.28}$$

i.e. $\{1, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots\}$. These functions are vectors in E. Do they form a *basis* of it?

One can check that:

$$\int_0^{2\pi} \cos n\theta \cos m\theta d\theta = 0\,, \qquad \text{for } n \neq m\,. \tag{B.29a}$$
$$\int_0^{2\pi} \sin n\theta \sin m\theta d\theta = 0\,, \qquad \text{for } n \neq m\,. \tag{B.29b}$$
$$\int_0^{2\pi} \sin n\theta \cos m\theta d\theta = 0\,, \qquad \text{for all } n, m\,. \tag{B.29c}$$

In vector space, this translates to:

$$\langle \cos n\theta, \cos m\theta\rangle = 0\,, \qquad \text{for } n \neq m\,. \tag{B.30a}$$
$$\langle \sin n\theta, \sin m\theta\rangle = 0\,, \qquad \text{for } n \neq m\,. \tag{B.30b}$$
$$\langle \sin n\theta, \cos m\theta\rangle = 0\,, \qquad \text{for all } n, m\,. \tag{B.30c}$$

Also,

$$\int_0^{2\pi} (\cos n\theta)^2 d\theta = \pi\,, \qquad \text{for } n \geq 1\,. \tag{B.31a}$$
$$\int_0^{2\pi} (\sin n\theta)^2 d\theta = \pi\,, \qquad \text{for } n \geq 1\,. \tag{B.31b}$$
$$\int_0^{2\pi} 1^2 d\theta = 2\pi\,. \tag{B.31c}$$

$$\Longrightarrow \|\cos n\theta\|^2 = \langle \cos n\theta, \cos n\theta\rangle = \pi\,, \qquad \text{for } n \geq 1\,. \tag{B.32a}$$
$$\|\sin n\theta\|^2 = \langle \sin n\theta, \sin n\theta\rangle = \pi\,, \qquad \text{for } n \geq 1\,. \tag{B.32b}$$
$$\|1\|^2 = 2\pi\,. \tag{B.32c}$$

Therefore,

$$\left\{\frac{1}{\sqrt{2\pi}}\right\} \cup \left\{\frac{1}{\sqrt{\pi}}\cos(n\theta)\right\}_{n=1}^{\infty} \cup \left\{\frac{1}{\sqrt{\pi}}\sin(n\theta)\right\}_{n=1}^{\infty}, \tag{B.33}$$

is an orthonormal set of functions in E.

Let $f \in E$:

$$\langle f, \frac{1}{\sqrt{2\pi}} \rangle = \int_0^{2\pi} f(\theta) \frac{1}{\sqrt{2\pi}} d\theta =: a_0 \,. \tag{B.34a}$$

$$\langle f, \frac{1}{\sqrt{2\pi}} \cos n\theta \rangle = \int_0^{2\pi} f(\theta) \frac{1}{\sqrt{2\pi}} \cos n\theta d\theta =: a_n \,, n > 0 \,. \tag{B.34b}$$

$$\langle f, \frac{1}{\sqrt{2\pi}} \sin n\theta \rangle = \int_0^{2\pi} f(\theta) \frac{1}{\sqrt{2\pi}} \sin n\theta d\theta =: b_n \,, n > 0 \,. \tag{B.34c}$$

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$, then $\mathbf{v} = \sum \langle \mathbf{v}, \mathbf{v}_n \rangle \mathbf{v}_n$, so we can express f as a linear combination of these functions with coefficients given by the inner product (projection) on each basis function:

$$f(\theta) = \frac{a_0}{\sqrt{2\pi}} + \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} (a_n \cos n\theta + b_n \sin n\theta) \qquad \text{(Fourier Series Representation)} \,.$$

That is, can we represent any function $f \in E$ as a linear combination of the functions Eq. (B.33)?

By working harder, one can prove that if,

$$P_N(\theta) = \frac{a_0}{\sqrt{2\pi}} + \frac{1}{\sqrt{\pi}} \sum_{n=1}^{N} (a_n \cos n\theta + b_n \sin n\theta) \,, \tag{B.35}$$

then, $\|f - P_N\| \to 0$ as $N \to \infty$ (the approximation is optimal in the sense of the $\mathbb{L}^2$ norm).

Which answers the question by a resounding yes (yes in the $\mathbb{L}^2$ sense): provided we add an infinity of sines and cosines, we can represent any periodic function this way. For some functions, just a few terms will suffice, but that will not be true in general, leading to phenomena like Gibb's oscillations (Lab 7). In other words, the space of periodic functions is a vector space, but it is of infinite dimension, which makes low-dimensional approximations challenging.

Appendix C

CIRCLES AND SPHERES

I Trigonometry

Trigonometric functions

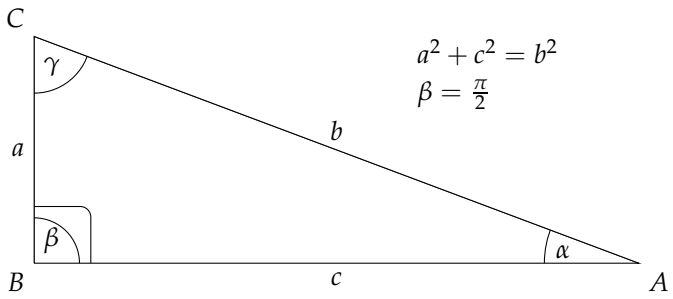


Figure C.1: Angle relations in the triangle

Sine

$$\sin(\alpha) = \frac{a}{b}$$
$$\alpha = \arcsin\left(\frac{a}{b}\right) = \sin^{-1}\left(\frac{a}{b}\right)$$

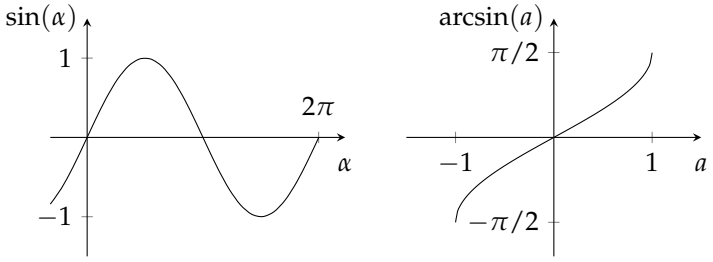


Figure C.2: Sine and arcsine functions

Cosine

$$\cos(\alpha) = \frac{c}{b}$$
$$\alpha = \arccos\left(\frac{c}{b}\right)$$

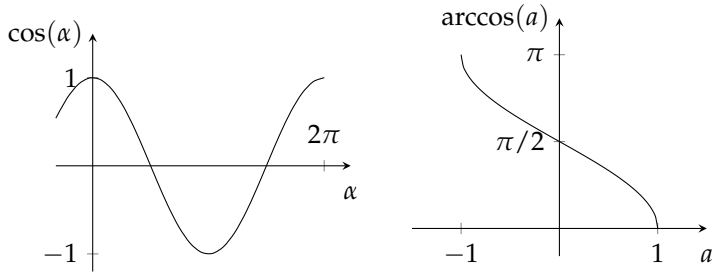


Figure C.3: Cosine and arccosine functions

Note *Angles are to be given in radians:*

$$\alpha[degrees] = \alpha[rad] \cdot \frac{180}{\pi}$$

Tangent

$$\tan(\alpha) = \frac{a}{c} \tag{C.1}$$
$$\alpha = \arctan\left(\frac{a}{c}\right) \tag{C.2}$$

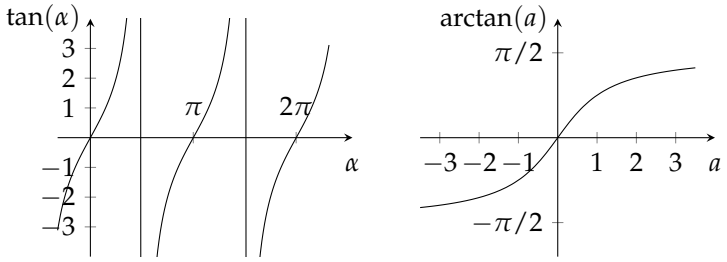


Figure C.4: Tangent and arctangent functions

Trigonometric relationships

$$\tan(\alpha) = \frac{\sin(\alpha)}{\cos(\alpha)} \tag{C.3}$$
$$\sin^2(\alpha) + \cos^2(\alpha) = 1 \tag{C.4}$$
$$\sin(\alpha) = \pm\sqrt{1 - \cos^2(\alpha)}; \qquad \cos(\alpha) = \pm\sqrt{1 - \sin^2(\alpha)} \tag{C.5}$$

Symmetry

$$\sin(-\alpha) = -\sin(\alpha) \tag{C.6}$$
$$\cos(-\alpha) = \cos(\alpha) \tag{C.7}$$
$$\tan(-\alpha) = -\tan(\alpha) \tag{C.8}$$

Periodicity

$$\sin\left(\alpha + \frac{\pi}{2}\right) = \cos(\alpha) \tag{C.9}$$
$$\sin(\alpha + \pi) = -\sin(\alpha) \tag{C.10}$$
$$\sin(\alpha + 2\pi) = \sin(\alpha) \tag{C.11}$$

etc.

Angle sum

$$\sin(\alpha \pm \beta) = \sin(\alpha)\cos(\beta) \pm \cos(\alpha)\sin(\beta) \tag{C.12}$$
$$\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta) \tag{C.13}$$

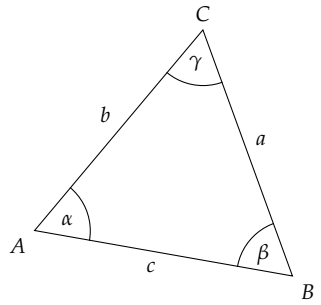


Figure C.5: Arbitrary triangle

Arbitrary triangles

Law of sines

$$c^2 = a^2 + b^2 - 2ab\cos(\gamma) \tag{C.14}$$

Law of cosines

$$\frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \tag{C.15}$$

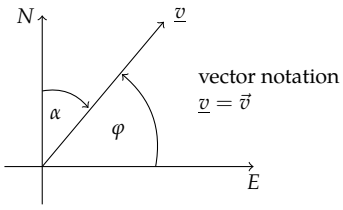


Figure C.6: Vector components

Example (Vector components)

$$v_E = |\underline{v}| \cdot \sin(\alpha)$$
$$v_N = |\underline{v}| \cdot \cos(\alpha)$$
$$|\underline{v}| = \sqrt{v_E^2 + v_N^2}$$
$$\alpha = \underbrace{\arctan\left(\frac{v_E}{v_N}\right)}_{resultin[0,\pi]} \overset{numerically!}{=} \underbrace{\arctan 2(v_E, v_N)}_{resultin[\pi,\pi]}$$

Example (Spherical coordinates)

Spherical coordinates: (τ, θ, φ) *v.s.* (x, y, z)

$$\theta = co\text{-}latitude \in [0; \pi] \tag{C.16}$$
$$\lambda = latitude = \pi - \theta \in \left[-\frac{\pi}{2}; \frac{\pi}{2}\right] \tag{C.17}$$
$$\varphi = longitude \in [0; 2\pi[\tag{C.18}$$

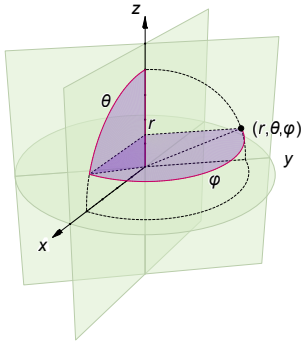


Figure C.7: Spherical coordinates

$$r = \sqrt{x^2 + y^2 + z^2} \qquad x = \tau \cos(\varphi) \sin(\theta)$$
$$\theta = \arccos\left(\tfrac{z}{r}\right) \qquad y = r \sin(\varphi) \sin(\theta)$$
$$\varphi = \arctan 2(y, x) \qquad z = r \cos(\theta)$$

Note that:

- *basis vectors $\underline{e}_r$, $\underline{e}_\theta$, $\underline{e}_\varphi$ are $f(\theta, \varphi)$; $\underline{e}_\varphi$ is undefined for $\theta = 0$ or $\theta = \pi$;*
- *in a Cartesian system:*

$$\underline{e}_r = \begin{pmatrix} \sin(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) \\ \cos(\theta) \end{pmatrix} \quad \underline{e}_\theta = \begin{pmatrix} \cos(\theta)\cos(\varphi) \\ \cos(\theta)\sin(\varphi) \\ -\sin(\varphi) \end{pmatrix} \quad \underline{e}_\varphi = \begin{pmatrix} \sin(\varphi) \\ \cos(\varphi) \\ 0 \end{pmatrix};$$

- *for integration, a surface element is:*

$$dA = r^2 \sin(\theta) d\varphi d\theta$$

Likewise for volume:

$$dV = r^2 \sin(\theta) dr d\theta d\varphi$$

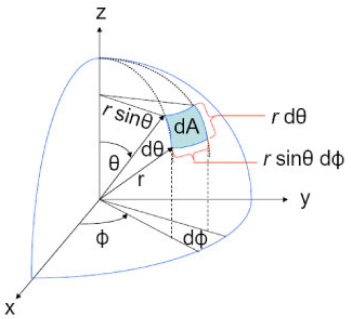


Figure C.8: Surface integration on a sphere

- *the gradient in spherical coordinates is:*

$$\underline{\nabla} f(r, \theta, \varphi) = \frac{\partial f}{\partial r}\underline{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\underline{e}_\theta + \frac{1}{r\sin(\theta)}\frac{\partial f}{\partial \varphi}\underline{e}_\varphi$$

Horizontal surface gradient ($r = 1$)

$$\underline{\nabla}_h = \begin{pmatrix} \partial_\theta \\ \frac{1}{\sin(\theta)}\partial_\varphi \end{pmatrix}$$

Example (Distance on sphere)
Given two points with longitude φ and latitude λ (φ, λ) and (φ_2, λ_2), find the great circle distance s in radians:

$$s = \arccos\left(\sin(\lambda), \sin(\lambda_2) + \cos(\lambda), \cos(\lambda_2)\cos(\varphi_1 - \varphi_2)\right)$$

This is a poor formula numerically for small s (why?), much better to use:

$$s = \arcsin\left[\left(\sin^2\left(\frac{\lambda_1 + \lambda_2}{2}\right) + \cos(\lambda), \cos(\lambda_2)\sin^2\left(\frac{\varphi_1 - \varphi_2}{2}\right)\right)^{\frac{1}{2}}\right]$$

Distance in m assuming sphere is:

$$d = R \cdot s, \qquad \text{with } R = 6371 \cdot 10^3 m \text{ for Earth}$$

For more, see the Aviation Formulary, by Ed Williams.

Example (Average directions and orientation)

1. *Given a set of n unit vectors, find the mean azimuth α:*

$$\langle \alpha \rangle = \arctan 2 \left(\frac{\sum v_i^E}{N}, \frac{\sum v_i^N}{N} \right)$$

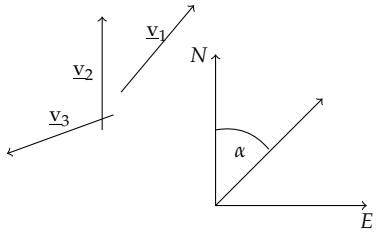


Figure C.9: Mean azimuth of unit vectors

2. *Given a set of N π-periodic orientations, find the mean azimuth:*

$$\alpha_i = \arctan 2 \left(o_i^E, o_i^N \right)$$
$$\tilde{o}_i^E = \sin(2\alpha_i)$$
$$\tilde{o}_i^N = \cos(2\alpha_i)$$
$$\langle \alpha \rangle_\pi = \frac{1}{2} \arctan 2 \left(\frac{\sum \tilde{o}_i^E}{N}, \frac{\sum \tilde{o}_i^N}{N} \right)$$

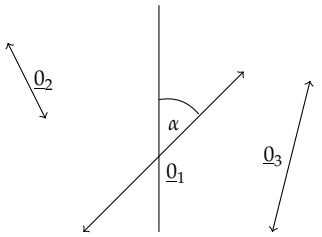
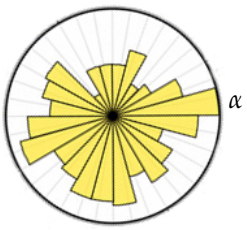
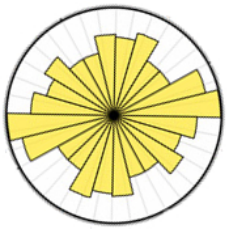


Figure C.10: Mean azimuth of π-periodic orientations

3. *Rose diagrams:* $A = \frac{\Delta\alpha}{2} r^2$



αr length



αA scaled

Figure C.11: Rose diagrams

II Complex Numbers

The Cardano-Tartaglia Equation

In sixteenth century Italy, there was a lot of interest in solving certain polynomial equations, in particular cubics[1]. Girolamo Cardano noticed something strange when he applied his formula to certain cubics. When solving $x^3 = 15x + 4$ he obtained an expression involving $\sqrt{-121}$. Cardano knew that you could not take the square root of a negative number yet he also knew that $x = 4$ was a solution to the equation. He wrote to Niccolo Tartaglia, another algebrist of the time, in an attempt to clear up the difficulty. Tartaglia certainly did not understand. In *Ars Magna*, Cardano's claims to fame, he gives a calculation with "complex numbers" to solve a similar problem but he really did not understand his own calculation which he says is "as subtle as it is useless."

Later it was recognized that if one just allowed oneself to write $i = \sqrt{-1}$, then $\sqrt{-121}$ would just be $i11$. If one could just bear the notion that i was a permissible thing to write, it turned out, a polynomial of

[1] this was generally in order to compute financial gains, so the research in this area of mathematics was rather handsomely subsidized by the nascent banking industry

Figure C.12: Gerolamo Cardano

degree n would always admit n solutions[2] – but the solutions would be "halfway between being and nothingness" (Leibniz), because they would involve this imaginary number i, which had no basis in reality.

[2] this is known as the Fundamental Theorem of Algebra

Figure C.13: Niccolò Tartaglia

Life in the complex plane

Complex Algebra

Liebniz's characterization was rather prescient, it turns out. Today we define a complex number z as

$$z = x + iy \tag{C.19}$$

with (x, y) two real numbers and $i = \sqrt{-1}$ as before (that is, $i^2 = -1$). We call x the real part ($\Re(z)$) and y the imaginary part ($\Im(z)$), so it really is "halfway between being and nothingness". That not does make it useless. In fact, it is applicable to so many areas of physics, mathematics & engineering that many people consider complex numbers to be just as "real" as real numbers. Excepts for integers, all of them are mental constructs anyway, so we might as well use mental constructs that can solve an astounding variety of problems. Here we will mostly use them to understand cyclicities.

A complex number may be charted in the *complex plane* (Fig. C.14), which is oriented by the real axis along $(1, 0)$ and the imaginary axis along $(0, i)$. Let us also define the define the *complex conjugate* of z, denoted $\bar{z}$ or z^*:

$$z^* = x - iy \tag{C.20}$$

That is, z^* is in the image of z by reflection along the real axis.

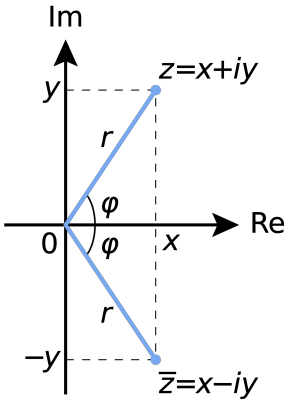


Figure C.14: Polar representation and complex conjugate. Two conjugates have the same modulus but opposite arguments. Credit: Wikipedia

If a complex number is just a point on a plane, why not just call it a vector of $\mathbb{R}^2$? In $\mathbb{R}^2$ we know how to add two vectors, or multiply them by a scalar. We even have an inner product (that turns two vectors into one scalar) and an outer product (that turns two vectors into a vector perpendicular to that plane, so outside the original space): in $\mathbb{R}^2$ there is no rule to multiply two elements and yet stay in $\mathbb{R}^2$. The space of complex numbers $\mathbb{C}$ does have such a rule. In addition to the usual properties:

$$z + z' = (x + x') + i(y + y') \quad \text{(addition)} \tag{C.21a}$$

$$\lambda z = \lambda x + i\lambda y \quad \text{(scalar multiplication)} \tag{C.21b}$$

there is an inner multiplication:

$$zz' = (xx' - yy') + i(x'y + y'x) \tag{C.22}$$

which, needless to say, is also in $\mathbb{C}$. Because of this, $\mathbb{C}$ is more like a richer version of $\mathbb{R}$, one that involves two coordinates but is endowed

with the same two rules of addition and multiplication. And while a polynomial of degree n may not always have roots in $\mathbb{R}$, it always has n roots in $\mathbb{C}$ – that is totally wild.

Polar representation

It is often extremely useful to represent z in terms of its distance to the origin (which we term the *modulus*, and denote $|z|$), and an angle ϕ, called its *argument*. The modulus verifies:

$$|z| = \sqrt{x^2 + y^2} = \sqrt{zz^*} \tag{C.23}$$

And the argument may be written (for any non-zero complex z):

$$\varphi = \arctan\left(\frac{y}{x}\right) \tag{C.24}$$

Any complex number (except zero) may be written in *polar form*:

$$z = re^{i\varphi} \tag{C.25}$$

where e is Euler's number as usual. What's magical about this is Euler's formula:

$$e^{i\varphi} = \cos\varphi + i\sin\varphi \tag{C.26}$$

so

$$z = r(\cos\varphi + i\sin\varphi) \tag{C.27}$$

Setting $r = 1$ defines the *unit circle* (Fig. C.17), numbers whose real part is given by $\cos\varphi$ and imaginary part given by $\sin\varphi$. Conversely, one may define sines and cosines this way:

$$\cos\varphi = \frac{e^{i\varphi} + e^{-i\varphi}}{2} \tag{C.28a}$$

$$\sin\varphi = \frac{e^{i\varphi} - e^{-i\varphi}}{2i} \tag{C.28b}$$

For the operations of multiplication, division, and exponentiation of complex numbers, it is generally much simpler to work with complex numbers expressed in polar form rather than rectangular form. From the laws of exponentiation, for two numbers $z_1 = r_1 e^{i\varphi_1}$ and $z_2 = r_2 e^{i\varphi_2}$:

Multiplication : $r_1 e^{i\varphi_1} \cdot r_2 e^{i\varphi_2} = r_1 r_2 e^{i(\varphi_1+\varphi_2)}$ (Fig. C.16)

Division : $\frac{r_1 e^{i\varphi_1}}{r_2 e^{i\varphi_2}} = \frac{r_1}{r_2} e^{i(\varphi_1-\varphi_2)}$

Exponentiation $(re^{i\varphi})^n = r^n e^{in\varphi}$

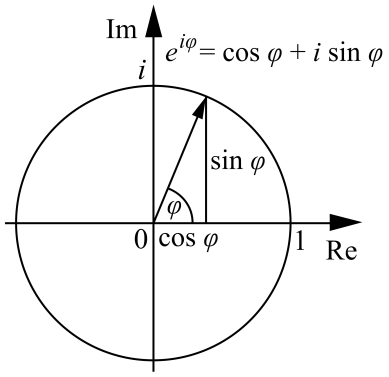


Figure C.15: Euler's formula. Credit: Wikipedia

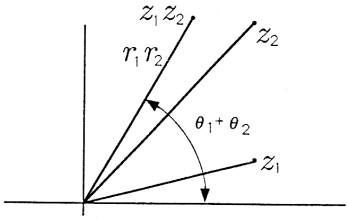


Figure C.16: Complex multiplication

the latter is known as De Moivre's formula, and explains a whole lot of trigonometric relations without lifting a finger. For instance, it explains $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\sin\theta\cos\theta$.

The notation also allows to re-express many famous numbers in terms of their argument. For instance,

$$\begin{aligned} 1 &= e^{i0} \\ -1 &= e^{i\pi} \\ i &= e^{i\pi/2} \\ -i &= e^{i3\pi/2} \end{aligned}$$

because who wouldn't want to use three symbols (two irrational numbers and an imaginary one) to write the number one?

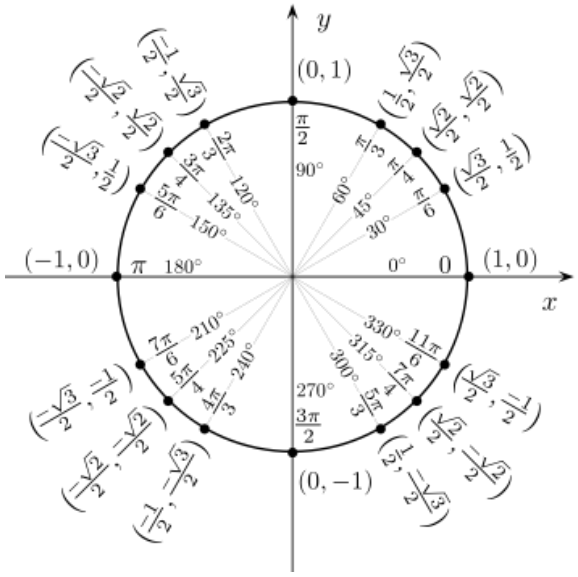


Figure C.17: The unit circle, defined as all complex numbers with unit modulus.

Further, one quickly notices that this notation is everything but unique. Indeed adding any multiple of 2π means going around the merry-go-round that many times, so we say that a complex argument is defined "modulo 2π". Hence we should have written $1 = e^{ik2\pi}$, where k is any positive or negative integer (*i.e.* , $k \in \mathbb{Z}$), and so on for the others. This "modulo" business simply expresses periodicity.

Roots of Unity

An n^{th} *root of unity*, where $n \in \mathbb{N}^*$, is a number z satisfying the deceitfully simple equation

$$z^n = 1 \tag{C.29}$$

Now, using the polar notation $z = re^{i\varphi}$, we have by De Moivre's formula:

$$r^n e^{in\varphi} = 1 \tag{C.30}$$

Two complex numbers are equal if and only if they have the same modulo and argument, so for this to work, $r^{1/n} = 1$, yielding $r = 1$: the solution must be on the unit circle (Fig. C.17). What about its argument? It must verify:

$$e^{i\varphi} = \left(e^{i2\pi k}\right)^{1/n} \tag{C.31}$$

Hence $\varphi = \frac{2\pi k}{n}$, $k \in \mathbb{Z}$. Now, $\mathbb{Z}$ is infinite, but clearly we can't have an infinity of solutions: every time we go around the the merry-go-round (*i.e.* when k is any multiple of n), we get back where we started. It is easy to show that they can be only n distinct solutions, called the primitive n^{th} roots of unity, verifying

$$\boxed{W_n^k = e^{i2\pi k/n}, \quad k \in \{0, \cdots, n-1\}} \tag{C.32}$$

These are some of the coolest numbers you'll ever meet. The third roots of unity are illustrated in Fig. C.18, and you can see that they bisect the unit circle in three equal slices. In general, n^{th} roots of unity split the unit circle in n equal slices[3]: they are *cyclotomic*. The roots are always located at the vertices of a regular n-sided polygon inscribed in the unit circle, with one vertex at 1.

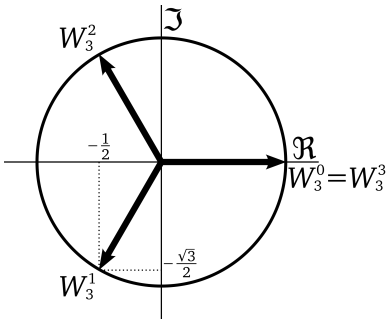


Figure C.18: Third roots of unity

[3] This proves invaluable in cutting pies without a protractor, *i.e.* in most social circumstances

Beyond their assistance in cutting pies, their periodicity makes them a cornerstone of Fourier analysis. Indeed, the sequence of powers

$$\cdots, W_n^{-1}, W_n^0, W_n^1, \cdots \tag{C.33}$$

is n-periodic (because $W_n^{j+n} = W_n^j \cdot W_n^n = W_n^j \cdot 1 = W_n^j$ for all values of j), and the n sequences of powers

$$s_k : \cdots, W_n^{k\cdot(-1)}, W_n^{k\cdot 0}, W_n^{k\cdot 1}, \cdots \tag{C.34}$$

for $k \in \{1, \cdots, n\}$ are all n-periodic (because $W_n^{k(j+n)} = W_n^{kj}$). Even more powerful, the set $\{s_1, \cdots, s_n\}$ of these sequences is a basis of the linear space of all n-periodic sequences. This means that any n-periodic sequence of complex numbers

$$\cdots, z_{-1}, z_0, z_1, \cdots \tag{C.35}$$

can be expressed as a linear combination of powers of a primitive n^{th} root of unity:

$$z_j = \sum_{k=0}^{n-1} Z_k \cdot W_n^{k\cdot j} = Z_1 \cdot W_n^{1\cdot j} + \cdots + Z_n \cdot W_n^{n\cdot j} \tag{C.36}$$

for some complex numbers $\{Z_1, \cdots, Z_n\}$ and every integer j. This is a form of Fourier analysis. If j is a (discrete) time variable, then k is a

frequency and Z_k is a complex amplitude. Choosing for the primitive n^{th} root of unity:

$$W = e^{i2\pi/n} = \cos\left(\frac{2\pi}{n}\right) + i\sin\left(\frac{2\pi}{n}\right) \tag{C.37}$$

allows z_j to be expressed as a linear combination of cos and sin functions:

$$\Re(z_j) = \sum_{k=0}^{n-1} A_k \cos\left(k\frac{2\pi j}{n}\right) + B_k \sin\left(k\frac{2\pi j}{n}\right) \tag{C.38}$$

where (A_k, B_k) are sequences of real numbers. This is a *discrete Fourier transform*, without which no modern telecommunication system would exist. The goal of Fourier analysis is to find those sequences (rapidly), and this is done at length in Chapter 7.

III Spherical harmonics

The harmonic equation

If you ever studied mechanics, you've doubtless encountered the harmonic equation, describing the small oscillations of a mass on a spring, or a pendulum, about their equilibrium position:

$$\ddot{x} + {\omega_0}^2 x = 0 \tag{C.39}$$

where ω_0 is a constant (in the case of the spring, ${\omega_0}^2$ is the ratio of its stiffness to the mass of the mobile). You may have seen that the solutions to such equations take the form

$$x(t) = A\cos(\omega_0 t + \phi) \quad \text{or;} \tag{C.40a}$$

$$x(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t) \tag{C.40b}$$

where A, B and ϕ are constants determined by the initial conditions. (the two formulations above are equivalent). A and B are called *amplitudes*, and ϕ is a *phase* (in radians).

Hence, sines and cosines are solutions to the harmonic equation, and are therefore called *harmonic functions*. In fact, many usual functions may be found as the solutions to well-known differential equations such as Eq. (C.39), which arise very often in physics, especially in the study of oscillations and vibrations. Spherical harmonics are a generalization of sines and cosines on a sphere.

Harmonics on the sphere

Solutions of Laplace's equation (aka the harmonic equation) describing free oscillations in a spherical medium:

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial f}{\partial\theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 f}{\partial\varphi^2} = 0$$

Eigenfunctions of orbital angular momentum operator in quantum mechanics (orbitals), normal modes of seismology, and form a natural set of orthonormal basis functions on the sphere.

$$Y_{lm} = \begin{cases} \sqrt{2}N_{lm}P_{lm}(\cos(\theta))\cos(m\varphi) & \text{for } m \geq 0 \\ \sqrt{2}N_{lm}P_{lm}(\cos(\theta))\sin(m\varphi) & \text{for } m < 0 \end{cases}$$

where l =degree, m =order and $m \in [-l; l]$.

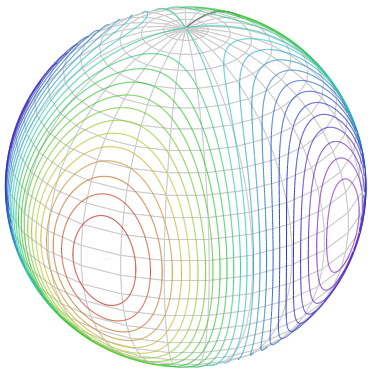


Figure C.19: The Y_{lm} spherical harmonic for $l = 2$, $m = 2$

Associated Legendre functions

$$P_{lm}(x) = (-1)^m(1-x^2)^{\frac{m}{2}}\frac{d^m}{dx^m}(P_l(x))$$
$$P_l(x) = \frac{1}{2^l l!}\frac{d^l}{dx^l}\left((x^2-1)^l\right)$$

where $P_{00} = 1$, $P_{10} = 0$, $P_{20} = \frac{1}{2}(3x^2-1)$, $P_{22} = 3(1-x^2)$.

Normalization

$$I = \int_0^\pi d\theta \int_0^{2\pi} d\varphi Y_{lm}Y_{l'm}, \quad \rightarrow \quad N_{lm} = \overset{\text{orthonormality}}{\sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}} \qquad \sin(\theta) = \delta_{ll}, \delta_{mm}$$

where δ =Kronecker δ, $\delta_{i,j} = \begin{cases} 1 \text{ for } i{=}j \\ 0 \text{ for } i{\neq}j \end{cases}$, used in physics and seismology.

In geodesy,

$$I = 4\pi\delta_{ll'}\delta_{mm'} \quad \rightsquigarrow \quad N^y_{lm} = \sqrt{(2l+1)\frac{(l-m)!}{(l+m)!}}.$$

In magnetics,

$$I = \frac{4\pi}{2l+1}\delta_{ll'}\delta_{mm'} \quad \rightsquigarrow \quad N^m_{lm} = \sqrt{\frac{(l-m)!}{(l+m)!}}.$$

An illustration is offered in Fig. C.19. For more extensive visualizations, see http://geodynamics.usc.edu/~becker/teaching-sh.html.

Spherical harmonics expansions (SHE)

In analogy to Fourier Transform in 2D, one can use Spherical Harmonics to convert a field given on the sphere to a harmonic representation:

$$f(\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} f_{lm} Y_{lm}(\theta,\varphi) \tag{C.41}$$

This holds for $\mathcal{L}^2$ norm convergence:

$$\lim_{L\to\infty} \int_0^{2\pi} d\varphi \int_0^{\pi} d\theta \left| f(\theta,\varphi) - \sum_{l=0}^{L} f_{lm} Y_{lm}(\theta,\varphi) \right|^2 \sin(\theta) = 0$$

Note *Compare* $|\underline{v}|_2 = \sqrt{\sum v_i^2}$ *with the expression above.*

Lab 1 illustrates how to obtain spherical harmonics expansion coefficients via numerical integration.

$$f_{em} = \int_{\Omega} d\Omega f(\theta,\varphi) Y_{lm}(\theta,\varphi) d\Omega = \int_0^{2\pi} d\varphi \int_0^{\pi} d\theta \sin(\theta) f(\theta,\varphi) Y_{lm}(\theta,\varphi)$$

Once fields are expressed as spherical harmonics, on can easily compute derived quantities like its power spectrum, derivatives, integrals, etc.

Appendix D

DIAGONALIZATION

Diagonal matrices are exceedingly easy to manipulate (multiply, invert, compute powers, etc). Diagonalization is all about transforming a non-diagonal matrix to diagonal form, to take advantage of these superpowers. This is used abundantly in Chapters 12 through 15.

I Earth Science Motivation

Example (Radioactive decay chain)

Consider the simple decay of a radioactive element to a stable daughter element, (e.g. $^{87}Rb \rightarrow {}^{87}Sr$). In this case, the number of atoms as a function of time $N(t)$ is given by

$$\frac{dN(t)}{dt} = -\lambda N \tag{D.1}$$

where λ is a constant.

We can also have a more involved case, e.g. $Bk \xrightarrow{\lambda_1} Bz \xrightarrow{\lambda_2} J \xrightarrow{\lambda_3} Rn$. In this case the numbers of atoms as a function of time are given by

$$\begin{aligned}
\frac{dN_1(t)}{dt} &= -\lambda_1 N_1 \\
\frac{dN_2(t)}{dt} &= -\lambda_2 N_2 + \lambda_1 N_1 \\
\frac{dN_3(t)}{dt} &= -\lambda_3 N_3 + \lambda_2 N_2
\end{aligned} \tag{D.2}$$

and we can rewrite this system of differential equations as

$$\frac{d}{dt}\begin{pmatrix} N_1(t) \\ N_2(t) \\ N_3(t) \end{pmatrix} = \underbrace{\begin{pmatrix} -\lambda_1 & 0 & 0 \\ \lambda_1 & -\lambda_2 & 0 \\ 0 & \lambda_2 & -\lambda_3 \end{pmatrix}}_{A} \begin{pmatrix} N_1(t) \\ N_2(t) \\ N_3(t) \end{pmatrix}. \tag{D.3}$$

After n time-steps the solution to the system of equations is given by the n^{th} power of the matrix A, A^n. If we could diagonalise A, that is find a matrix Λ such that

$$A = V\Lambda V^{-1} \tag{D.4}$$

then we would have

$$A^n = V\Lambda^n V^{-1} \tag{D.5}$$

and the problem would be solved.

II Eigendecomposition

Eigenvalues and eigenvectors

Given the matrix A, λ is an *eigenvalue* of A if and only if $\exists\ v \neq 0$ such that $Av = \lambda v$. The case $v = 0$ is trivial since it's always true. We can rewrite $Av = \lambda v$ as follows

$$\begin{aligned} &Av = \lambda v \Leftrightarrow Av - \lambda v = 0 \Leftrightarrow (A - \lambda I)v = 0 \Leftrightarrow \\ &v \in \mathcal{N}(A) \text{ (null space of } A\text{).} \end{aligned} \tag{D.6}$$

$\mathcal{N}(A)$ is non-trivial only if the matrix $(A - \lambda I)$ is singular, i.e.

$$det(A - \lambda I) = P(\lambda) = 0 \tag{D.7}$$

where $P(\lambda)$ is called *characteristic polynomial*.

A simple example

Consider the case

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad A - \lambda I = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda. \end{pmatrix} \tag{D.8}$$

We have to solve

$$det(A - \lambda I) = (2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = 0 \tag{D.9}$$

We have

$$\sqrt{\Delta} = \sqrt{b^2 - 4ac} = 2 \in \mathbb{R} \tag{D.10}$$

so we'll have two real solutions. These are given by

$$\lambda_{1,2} = \frac{-b \pm \sqrt{\Delta}}{2a} = 2 \pm 1 = \begin{cases} \lambda_1 = 3 \\ \lambda_2 = 1 \end{cases}. \tag{D.11}$$

What are the eigenvectors? We have to look for solutions v such that $Av = \lambda_{1,2}v$.

- For λ_1 we have

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = 3 \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \Leftrightarrow \begin{cases} 2x_1 + y_1 = 3x_1 \\ x_1 + 2y_1 = 3y_1 \end{cases} \Leftrightarrow x_1 - y_1 = 0 \Leftrightarrow$$

$$v_1 = c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad c_1 \in \mathbb{R}^* \tag{D.12}$$

- For λ_2 we have

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = 1 \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \Leftrightarrow \begin{cases} 2x_2 + y_2 = x_2 \\ x_2 + 2y_2 = y_2 \end{cases} \Leftrightarrow x_2 + y_2 = 0 \Leftrightarrow$$

$$v_2 = c_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad c_2 \in \mathbb{R}^* \tag{D.13}$$

As we can see, eigenvectors are arbitrary up to a non-zero constant, i.e. what matters is only their direction.
The two eigenvetors are orthogonal to each other

$$v_1 \cdot v_2 = v_1^\top v_2 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 0. \tag{D.14}$$

We can define two *unit vectors* v_1', v_2' associated to the eigenvalues we have found

$$v_1 \cdot v_1 = v_1^\top v_1 = \|v_1\| = 2 \Rightarrow v_1' = \frac{1}{\sqrt{2}} v_1$$

$$v_2 \cdot v_2 = v_2^\top v_2 = \|v_2\| = 2 \Rightarrow v_2' = \frac{1}{\sqrt{2}} v_2 \tag{D.15}$$

$$\tag{D.16}$$

Then we can write

$$v_i' \cdot v_j' = \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases} \tag{D.17}$$

so the two vectors v_1', v_2' are orthonormal. Out of these two vectors we can build a matrix $V = (v_1' \quad v_2')$ which is *orthogonal*, i.e.

$$V^\top V = V V^\top = I_2 \text{ (2-dimensional identity matrix).} \tag{D.18}$$

This means that V is invertible and the inverse matrix is given by $V^\top$. Defining

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{D.19}$$

we have, by definition, $AV = \Lambda V$ and, consequently, we have the *canonical expansion*

$$A = V\Lambda V^{-1} \text{ or } \Lambda = V^{-1}AV. \tag{D.20}$$

The original matrix A has been diagonalized.

General case

The previous example represented a very special case. In fact

- A was a 2×2 matrix so $P(\lambda)$ was a polynomial of order 2 in λ and the eigenvalues have been calculated easily. Things are not quite so simple for dimension $n > 2$ – finding the roots of the polynomial needs to happen numerically.
- A was real and symmetric (i.e. $A^\top = A$). It can be shown that in such a case one can always find two non-negative eigenvalues and their eigenvectors are orthogonal.

In general, one always starts with a characteristic polynomial defined as $P(\lambda) = det(A - \lambda I_n)$ where $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ with $\mathcal{M}_{n\times n}(\mathbb{R})$ the space of real $n \times n$ matrices. Then, in order to find the eigenvalues and eigenvectors, the following steps have to be performed.

1. The characteristic polynomial can always be factorized as follows

   $$P(\lambda) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2} \dots (\lambda - \lambda_k)^{n_k} \tag{D.21}$$

   where $\lambda_1, \lambda_2 \dots \lambda_k$ are the roots of the of characteristic polynomial and $n_1, n_2 \dots n_k$ are the corresponding multiplicities. The multiplicities are such that

   $$\sum_{i=1}^{k} n_i = n. \tag{D.22}$$

   This decomposition is always possible in $\mathbb{C}$, not always in $\mathbb{R}$.
   Once the eigenvalues have been identified by solving the equation $P(\lambda) = 0$ (numerically, if necessary) one must find the corresponding eigenvectors $V = (v_1, v_2, \dots, v_k)$. The diagonalization of matrix A yields a matrix with the following structure:

   $$\Lambda = \begin{pmatrix} \Lambda_1 & 0 & \dots & 0 \\ 0 & \Lambda_2 & \dots & 0 \\ \dots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \Lambda_k \end{pmatrix} \text{ where } \Lambda_i = \begin{pmatrix} \lambda_i & & 0 \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix}.$$

   (n_i times)

2. Solve for $Av_i = \lambda_i v_i$ for each $i \leq i \leq k$ (Gaussian elimination will yield v_i). The v_i's are not repeated, as v_i can be a matrix $n_i \times n_i$ if $n_i > 1$.

3. If desired, normalize all v_i's.

What are the advantages of going through this procedure? There are several of them:

- once the diagonal matrix has been found it is easy to raise the matrix to any power m: $A^m = V\Lambda^m V^{-1}$ where Λ^m can be readily obtained from Λ

$$\Lambda^m = \begin{pmatrix} \lambda_1^m & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n^m \end{pmatrix} \tag{D.23}$$

  and for $m = -1$ we get the inverse matrix

$$A^{-1} = V \begin{pmatrix} \frac{1}{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_n} \end{pmatrix} V^{-1} \tag{D.24}$$

  which is possible only in the case in which all of the λ_i are non-zero;

- a zero eigenvalue means that $\exists\, v \neq 0 \mid Av = 0$. v is in the null space of A and it is denoted as $\mathcal{N}(A)$;

- the number r of non-zero eigenvalues ($r \leq n$) yields the *rank* of A, it corresponds to the number of linearly independent matrix columns;

- we can identify $(n - r)$ vectors $(v_{r+1}, \dots, v_n)$ which define a basis for $\mathcal{N}(A)$. This basis is orthogonal if A is real and symmetric.

All of this can be done numerically, of course, though it may take a long time for very large matrices[1].

[1] `np.linalg.eig()`

III Singular value decomposition

Definition

Eigendecomposition is the privilege of square matrices. For non-square matrices, similar benefits may be obtained via the *singular value decomposition* (SVD for short). Any matrix $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ (or $A \in \mathcal{M}_{m\times n}(\mathbb{C})$) admits a singular value decomposition as

$$A = U\, \Sigma\, V^\top \tag{D.25}$$

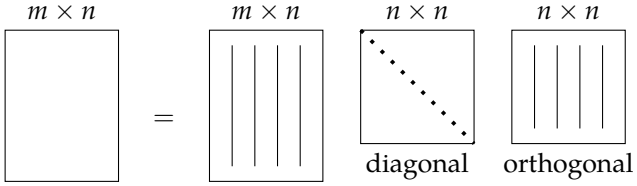
For $m \geq n$ we have

$$\Sigma = \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n \end{pmatrix} \text{ with } \sigma_i \text{the } \textit{singular values} \tag{D.26}$$

$$VV^\top = V^\top V = I_n \textit{ (right singular vectors)} \tag{D.27}$$

$$U^\top U = I_n \textit{ (left singular vectors)} \tag{D.28}$$

so

$A =$  .

If A is "fat" $(m < n)$, a similar form exists with

$$\Sigma = \begin{pmatrix} \Sigma_{n\times n} & 0 \\ 0 & \underbrace{0}_{(m-n) \text{ columns of zeroes}} \end{pmatrix}$$

Properties

What does the singular value decomposition do for us? The number of non-zero singular values defines the *rank* of matrix A. In practice, how do we find U, Σ and V? The matrix $(A^\top A)$ is real and symmetric, so it is diagonalizable and admits orthonormal eigenvectors, i.e.

$$A^\top A = V' \Lambda V'^T, \quad V'^T V = I \tag{D.29}$$

so, if $A = U\Sigma V^\top$ then

$$A^\top A = (V\Sigma^\top U)(U\Sigma V^\top) = V\Sigma \underbrace{U^\top U}_{I_n} \Sigma V^\top = V\Sigma^2 V^\top \equiv V' \Lambda V'^\top \tag{D.30}$$

so

- $V' \equiv V$: the eigenvectors of $A^\top A$ are the right singular vectors of A,
- $\Lambda \equiv \Sigma^2$: the eigenvalues of $A^\top A$ are the squared singular values of A.

What about U?

$$A = U\Sigma V^\top \Rightarrow AV = U\Sigma \underbrace{V^\top V}_{I_n} = U\Sigma \tag{D.31}$$

so $Av_i = u_i\sigma_i$ for $i \in [1, n]$. For each σ_i we have two possibilities

$$\begin{cases} \sigma_i \neq 0 \Rightarrow u_i = \frac{1}{\sigma_i} A v_i & \text{(i)} \\ \sigma_i = 0 \Rightarrow u_i \text{ is arbitrary (provided that } U^\top U = I_n) & \text{(ii)} \end{cases} . \tag{D.32}$$

Because of (i) and the orthogonality of the columns of U, the set of vectors $\{u_1, \ldots, u_r\}$ represents a basis for $R(A)$ which is the *range* of $A = \{y \in \mathbb{R}^n \mid \exists x \in \mathbb{R}^m \mid y = Ax\}$[2]. Similarly, $\{v_{r+1}, \ldots, v_n\}$ are a basis for $\mathcal{N}(A)$. One can show that

- $\{v_1, \ldots, v_r\}$ is a basis for $R(A^\top)$
- $\{u_{r+1}, \ldots, u_n\}$ is a basis for $\mathcal{N}(A^\top)$.

So the singular value decomposition gives us *everything* we want to know about A

- its diagonal form
- its rank
- a basis for $\mathcal{N}(A)$, $R(A)$, $\mathcal{N}(A^\top)$, $R(A^\top)$ which are the four fundamental subspaces. All this with an incredibly efficient algorithm[3].

[2] In English, the range of A the set of vectors in the arrival space that can be written as A times a vector from the source space. In other words, it is the image of the source space by the transformation A.

[3] `np.linalg.svd()`

Low Rank Approximation

The idea is to approximate a matrix by one with low-rank, with fit measured by the Frobenius norm[4], i.e. find the matrix $\widehat{D}$ such that

$$\|D - \widehat{D}\|_{\mathrm{F}} \text{ is minized, subject to } \quad \operatorname{rank}\left(\widehat{D}\right) \leq r \tag{D.33}$$

The result is referred to as the matrix approximation lemma or Eckart-Young-Mirsky theorem, for those who want to show off at cocktail parties. This minimization problem may be solved via (you guessed it) SVD. Let $D = U\Sigma V^\top \in \mathbb{R}^{m\times n}$, $m \leq n$ be the singular value decomposition of D and partition U, $\Sigma =: \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$, and V as follows:

$$U =: \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \quad \Sigma =: \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \quad \text{and} \quad V =: \begin{bmatrix} V_1 & V_2 \end{bmatrix}, \tag{D.34}$$

where Σ_1 is $r \times r$, U_1 is $m \times r$, and V_1 is $n \times r$. Then the rank-r matrix, obtained from the **truncated singular value decomposition**

$$\widehat{D}^* = U_1 \Sigma_1 V_1^\top, \tag{D.35}$$

[4] The Frobenius norm of a matrix A is such that: $\|A\|_F{}^2 = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 = \operatorname{trace}(A^* A) = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2$. It is very useful in numerical linear algebra, being invariant under unitary transformations like rotations

is such that:

$$\|D - \widehat{D}^*\|_{\mathrm{F}} = \min_{\mathrm{rank}(\widehat{D}) \leq r} \|D - \widehat{D}\|_{\mathrm{F}} = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_m^2}. \qquad \text{(D.36)}$$

The minimizer $\widehat{D}^*$ is unique if and only if $\sigma_{r+1} \neq \sigma_r$. In other words, by retaining only r singular values, one gets a matrix of rank r, with a misfit to the original matrix that is readily quantified by the quadratic sum of the remaining singular values. This is better than whiskey without hangovers! It means that out of every ill-conditioned system, one can find a low-rank solution that is better conditioned, therefore invertible. Of course, this comes at the price of a misfit of the original matrix, but at least we can directly optimize that misfit and use the dual information to choose the truncation level r that best accomplishes our goals.

BIBLIOGRAPHY

Abramowitz, M., and I. A. Stegun (1965), *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Math # 55, Dover Publications.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anchukaitis, K., and J. Tierney (2013), Identifying coherent spatiotemporal modes in time-uncertain proxy paleoclimate records, *Climate Dynamics*, *41*(5-6), 1291–1306, DOI: 10.1007/s00382-012-1483-0.

Ault, T. R., J. E. Cole, M. N. Evans, H. Barnett, N. J. Abram, A. W. Tudhope, and B. K. Linsley (2009), Intensified decadal variability in tropical climate during the late 19th century, *Geophys. Res. Lett.*, *36*(8).

Ault, T. R., C. Deser, M. Newman, and J. Emile-Geay (2013), Characterizing decadal to centennial variability in the equatorial pacific during the last millennium, *Geophysical Research Letters*, *40*, 3450–3456, DOI: 10.1002/grl.50647.

Backus, G., and F. Gilbert (1968), The Resolving Power of Gross Earth Data, *Geophysical Journal International*, *16*, 169–205, DOI: 10.1111/j.1365-246X.1968.tb00216.x.

Bretherton, C. S., C. Smith, and J. M. Wallace (1992), An Intercomparison of Methods for Finding Coupled Patterns in Climate Data, *J. Climate*, *5*, 541–560.

Chave, A. D., D. J. Thomson, and M. E. Ander (1987), On the robust estimation of power spectra, coherences, and transfer functions, *J. Geophys. Res.*, *92*, 633–648, DOI: 10.1029/JB092iB01p00633.

Comboul, M., J. Emile-Geay, M. N. Evans, N. Mirnateghi, K. M. Cobb, and D. M. Thompson (2014), A probabilistic model of chronological errors in layer-counted climate proxies: applications to annually banded coral archives, *Climate of the Past*, *10*(2), 825–841, DOI: 10.5194/cp-10-825-2014.

Cook, E. R., and K. Peters (1981), The smoothing spline: A new approach to standardizing forest interior tree-ring width series for dendroclimatic studies, *Tree-Ring Bulletin*, *41*, 45–53.

Dommenget, D. (2007), Evaluating eof modes against a stochastic null hypothesis, *Climate Dynamics*, *28*(5), 517–531, DOI: 10.1007/s00382-006-0195-8.

Emile-Geay, J. (2006), ENSO dynamics and the Earth's climate : from decades to Ice Ages, Ph.D. thesis, Columbia University.

Emile-Geay, J., and J. A. Eshleman (2013), Toward a semantic web of paleoclimatology, *Geochemistry, Geophysics, Geosystems*, *14*(2), 457–469, DOI: 10.1002/ggge.20067.

Foster, G. (1996), Wavelets for period analysis of unevenly sampled time series, *Astron. Jour.*, *112*, 1709, DOI: 10.1086/118137.

Gelman, A., and H. Stern (2006), The difference between "significant" and "not significant" is not itself statistically significant, *The American Statistician*, *60*(4), 328–331, DOI: 10.1198/000313006X152649.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2013), *Bayesian Data Analysis*, 2nd ed., 675 pp., Chapman and Hall, New York, NY.

Ghil, M., R. M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou (2002), Advanced spectral methods for climatic time series, *Rev. Geophys.*, *40*(1), 1003–1052, DOI: 10.1029/2000RG000092.

Golub, G. H., and C. F. van Loan (1993), *Matrix Computations*, 642 pp pp., 2d ed. Johns Hopkins University Press.

Golub, G. H., M. Heath, and G. Wahba (1979), Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, *21*(2), 215–223.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson (2007), Empirical orthogonal functions and related techniques in atmospheric science: A review, *International Journal of Climatology*, *27*(9), 1119–1152, DOI: 10.1002/joc.1499.

Hasselman, K. (1976), Stochastic climate models. part i. theory, *Tellus*, *28*, 473–485.

Hastie, T., R. Tibshirani, and J. Friedman (2008), *The elements of statistical learning: data mining, inference and prediction*, 2 ed., Springer.

Hu, J., J. Emile-Geay, and J. Partin (2017), Correlation-based interpretations of paleoclimate data – where statistics meet past climates, *Earth and Planetary Science Letters*, *459*, 362–371, DOI: 10.1016/j.epsl.2016.11.048.

Hurst, H. E. (1951), Long term storage capacities of reservoirs, *Trans. ASCE*, *116*, 776–808.

Huybers, P., and W. Curry (2006), Links between annual, milankovitch and continuum temperature variability, *Nature*, *441*(7091), 329–332.

Jaynes, E. T. (2004), *Probability Theory: The Logic of Science*, 727 pages, Cambridge University Press, Cambridge.

Johnstone, I. M., and A. Y. Lu (2009), Sparse Principal Components Analysis, *ArXiv e-prints*.

Kalnay, E. & coauthors. (1996), The NCEP/NCAR 40-year Reanalysis Project, *Bull. Amer. Meteor. Soc.*, *77*, 437–471.

Kim, K., and G. Shevlyakov (2008), Why gaussianity?, *Signal Processing Magazine, IEEE*, *25*(2), 102–113, DOI: 10.1109/MSP.2007.913700.

Kug, J.-S., F.-F. Jin, and S.-I. An (2009), Two types of El Niño events: Cold tongue El Niño and warm pool El Niño, *Journal of Climate*, *22*(6), 1499–1515, DOI: 10.1175/2008JCLI2624.1.

Lee, T., and M. J. McPhaden (2010), Increasing intensity of el niño in the central-equatorial pacific, *Geophysical Research Letters*, *37*(14), n/a–n/a, DOI: 10.1029/2010GL044007.

Lorenz, E. N. (1956), Empirical orthogonal functions and statistical weather prediction, *Scientific Report 1, Statistical Forecasting Project. 110268*, Massachusetts Institute of Technology Defense Doc. Center.

Lovejoy, S. (2015), A voyage through scales, a missing quadrillion and why the climate is not what you expect, *Climate Dynamics*, *44*(11-12), 3187–3210, DOI: 10.1007/s00382-014-2324-0.

Lovejoy, S., D. Schertzer, M. Lilley, K. B. Strawbridge, and A. Radkevich (2008), Scaling turbulent atmospheric stratification. i: Turbulence and waves, *Quarterly Journal of the Royal Meteorological Society*, *134*(631), 277–300, DOI: 10.1002/qj.201.

Mann, M. (2011), On long range dependence in global surface temperature series, *Climatic Change*, *107*, 267–276, 10.1007/s10584-010-9998-z.

Mann, M., and J. Lees (1996), Robust estimation of background noise and signal detection in climatic time series, *Clim. Change*, *33*, 409–445.

Meinshausen, M., N. Meinshausen, W. Hare, S. C. B. Raper, K. Frieler, R. Knutti, D. J. Frame, and M. R. Allen (2009), Greenhouse-gas emission targets for limiting global warming to 2c, *Nature*, *458*(7242), 1158–1162.

Mudelsee, M., D. Scholz, R. Röthlisberger, D. Fleitmann, A. Mangini, and E. W. Wolff (2009), Climate spectrum estimation in the presence of timescale errors, *Nonlinear Processes in Geophysics*, *16*(1), 43–56, DOI: 10.5194/npg-16-43-2009.

Nadakuditi, R., and A. Edelman (2008), Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples, *Signal Processing, IEEE Transactions on*, *56*(7), 2625–2638, DOI: 10.1109/TSP.2008.917356.

Neumaier, A., and T. Schneider (2001), Estimation of parameters and eigenmodes of multivariate autoregressive models, *ACM Trans. Math. Softw.*, *27*, 27–57.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng (1982), Sampling errors in the estimation of empirical orthogonal functions, *Mon. Weather Rev.*, *110*, 699–706.

O'Hagan, T. (2006), Bayes factors, *Significance*, *3*(4), 184–186, DOI: 10.1111/j.1740-9713.2006.00204.x.

Overland, J. E., and R. W. Preisendorfer (1982), A significance test for principal components applied to a cyclone climatology, *Monthly Weather Review*, *110*(1), 1–4, DOI: 10.1175/1520-0493(1982)110<0001:ASTFPC>2.0.CO;2.

Partin, J., T. Quinn, C.-C. Shen, J. Emile-Geay, F. Taylor, C. Maupin, K. Lin, C. Jackson, J. Banner, D. Sinclair, and C.-A. Huh (2013), Multidecadal rainfall variability in south pacific convergence zone as revealed by stalagmite geochemistry, *Geology*, DOI: 10.1130/G34718.1.

Rasmussen, E., and T. Carpenter (1982), Variations in tropical sea-surface temperature and surface winds associated with the Southern Oscillation/ El Niño, *Mon. Weather Rev.*, *110*, 354–384.

Reynolds, R. W., and T. M. Smith (1994), Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation., *J. Climate*, *7*, 929–948.

Sarachik, E. S., and M. A. Cane (2010), *The El Niño-Southern Oscillation Phenomenon*, 384 pp., Cambridge University Press, Cambridge, UK.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, *6*, 461–464.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*, 1 ed., Wiley.

Silver, P. G., and T. H. Jordan (1981), Fundamental spheroidal mode observations of aspherical heterogeneity, *Geophysical Journal International*, *64*(3), 605–634.

Silverman, B. W. (1986), *Density estimation: for statistics and data analysis*, Chapman & Hall, London.

Stoica, P., and Y. Selen (2004), Model-order selection: a review of information criterion rules, *Signal Processing Magazine, IEEE*, *21*(4), 36–47, DOI: 10.1109/MSP.2004.1311138.

Takahashi, K., A. Montecinos, K. Goubanova, and B. Dewitte (2011), Enso regimes: Reinterpreting the canonical and modoki el niño, *Geophysical Research Letters*, *38*(10), n/a–n/a, DOI: 10.1029/2011GL047364.

Tarantola, A. (2004), *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Thomson, D. J. (1982), Spectrum estimation and harmonic analysis, *Proc. IEEE*, *70(9)*, 1055–1096.

Tukey, J. W. (1977), *Exploratory data analysis*, Addison-Wesley.

Wahba, G. (1990), Spline models for observational data, in *CBMS-NSF Regional Conference Series in Applied Mathematics, Based on a series of 10 lectures at Ohio State University at Columbus, March 23-27, 1987, Philadelphia: Society for Industrial and Applied Mathematics, 1990*.

Wahba, G., and Y. Wang (1995), Behavior near zero of the distribution of GCV smoothing parameter estimates, *Stat. Probabil. Lett.*, *25*, 105–111.

Wallace, J. M., and D. S. Gutzler (1981), Teleconnections in the geopotential height field during the Northern Hemisphere winter, *Mon. Weather Rev.*, *109*, 784–812.

Wax, M., and T. Kailath (1985), Detection of signals by information theoretic criteria, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *33*(2), 387–392, DOI: 10.1109/TASSP.1985.1164557.

Weinert, H. L. (2009), A fast compact algorithm for cubic spline smoothing, *Computational Statistics & Data Analysis*, *53*(4), 932–940, DOI: 10.1016/j.csda.2008.10.036.

Wikle, C. K., and L. M. Berliner (2007), A bayesian tutorial for data assimilation, *Physica D: Nonlinear Phenomena*, *230*(1–2), 1 – 16, DOI: http://dx.doi.org/10.1016/j.physd.2006.09.017, data Assimilation.

Wilks, D. S. (2011), *Statistical Methods in the Atmospheric Sciences: an Introduction*, 676 pp., Academic Press, San Diego.

Wunsch, C. (2000), On sharp spectral lines in the climate record and the millennial peak, *Paleoceanography*, *15*(4), 417–424, DOI: 10.1029/1999PA000468.

Zou, H., T. Hastie, and R. Tibshirani (2006), Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, *15*(2), 265–286, DOI: 10.1198/106186006X113430.

INDEX