Notes: Practical Statistics for Physics & Astronomy

R. Benton Metcalf

Alma Mater Studiorum - Universitá di Bologna

May 19, 2022

Contents

Chapter 1

What is Probability?

Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means. — Bertrand Russell

Although this question could be considered purely philosophical and not practical, it turns out that differences in the basic conception of what probability is have lead to very concrete differences in how it is used to analyze data. For this reason let us briefly consider it.

Of course the concept of certain outcomes being more probable than others has been with us for as long as people have been around. In fact you can make the argument that even some animals have an understanding of probability in that they will anticipate an event without fully committing to it happening. For example my dog will prepare herself to chase a ball when I pick one up, but wont go running off until I actually have thrown it. Probability in the sense of expressing preferences based on inconclusive information can be consider the basis of intelligence. This is why probability and statistics are inextricably linked to machine learning and artificial intelligence. As scientists we seek to use probability to link uncertain measurements and observations to insights about the natural world and/or predictions of future measurements or observations.

The intellectual difficulties and disagreements arise when one tries to make probability quantitative and systematic. There have several approaches to putting probability theory on a solid foundation. We will consider the two most commonly cited.

1.1 Frequentist interpretation of probability

Imagine there is some event, instance or outcome of an experiment or observation called A. The probability of A is the fraction of times A occurs when the experiment or observation is repeated in the same way or circumstances an *infinite* number of

times. We can write this symbolically as

$$P(A) = \lim_{N\to\infty} \frac{\text{number of trials where A is true}}{N(\text{total number of trials})} \tag{1.1.1}$$

This is the traditional definition of probability used by Fermat and Pascal in there famous correspondence on gambling problems in 1654, by Jacob Bernoulli in his *Ars Conjectandi (The art of Conjecturing)* 1713 and by Laplace in his *Théorie analytique des probabilités (Analytic Theory of Probability)* 1812. It was almost universally used for centuries despite no one ever having done anything *exactly* the same way twice let along an *infinite* number of times.

Applying this definition to any physical phenomenon requires a partitioning of the world into things that are known and fixed on each repetition of the observation and those things that are not known and change every repetition. If nature is deterministic and an experiment could be set up *exactly* the same way in all respects than the outcome would always be the same and probability would not apply. Of course even in classical physics it is not possible to know the state of every atom and photon that might possibly influence your measurement apparatus (or brain). It is these things that change when repeating the observation.

This partitioning between known and unknown factors seems reasonable when we talk about the positions and momenta of particles in a gas or the flipping of a coin, but in many other common situations where probability is used it seems less well defined. Say someone tells you that there is a 30% probability that candidate A will win an election tomorrow. Of course an identical election will never be run again and was never run in the past. There are many factors, known and unknown, that could affect an election. This statement was probably based on polling data. By the above definition of probability, this means that if the election were held an infinite number of times in which the polling data were exactly the same the candidate would win 30% of them. This seems like a completely unverifiable claim. If scientific knowledge must be reproducible to be considered true then it would seem that any such argument should be considered unscientific. And yet probability through statistics is at the foundation of all quantitative measurements.

Let us be a bit more practical. Let's say we don't need an infinite number of trials, but just a very *large number* of them. Let's say we flip a coin a *very large number* of times. If we did it say one billion times we would not expect that *exactly* 500 million times it would be heads. We would expect that roughly half, but not exactly half of the times it would be heads even if the probability of getting heads in each flip is 1/2. We might try to quantify how close the number of heads should be to 500 million, but in doing so we would need to use a probabilistic argument that would use the very concept we are trying to define.

Many statisticians and philosophers have found this definition of probability problematic. Despite this it is the definition usually used by scientists when they are forced to addressing this subject.

1.2 Subjective or Bayesian interpretation of probability

Thomas Bayes (1701 - 1761) (and initially Jacob Bernoulli 1655-1705) had a different conception of what probability is although the idea was not put on a firm theoretical foundation until the 1940's and 50's by G. Polya, R.T. Cox and E.T. Jaynes. It did not make its way into common use in science, in the form of Bayesian statistics, until relatively recently (80s and 90s for astrophysics).

In this school of thought, probability theory is an extension of formal logic to situations where the truth or falsehood of a proposition (e.g. "It will rain tomorrow." or "The mass of the Earth is between 5.972×10^{24} kg and 5.978×10^{24} kg.") cannot be deduced conclusively by deductive reasoning. A proposition or statement about the world has a probability assigned to it that depends on the evidence for and against its truth. When deductive reasoning can be applied conclusively this function is either zero (false) or one (true). In this way symbolic logic is a limiting case of probability theory. Surprisingly, from just the following requirements (or *desiderata*) on the probability function of a proposition you can deduce the rules of probability (section 1.5) and show that they are complete without ever mentioning randomness or repetition of experiments.

Desiderata:

1. Degrees of plausibility are represented by real numbers.

2. The measure of plausibility must exhibit qualitative agreement with rationality. This means that as new information supporting the truth of a proposition is supplied, the number which represents the plausibility will increase continuously and monotonically. Also, to maintain rationality, the deductive limit (plausibility 1 and 0) must be obtained where appropriate.

3. Consistency

   (a) *Structured consistency* : If the conclusion can be reasoned out in more than one way, every possible way must lead to the same result. (Logically equivalent statements must have the same weight.)

   (b) *Propriety*: The theory must take account of all information that is relevant to the question.

(c) *Jaynes consistency*: Equivalent states of knowledge must be represented by equivalent plausibility assignments. For example, if $(A, B)||C = B||C$, then the plausibility of $(A, B)||C$ must equal the plausibility of $B||C$. (Here $||$ is logical "or" and , is logical "and").

(taken from Gregory (2006)).

These foundational proofs are very interesting, but outside the scope of this course (for those that are interested see chapter 2 of Gregory (2006) or, more comprehensively, Jaynes (2003)). One thing that is of importance here is that this definition allows one to define the probability of something that would not usually be considered a *random variable* or a repeated even. It also establishes the accumulation of supporting evidence as central to the meaning of probability. Probability is a measure of knowledge, or ignorance, of an event and not a property of the event itself. These principles are central to the Bayesian method of parameter estimation and model selection that we will study later.

1.3 classical interpretation of probability

The "classical interpretation" of probability is more of a prescription for calculating probabilities than a definition of probability. It relies on identifying events that are equally likely or probable and then grouping the events to find the probabilities of more complicated events. The **principle of indifference** holds that each of n mutually exclusive events that encompass all possibilities (collectively exhaustive) should be given probability $1/n$ if there is no reason to favor one over any other. This is often the argumentation used in classical statistical mechanics where each micro-state of the system is taken to be equally probable. If one then says that the probability of being in either of two mutually exclusive states is the sum of their probabilities and that the some of the probabilities of being in all possible states is one then you can find a numerical value for the probability of each state. A macro-state (one with temperature equal to some value or total energy equal to some value) corresponds to many micro-states so by adding up their probabilities you can find the probabilities of macro states which will not necessarily be equal.

This approach is limited to a restricted class of problems where these fundamental, equally probably states can be identified and, by itself, leaves open some questions of interpretation. What does it mean that two states are equally probable? What does the probability of a macro-state mean? Another problem is that it is not obvious that all events that we commonly apply probability to can be reduced in this way to a collection of equally probable, mutually exclusive events. How do you apply this to the election? Or an unfair coin? This also presupposes the rules for combining

probabilities, which we will get to in a moment, without any justification for them.

1.4 Quantum mechanical probability

Probability according to the Copenhagen interpretation of quantum mechanics is a fundamentally different thing than the probability that was in use before. In the frequentist interpretation of probability it is assumed that there are some "hidden variables" that are different every trial. It is often said that Bell's inequalities prove that hidden variable cannot exist. This is not actually true. Bell's inequalities and their experimental verification show that any complete quantum theory must be non local. Viable deterministic, hidden variable extensions to quantum theory exist, most notably the theories of David Bohm. In these theories probability would have its classical meaning - hidden variables have definite values we just don't know them, and in some cases can't know them. The price is a seeming violation of special relativity.

In the more traditional "Copenhagen interpretation" of QM probability plays a strange role. When a measurement is made the square of the wave function gives the probability of an observation, but up to that point the outcome is not determined, not just difficult to determine. This makes probability a property of physical systems and not solely a property of the observer's ignorance. This seems to imply an intimate connection between physical laws and human thought! There are extensions to QM with "spontaneous collapse" where a definite physical prescription is given of how and when a wave function collapses. Probability still seems to be a physical thing in these theories in contradiction to the traditional interpretations of it. Still further out there are Everett's many-worlds interpretation and the many-minds interpretation.

This is obviously a subject for a different course (or a *Star Trek* episode) so I will go no further here, but those that are interesting might consult Norisen (2017) for an interesting treatment of the subject.

1.5 the rules of probability

Suppose A, B, ... are events that either occur or don't occur, that is they have values true or false (or 0 and 1 if you prefer). $P(A)$ is the probability of A occurring or being true. We can combine events in one of two ways. (A, B) means "A *and* B". It is true if both of them are true and false otherwise. $(A||B)$ means "A *or* B" it is true if either A or B is true. It is true if both are true. $\overline{A}$ means "not A". Note that $\overline{A||B} = \overline{A}, \overline{B}$ and $\overline{A, B} = \overline{A}||\overline{B}$ in the sense that there are no combinations of trues and falses for A and B that give different answers on either side of the equality. See

A	B	A,B	$\overline{A,B}$	$\overline{A}\|\|\overline{B}$	$A\|\|B$	$\overline{A\|\|B}$	$\overline{A},\overline{B}$	$\overline{A},B$	$A,\overline{B}$	$\overline{A}\|\|B$	$\overline{A}\|\|\overline{B}$
F	T	F	T	T	T	F	F	T	F	T	F
F	F	F	T	T	F	T	T	F	F	T	T
T	T	T	F	F	T	F	F	F	F	T	T
T	F	F	T	T	T	F	F	F	T	F	T

Table 1.1: The truth table for binary logical expressions. Statements with the same truth table are logically equivelent. Note that $\overline{A,B} = \overline{A}||\overline{B}$ and $\overline{A||B} = \overline{A},\overline{B}$ because their truth tables are the same.

table 1.1. In the language of Boolean algebra, they have the same truth table and are therefor equivalent statements. There probabilities must also be the same.

$P(A,B)$ is called the **joint probability** of events A and B. $P(A||B)$ is often called the **disjoint probability** of events A and B.

$P(A|B)$ is called a **conditional probability**. It means the probability of A *given* that B is true. You can imagine every probability being a conditional probability where it is "conditioned" on everything that you assume about the state of the Universe. Some of these things are assumed to be irrelevant and are left out. Some might be relevant but are taken for granted so they are left out. The probability that a coin comes up heads does not depend on the time of day. It does depend on the assumption that it is a fair coin - no more likely to be heads than tails - although it might not always be stated. This is a simple example of a **statistical model** for the experiment. The two fundamental rules of probability theory are

$$\begin{array}{ll} P(A) \geq 0 & \textbf{positive semidefinite} \\ P(A,B) = P(A)P(B|A) & \textbf{product rule} \\ P(A) + P(\overline{A}) = 1 & \textbf{sum rule} \end{array} \tag{1.5.1}$$

These rules can be derived from the basic requirements or "desidariata" stated before in section 1.2 , but they can also be take as axioms. From these two rules and logic rules we can derive all the necessary properties of probability.

There are several particularly useful results that follow from these rules. From the logical requirement that (A,B) is the same as (B,A) and the product rule we get

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \textbf{Bayes' theorem} \tag{1.5.2}$$

Applying the sum rule to $(A||B)$ gives

$$\begin{aligned}
P(A||B) &= 1 - P(\overline{A||B}) && && (1.5.3)\\
&= 1 - P(\overline{A}, \overline{B}) && \text{see table 1.1} && (1.5.4)\\
&= 1 - P(\overline{A})P(\overline{B}|\overline{A}) && \text{product rule} && (1.5.5)\\
&= 1 - P(\overline{A})\left[1 - P(B|\overline{A})\right] && \text{sum rule} && (1.5.6)\\
&= 1 - P(\overline{A}) - P(\overline{A})P(B|\overline{A}) && && (1.5.7)\\
&= P(A) + P(\overline{A})P(B|\overline{A}) && \text{sum rule} && (1.5.8)\\
&= P(A) + P(\overline{A}, B) && \text{product rule} && (1.5.9)\\
&= P(A) + P(B)P(\overline{A}|B) && \text{product rule} && (1.5.10)\\
&= P(A) + P(B)\left[1 - P(A|B)\right] && \text{sum rule} && (1.5.11)\\
&= P(A) + P(B) - P(B)P(A|B) && && (1.5.12)\\
P(A||B) &= P(A) + P(B) - P(B, A) && \textbf{extended sum rule} && (1.5.13)
\end{aligned}$$

In words, the disjoint probability of two events is equal to the sum of their probabilities minus their joint probability.

If A and B are **independent** then the probability of A occurring does not depend on whether B has occurred so $P(A|B) = P(A)$ through the product rule this implies $P(B|A) = P(B)$ and

$$P(A, B) = P(A)P(B) \quad \text{independent events} \tag{1.5.14}$$

If two events are **mutually exclusive**, that is they cannot occur at the same time (the first flip of a coin cannot be both heads and tails) then $P(A, B) = 0$ and the extended sum rule becomes

$$P(A||B) = P(A) + P(B) \quad \text{mutually exclusive events} \tag{1.5.15}$$

Example: If you roll a die once the probability of getting a 6 *or* a 5 is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. If you roll a die twice the probability of getting a 6 *and then* a 5 is $\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$. The probability of getting a 6 *and* a 5 is twice this, $\frac{1}{18}$, because there are two ways of doing this, a 6 first or a 5 first.

This second case can be calculated in an alternative way. In the first roll we must get a 5 or a 6. We have calculated that the probability of this is $\frac{1}{3}$ (sum rule). Once this is done in the second roll we most get whichever number we didn't get in the first roll, one number out of 6, probability $\frac{1}{6}$. The probability of these two independent events happening is then given by the product rule $\left(\frac{1}{3}\right)\left(\frac{1}{6}\right) = \frac{1}{18}$.

Now say we have a set of observations $\{A_I\}$ that are all mutually exclusive and together they include all possible outcome then

$$\begin{aligned} 1 &= P(A_1||A_2||A_3||\ldots|B) + P(\overline{A_1||A_2||A_3||\ldots}|B) && (1.5.16)\\ &= P(A_1||A_2||A_3||\ldots|B) + 0 && (1.5.17)\\ &= P(A_1|B) + P(A_2||A_3||\ldots|B) && (1.5.18)\\ &= P(A_1|B) + P(A_2|B) + P(A_3||\ldots|B) && (1.5.19)\\ &= \sum_i P(A_i|B) && (1.5.20) \end{aligned}$$

This is the origin of the normalization requirement on any probability distribution function (PDF). Note that I have put a B in as a condition on all the probabilities, but this would hold without them.

Another important result along these lines is

$$\sum_i P(B|A_i)P(A_i) = \sum_i P(B, A_i) = \sum_i P(A_i|B)P(B) = P(B)\sum_i P(A_i|B) = P(B) \tag{1.5.21}$$

with the same requirements on the set $\{A_i\}$. This is the origin of what we will later call **marginalization**.

Problem 1. *We know the probability of a person having red hair is $P(R)$, the probability of a person having blue eyes is $P(B)$ and that the probability of a red headed person having blue eyes is $P(B|R)$.*

1. *What is the probability that a blue eyed person will have red hair?*
2. *What is the probability that a person will have both blue eyes and red hair?*
3. *What is the probability that a person will have either blue eyes or red hair?*

Problem 2. *Say we have developed a new cheaper test for cancer. We test it on patients that we know have cancer and find that 90% of them get a positive result. Then we test it on patients that we know don't have cancer and we find that 90% of them get a negative test result. The test shows only negative or positive, not undetermined. From previous research we know that the cancer rate in the general population is 1 in 10,000.*

1. *If we use this test in the general population what is the chance of a person with a positive test of actually having cancer?*
2. *What is the chance of a random person having cancer and a false test result?*

Chapter 2

Some warm up problems

There are a large class of problems, classical statistical physics included, for which individual states are considered equally probable and the question is how many states out of all possible states have a certain property. The property could be the temperature, pressure or having a full house in your poker hand and states could be the positions of each atom in a gas, the spin state of each atom in a metal or the identity of the five cards you are dealt in poker. Here are some very simple problems that illustrate some of the counting techniques used throughout statistics and in the process introduce the two most important distributions in statistics: the binomial and multinomial distributions[1].

2.1 Flipping coins & the Binomial Distribution

Let us flip a fair coin 4 times. What is the probability of all 4 being heads? Let's call the outcome of each flip being H or T. Since each flip is *independent* we can use the product rule

$$P(H,H,H,H) = P(H)P(H)P(H)P(H) = P(H)^4 \tag{2.1.1}$$

Each flip will have a probability of 0.5 of being heads so the probability is $(0.5)^4 = 0.0625$. Simple enough.

Now let us look at the probability of getting all heads but one

$$P(\text{one tail}) = P((\#1 \text{ is T})||(\#2 \text{ is T})||(\#3 \text{ is T})||(\#4 \text{ is T})) \tag{2.1.2}$$

$$= P(T,H,H,H) + P(H,T,H,H) + P(H,H,T,H) + P(H,H,H,T) \tag{2.1.3}$$

[1]The binomial is actually a special case of the multinomial so these could be considered one distribution.

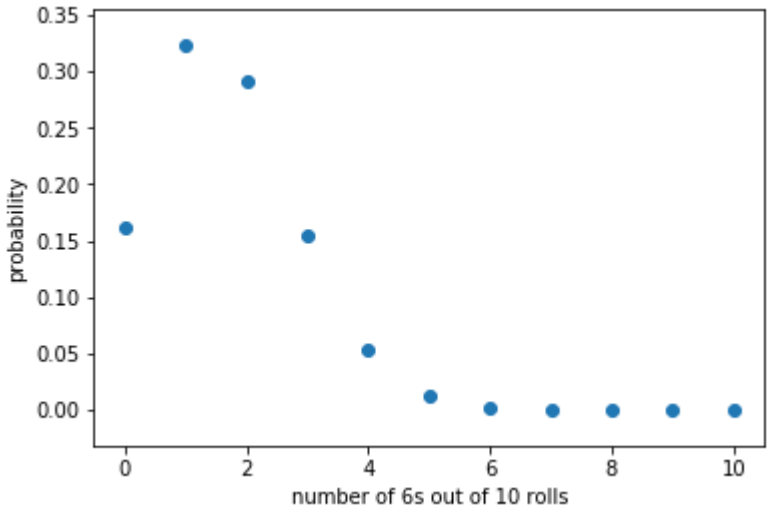

Figure 2.1: The binomial distribution for the number of 6s in ten rolls of a die or one roll of ten dice. $N = 10$, $k = 0 \ldots 10$, $p = 1/6$

where the sum rule has been used. Since each of these probabilities must be equal

$$P(\text{one tail}) = 4\,P(T, H, H, H) = 4\,P(T)P(H)^3 \tag{2.1.4}$$

where the product rule has been used.

You can see how this can be extended to any number of tails and heads even if the coin is not fair. Note that because there are only two possible outcomes $P(T) = 1 - P(H)$ - let's just call $P(T)$ p - the probability of getting n tails out of N will be $p^n(1-p)^{N-n}$ times the number of ways you can draw tails n times out of N flips. Since we are not concerned with the order of the flips, the number of cases is given by the **binomial coefficient**

$$\binom{N}{n} \equiv \frac{N(N-1)(N-2)\cdots(N-n-1)}{n!} = \frac{N!}{n!(N-n)!} \tag{2.1.5}$$

This is often expressed as "N choose n".

So the probability of getting n outcomes of probability p out of N trials is

$$p(n|N) = \binom{N}{n} p^n (1-p)^{N-n} \qquad n \leq N \tag{2.1.6}$$

which is called the **binomial distribution**. The case of $N = 10$ and $p = 1/6$ is shown in figure 2.1.

We can also think of the binomial distribution as the solution to the problem of "drawing with replacement". Imagine a bag full of green and blue balls. Each trial you take one out, record its color and put it back in the bag. The probability of getting a green ball in each draw is p = (number of green balls) / (total number of balls). The chance of getting n green balls in N draws is given by the binomial distribution.

The **Bernoulli distribution** is the special case of $N = 1$

$$p(n) = \begin{cases} p & , \quad n = 1 \\ (1-p) & , \quad n = 0 \end{cases} \tag{2.1.7}$$

an almost trivial case, but perhaps the first probability distribution written down. These types of experiments where each trial is independent of the others and there are two possible outcomes (or perhaps the outcomes are divided into two categories) are called **Bernoulli trials**. The binomial distribution is important for calculating the distribution of any finite sample of observations and comes up very often as we will see.

Note that the binomial coefficient gets its name because of the **binomial expansion**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k} \tag{2.1.8}$$

The mean and variance of the binomial distribution are

$$\langle n \rangle = Np \tag{2.1.9}$$

$$\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = Np(1-p) \tag{2.1.10}$$

2.2 Rolling Dice & the Multinomial Distribution

Let us look at a more complicated case. A roll of a die has 6 possible outcomes. If we rolled the die N times and we asked what the probability of getting n 6s we could use the binomial distribution because there are only two outcomes to each trail that count– 6 and not 6.

But if we asked what the probability of getting one 1, two 2s and two 5s in five rolls is we will have to work a little harder. By the product rule for independent trials,

$$P(1,2,2,5,5) = p_1 p_2 p_2 p_5 p_5 = (p_1)(p_2)^2 (p_5)^2 \tag{2.2.1}$$

The order of the rolls is not important so all permutations of rolls must have equal probability and must be added together. There are $n!$ permutations, or orderings, of

n things so there are 5! ways of getting one 1, two 2s and two 5s and they all have the same probability. But some of the outcomes are the same so we cannot count them as distinct combinations - you could reorder the p_2's in the above and each one would be identical. If n_2 is the number of 2's then there are $n_2!$ such identical re-orderings. Likewise there are $n_5!$ indistinct re-orderings of the 5s. Adding all the distinct permutations together gives

$$P\left(\{1,2,2,5,5\}\right) = \frac{5!}{1!2!2!}(p_1)(p_2)^2(p_5)^2. \tag{2.2.2}$$

The brackets $\{\dots\}$ indicate a set which has no specified order.

You can see that in general if there are k possible outcomes with probabilities p_1, p_2, ... p_k (these are all the possible outcomes so $\sum_i p_i = 1$) the probability of each of these occurring n_1, n_2, ... n_k times in N trials ($\sum_{i=1}^{k} n_i = N$) is

$$P(n_1, n_2, \dots, n_k | N, \{p_i\}) = \frac{N!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \tag{2.2.3}$$

$$= \frac{N!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i} \tag{2.2.4}$$

This is called the **multinomial distribution**. The mean and variance of the distribution are

$$E[x_i] = Np_i \tag{2.2.5}$$

$$Var[x_i] = Np_i(1-p_i) \tag{2.2.6}$$

Problem 3. *You have a set of 6 dice that have been altered so that the probability of rolling a 6 is twice as large as getting any other number. The other numbers have equal probability. If you roll the set of 6 dice what is the probability of getting two 6s, one 5?*

2.3 Birthday Paradox

This is another widely known problem for which many people go down the wrong path and get confused. The "paradox" is that in a relatively small group of people there is a surprisingly high probability that two of them will have the same birthday.

Let's say there are n people at the party. There are 365 choices for the birthday of each person (not including leap years) so there are 365^n combinations of n birthdays. We will assume these are all equally likely. Instead of finding the number

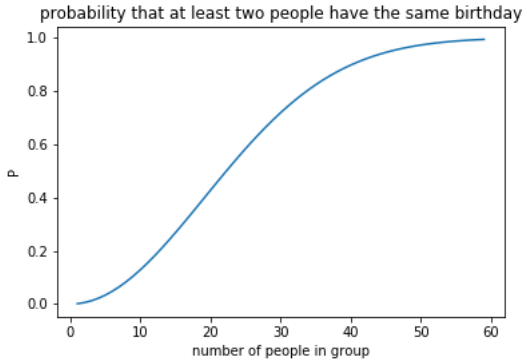

Figure 2.2: Probability of more than one person having the same birthday.

of combinations with repeat birthdays let's find the number of combinations with no repeats. There are 356 choices for the first person, then 364 choices for the second etc. until you get to the last person so the number of cases with no repeats is $365 \times 364 \times ... \times (365 - n + 1) = 365!/(365 - n)!$. So the total probability is

$$P(\text{at least two the same}) = 1 - P(\text{no two the same}) = 1 - \frac{365!}{365^n (365 - n)!}. \quad (2.3.1)$$

If you try to calculate this number directly with a computer you will find that some of these numbers are too big to store. The python scipy factorial function (scipy.special.factorial) will give infinity for 356! for example. But the quotient of these numbers is something reasonable. This problem often comes up in this kind of problem. We will need an approximation to complete the calculation. Taking the log of a quotient often helps you cancel some things out. And taking **Stirling's approximation** ($\ln N! \simeq N \ln N - N$) often helps simplify factorials.

$$\ln\left(\frac{N!}{N^n (N-n)!}\right) = \ln N! - \ln(N-n)! - n \ln N \quad (2.3.2)$$

$$= N \ln N - N - (N-n)\ln(N-n) - (N-n) - n \ln N \quad (2.3.3)$$

$$= (N-n)\ln N - (N-n)\ln(N-n) - n \quad (2.3.4)$$

$$= (N-n)\ln\left(\frac{N}{(N-n)}\right) - n \quad (2.3.5)$$

We can then take the exponential of this to get

$$P(\text{at least two the same}) \simeq 1 - \left(\frac{N}{N-n}\right)^{N-n} e^{-n} \quad (2.3.6)$$

This is plotted in figure 2.2. For a group of 23 people there is a 50% chance that at least 2 of them will have the same birthday.

The multinomial distribution can be used to find the probability of getting exactly so many birthdays on specific days. To find the probability of getting some class of combinations that satisfy a criterion (for example the probability that three birthdays are on the any day or that there are two days with two peoples birthdays on each) you can find the probability for one example of it and then multiply it by the number of ways that the criterion can be satisfied.

Problem 4. *What is the probability of at least 2 people out of n having the same birthday taking leap days into account.*

2.4 Poker

A deck of poker cards consists of 52 cards. There are four suits - diamonds (♢), hearts (♡), spades (♠) and clubs (♣). In each suit there are an ordered sequence of 13 cards (we will take the ace to be greater than the king). A poker hand consists of 5 cards. In "five card stud" you are dealt five cards and you are not allowed to exchange any. This version of poker is almost never played because it relies too much on chance and not skill, but we will consider it here because it is simple.

What is the probability of getting a flush (five cards of the same suit) in five card stud? You might at first think this is just like the dice rolling problem and say it is $4(1/4)^5 \simeq 0.0039$, but this would be wrong because the draws are **not independent**. If your first card is a ♣ there will be fewer ♣ in the deck and the deck will be smaller so the probability of getting a club the second time will be $(13-1)/(52-1)$.

$$P(\text{flush}) = \frac{4}{4}\frac{12}{51}\frac{11}{50}\frac{10}{49}\frac{9}{48} = 0.00198\ldots \tag{2.4.1}$$

Significantly less probable than we would get if there were replacement.

We can also calculate the probability of getting 5 diamonds by repeated applications of the product rule

$$\begin{aligned}
P(\diamondsuit_1, \diamondsuit_2, \diamondsuit_3, \diamondsuit_4, \diamondsuit_5) &= P(\diamondsuit_1)P(\diamondsuit_2, \diamondsuit_3, \diamondsuit_4, \diamondsuit_5|\diamondsuit_1) && (2.4.2)\\
&= P(\diamondsuit_1)P(\diamondsuit_2|\diamondsuit_1)P(\diamondsuit_3, \diamondsuit_4, \diamondsuit_5|\diamondsuit_1, \diamondsuit_2) && (2.4.3)\\
&= P(\diamondsuit_1)P(\diamondsuit_2|\diamondsuit_1)P(\diamondsuit_3|\diamondsuit_1, \diamondsuit_2)P(\diamondsuit_4, \diamondsuit_5|\diamondsuit_1, \diamondsuit_2, \diamondsuit_3) && (2.4.4)\\
&= P(\diamondsuit_1)P(\diamondsuit_2|\diamondsuit_1)P(\diamondsuit_3|\diamondsuit_1, \diamondsuit_2)P(\diamondsuit_4|\diamondsuit_1, \diamondsuit_2, \diamondsuit_3) && (2.4.5)\\
&\quad \times P(\diamondsuit_5|\diamondsuit_1, \diamondsuit_2, \diamondsuit_3, \diamondsuit_4) && (2.4.6)
\end{aligned}$$

Unlike in the case of rolling dice the probabilities are conditional. This expresses the fact that each draw is not independent. The probability of getting any of the

four flushes is then found with the extended sum rule for mutually exclusive events (equation 1.5.15).

What is the probability of a straight? This is getting five sequential cards, for example 8, 9, 10 ,J ,Q. The probability of drawing them all in a row must be the same as the probability of drawing them in any other order so we can calculate the probability of drawing them in order and then multiply by the number of permutations. First we need to draw a card of 10 or lower or there wont be enough cards of higher value. That probability is $4 \times 9/52$ (remember there are no 1 cards). Then there are 4 cards of one higher value out of 51 remaining cards, etc.. Then for each case there are 5! permutations.

$$P(\text{straight}) = 5!\frac{36}{52}\frac{4}{51}\frac{4}{50}\frac{4}{49}\frac{4}{48} = 0.003546\ldots \tag{2.4.7}$$

Somewhat more likely than a flush which is why this hand is worth less. If we count the ace-low straight this is 0.00394.... This includes straight-flushes and royal-straight-flushes which are actually higher hands.

What is the probability of a full house? A full house is two of a kind (two 10's or two kings for example) and three of another kind (three aces or three twos).

Let's do this one a little differently. Let's count the total number of distinct five card hands and then count the number of distinct full houses. The probability will be the ratio of these since every hand is equally probable. Let's make this a little more abstract. There are N distinct objects (cards) we have N ways of choosing the first one. There are $N-1$ objects left when we pick the next one, etc. So there are $N \cdot (N-1) \ldots \cdot (N-n+1)$ distinct ways of choosing n objects out of N. This can also be written $N!/(N-n)!$. This counts combinations of objects in different orders as distinct (123 is different than 213). If we wish to count different permutations of the same objects as the same set then we need to divide by the number of permutations of n objects which is $n!$. Again we get the **binomial coefficient** for the number of these distinct sets

$$\binom{N}{n} \equiv \frac{N!}{n!(N-n)!} \tag{2.4.8}$$

Let's use it on our problem.

There are $\binom{52}{5}$ distinct five card hands. There are four cards of each type, one for each suit, so there are $13 \cdot \binom{4}{2}$ distinct pairs of cards of the same kind. The three of a kind needs to be different than the pair so there are $12 \cdot \binom{4}{3}$ of them. So the probability of a full house is

$$P(\text{full house}) = \frac{\binom{4}{2} \cdot \binom{4}{3} \cdot 13 \cdot 12}{\binom{52}{5}} = 0.00144\ldots \tag{2.4.9}$$

Very similar logic will lead you to the probabilities of getting two pair or four of a kind.

Calculating the probabilities for poker may seem frivolous, but the calculation of odds for gambling actually played a very important role in the development of statistics. Pascal and Fermat had a famous correspondence in the 17th century on games of chance in which they developed the foundations of probability theory.

drawing without replacement, the hypergeometric distribution

Related to poker hands is the case where there are a finite number of objects of two types which are selected at random and not replaced before selecting the next. In this case each trial will not be independent of the ones before it (or the ones after it). We have a bag containing N balls with R red ones and $N - R$ blue ones. The probability of getting r red ones out of n tries *without replacement* is

$$p(r|n, N, R) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \tag{2.4.10}$$

Note that $p(r|1, N, R) = R/N$ and $p(r|N, N, R) = \delta^K_{Rr}$ as they should (δ^K is the Kronecker delta). The probability of a flush in 5 card stud would be $4 \times p(5|5, 52, 13)$ and in 7 card stud $4 \times [p(5|7, 52, 13) + p(6|7, 52, 13) + p(7|7, 52, 13)]$.

Problem 5. Monty Hall Problem *This is a classic problem based on an old American TV game show. It was before my time, but apparently the host of the show was named Monty Hall. There are variations of this game show on Italian TV also. In this game the contestant can choose between three doors. He knows that behind one of the doors is something nice like a new car and behind the other two are things that are not so nice like a chicken or an old shoe. The contestant chooses one door, but does not open it. Monty then eliminates one of the doors that were not chosen and shows that it has the shoe or chicken. The contestant then has a chance to change his choice or remain with his first choice.*

What are the probabilities of getting the prize for each choice?

1. *Stay with the first choice :*

2. *Change doors :*

Problem 6. *You have a bag of 100 blue and yellow balls. 60 of them are blue and 40 of them are yellow.*

$N!$	# of permutations (orderings) of N objects
$\frac{N!}{(N-n)!}$	# of ordered combinations of n out of N objects, no replacement.
$\binom{N}{n} \equiv \frac{N!}{n!(N-n)!}$	# of unordered combinations of n out of N objects, no replacement. Called "N choose n" or the binomial factor.
N^n	# of ordered combinations with replacement, i.e objects can can be repeated but there are only N types.
$\binom{n+N-1}{n}$	# of unordered combinations of n objects out of N possibilities with replacement.

1. *What is the probability of drawing 5 yellow balls in a row out of the bag without looking?*

2. *What is the probability of 6 draws out of 10 being yellow?*

Problem 7. *There are f flavors of gelato. You get a bowl of n scoops. Show that there are*

$$\binom{n+f-1}{n} \tag{2.4.11}$$

combinations of flavors you could order.

Chapter 3

Probability distributions

In this section we will look at some frequently used probability distributions and probability distribution functions (PDFs) and what they are meant to represent. There are many, many named distributions that have been used to model many different things. I will discuss only a few of the most widely applicable distributions that come up very often in statistics. Most other distributions can be derived from these, are limiting cases of these or can be derived using the kind of arguments that I will use to derive them. In practical cases one might need to derive a statistical model that fits the question or the physical theory might dictate a probability distribution for an observable quantity that is not one of the classical distributions. For this reason it is critical that a good scientist have a good understanding of how the classical distributions are derived and what they represent.

3.1 properties of a probability distribution function (PDF)

So far we have considered the probabilities of discrete events - the probability of getting a 5 or 6. If we consider a continuous variable x we can define the probability of being within an infinitesimal range x to $x + dx$ as $p(x)dx$. This probability must be positive.

$$p(x) \geq 0 \tag{3.1.1}$$

There are an infinite number of these bins across the range of x. A measurement of x will be in only one of them so we can apply the sum rule for mutually exclusive events (1.5.16) to these bins. In the infinitesimal limit the sum becomes an integral

$$\int_{-\infty}^{\infty} dx\ p(x) = 1 \tag{3.1.2}$$

All valid PDFs must satisfy these two requirements. Sometimes people call the PDF the **probability mass function**. They mean the same thing.

In the frequentist tradition x is called a **random variable**. A strict Bayesian might avoid using the term. He/she might say that there is an event were the value x is observed and we can attach a probability to this event given our prior knowledge and statistical model. There is no "randomness" about it. I will take a practical approach and ignore the linguistic distinctions as most scientists do.

3.2 mean, median, mode ...

Before we get started with the specific distributions, it will be useful to define some terms and quantities that are used to describe the properties of distributions.

- **cumulative distribution function** - the function of x describing the probability of the measured value being $< x$:

$$F(x) = \int_{-\infty}^{x} dx' p(x') \tag{3.2.1}$$

  By definition $F(-\infty) = 0$ and $F(+\infty) = 1$. The cumulative distribution for a discrete distribution is defined in the obvious way.

- **Quantile function** is the inverse of the cumulative distribution function

$$Q(u) = F^{-1}(u) \quad 0 \leq u \leq 1 \tag{3.2.2}$$

  There is a probability u that the random variable will be $x < Q(u)$.

- **expectation value** - The "average" of any function of the random variable. This is denote by $E[\dots]$ or $\langle \dots \rangle$. The expectation value of $f(x)$ is

$$E\left[f(x)\right] = \langle f(x) \rangle = \begin{cases} \sum_x p(x)\ f(x) \\ \int_{-\infty}^{\infty} dx\ p(x)\ f(x) \end{cases} \tag{3.2.3}$$

- **mode** - A point where a distribution has a maximum. **Unimodal** distributions have one mode and **multimodal** distributions have more than one.

- **median** - The point $x_m = Q(1/2)$, or alternatively $F(x_m) = 1/2$. The probability that x will be less than the median is equal to the probability that it will be more than the median. In a sample or data set the median is the data point that has equal numbers of data points larger than and less than it. For a set with an even number of points the arithmetic mean between the two points closest to having this property is often used.

- **mean** - The mean is the expectation value of the random variable itself, $E[x]$. This will often be represented by μ.
- **moments** - The nth moment of a distribution is $E[x^n]$.
- **central moments** - The nth central moment is $E[(x-\mu)^n]$
- **variance** - The variance is the second central moment $E[(x-\mu)^2]$. It is often denoted by $Var[x]$ or σ^2. This is a measure of the width of the distribution. Note that

$$\begin{aligned} Var[x] = E\left[(x-\mu)^2\right] &= E\left[x^2 - 2x\mu + \mu^2\right] && (3.2.4)\\ &= E\left[x^2\right] - 2E\left[x\right]\mu + \mu^2 && (3.2.5)\\ &= E\left[x^2\right] - \mu^2. && (3.2.6) \end{aligned}$$

- **standard deviation** - the square root of the variance. It is often denotes by σ. An equivalent measure of the width of the distribution in the same units as the random variable.
- **mean deviation** $E[|x-\mu|]$. This is an alternative measure of the width of a distribution. It is often more robustly estimated from a small sample especially when the distribution has large "tails" (much of the probability lies far away from the mode or beyond $\sim \sigma$ from it.).
- **skewness** - $E[(x-\mu)^3]/\sigma^3$. This is a unitless measure of the asymmetry of the distribution.
- **kurtosis** - $E[(x-\mu)^4]/\sigma^4$. This is a measure of the relative importance of outliers (point differing from the mean by larger than several σ). If the kurtosis is larger than 1 the "tails" of the distribution are more important than for a Gaussian. This also reflects the "boxyness" of the distribution.
- **standardized variable** - It is often useful to rescale a random variable with the standard deviation and mean of its distribution

$$X = \frac{(x-\mu)}{\sigma}. \qquad (3.2.7)$$

This variable will always have a mean of 0 and a variance of 1.

Although the moments of a distribution are often used to describe a distribution, and it is true that two distributions with the same moments must be the same distribution, it is possible for a distribution to have no moments. An example of this

that is of particular interest in physics and astronomy it the Cauchy or Lorentzian distribution:

$$p(x) = \frac{\gamma}{\pi\left[(x-x_o)^2+\gamma^2\right]} \quad \text{Cauchy-Lorentz distribution.} \tag{3.2.8}$$

Among other things, this is the natural profile of a spectral line because of the finite lifetime of the excited state. It is also the distribution of the ratio of two normally distributed variable with zero means. Also if you have a point on a plane and you shoot rays out from it in random directions their intercepts with any line not going through the point will have this distribution (Try proving this). It is also the $n = 1$ case of the student-t distribution (section 3.16).

This distribution is normalized and it is symmetric around its mode at $x = x_o$, but the integrals that define all the moments, including the mean, are divergent. Later we will ask what would happen if we tried to estimate the mean or variance using a sample drawn from this distribution.

3.3 changing of variables

Say we have a variable x and the probability of it being between x and $x+dx$ is $p(x)dx$. Now say we have another variable y that is related to x by $x = f(y)$ where $f(y)$ is single valued and differentiable. Then for a change dy, x changes by $dx = \left[\frac{d}{dy}f(y)\right] dy$. The probability of being within this range should not depend on which variable is used to measure the range so it must be that

$$p_x(x)dx = p_x\left(f(y)\right)\left|\frac{df}{dy}\right| dy = p_y(y)dy \tag{3.3.1}$$

In this way the pdf for one variable can be transformed into the pdf for another. For example if the PDF of x is $p_x(x)$, the PDF of $y = x^2$ is $p_y(y) = \frac{1}{2}p_x(\sqrt{y})/\sqrt{y}$.

This is really just the same as a change of variables in an integral of course. For a multivariate pdf variables can be changed in the usual way

$$p_x(x_1, x_2, \dots)dx_1dx_2\cdots = p_x(y_1, y_2, \dots)\left|\frac{\partial x}{\partial y}\right| dy_1dy_2\dots \tag{3.3.2}$$

where $\left|\frac{\partial x}{\partial y}\right|$ is the determinant of the Jacobian matrix relating the volume element in one coordinate system to another.

For example if the probability of a galaxy existing at a point in three dimensional space is $p(x, y, z)dxdydz$ then the probability in spherical coordinates is

$$p\left(x = r\sin(\theta)\cos(\phi), y = r\sin(\theta)\cos(\phi), z = r\cos(\theta)\right)\ r^2\sin(\theta)drd\theta d\phi. \tag{3.3.3}$$

3.4 moment generating function

The **moment generating function** (MGF) of a distribution is defined in the discrete and continuous cases as

$$m_x(t) = \langle e^{tx} \rangle = \begin{cases} \sum_x e^{tx} p(x) \\ \int_{-\infty}^{+\infty} dx \; e^{tx} p(x) \end{cases} \tag{3.4.1}$$

From this we can easily see that the moments of a distribution can be calculated by taking the derivatives of the MGF

$$\left. \frac{d^n m_x(t)}{dt^n} \right|_{t=0} = \langle x^n \rangle \tag{3.4.2}$$

This can be very useful for cases where the MGF can be found analytically. With a change in sign of t this is the same thing as the Laplace transform.

For example, the moment generating function for the binomial distribution is

$$m_x(t) = \sum_{n=0}^{\infty} e^{tn} \binom{N}{n} p^n (1-p)^{N-n} \tag{3.4.3}$$

$$= \sum_{n=0}^{\infty} \binom{N}{n} (e^t p)^n (1-p)^{N-n} \tag{3.4.4}$$

$$= \left(e^t p + 1 - p\right)^N \tag{3.4.5}$$

3.5 characteristic function

The characteristic function is essentially the same thing as the moment generating function only it is the Fourier transform of the PDF instead of the Laplace transform

$$\phi_x(t) = \langle e^{itx} \rangle = \begin{cases} \sum_x e^{itx} p(x) \\ \int_{-\infty}^{+\infty} dx \; e^{itx} p(x) \end{cases} \tag{3.5.1}$$

The only difference is the i.

The characteristic function has the following properties under transformations of the random variable

$$\begin{array}{lcll} \phi_z(t) & = & \phi_x(t/n) & z = x/n \\ \phi_z(t) & = & e^{-it\mu/\sigma} \phi_x(t/\sigma) & z = \frac{x-\mu}{\sigma} \\ \phi_z(t) & = & \phi_x(t)\phi_y(t) & z = x + y \end{array} \tag{3.5.2}$$

where x and y are independently distributed. They are readily derived and recognizable as properties of the Fourier transform. The last one comes from the convolution theorem.

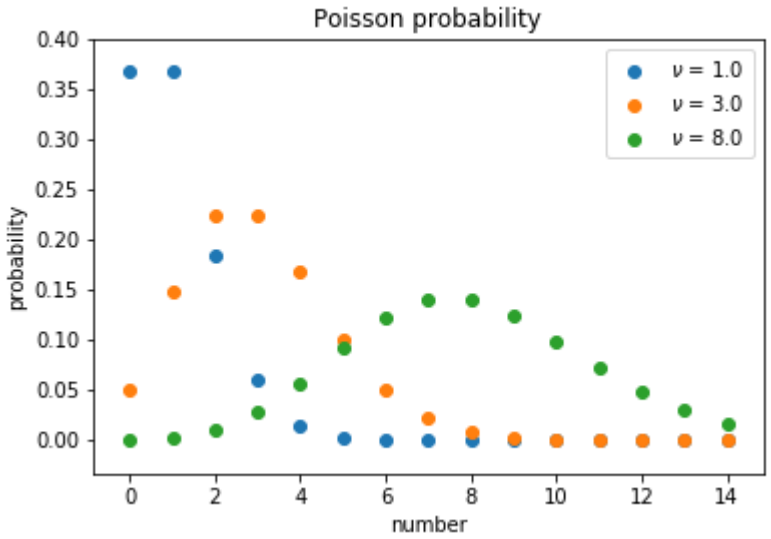

Figure 3.1: The Poisson distribution for several rates ν.

3.6 Poisson distribution

Let's say the probability of an event happening within t and $t + dt$ is a constant rdt. We want to know the probability of N of these events happening within a finite range of time.

First let us find the probability of *no* events happening within a finite range, t_o to $t + dt$. Let's call it $p(0|t_o, t + dt)$. The probability that no event happens between t and $t + dt$ is $1 - rdt$. We can express the joint probability of no events happening in the range t_o to t and no events happening within t to $t + dt$ using the product rule for statistically independent events

$$p(0|t_o, t + dt) = p(0|t_o, t)\,[1 - rdt] \tag{3.6.1}$$

Rearranging this we can obtain the differential equation

$$\frac{p(0|t_o, t + dt) - p(0|t_o, t)}{dt} = \frac{d}{dt}p(0|t_o, t) = -p(0|t_o, t)r \tag{3.6.2}$$

The solution to this is $p(0|t_0, t) = Ae^{-rt}$. We can find the normalization be requiring that $p(0|t_o, t_o) = 1$, there will always be no events in a range of zero length. The results is,

$$p(0|t_o, t) = e^{-r(t-t_o)} \tag{3.6.3}$$

Now for a finite number of events. The probability of n events occurring at ordered times $t_1 \dots t_n$ all less than t (which will also be t_{n+1} in this notation) can also be found by the product rule:

$$p(0 < t_1 < t_2 < \cdots < t_n < t) = p(0|0,t_1) r dt_1 \Theta(t_1 < t_2) \times p(0|t_1,t_2) r dt_2 \Theta(t_2 < t_3) \cdots \times p(0|t_n,t) dt_n \Theta(t_n < t) \tag{3.6.4}$$

$$= r^n e^{-rt} \prod_{i=1}^{n} dt_i \Theta(t_i < t_{i+1}) \tag{3.6.5}$$

where

$$\Theta(x < y) = \begin{cases} 1 & , \quad x \leq y \\ 0 & , \quad x > y \end{cases} \tag{3.6.6}$$

Using the sum rule we know that the probability of n events occurring is the sum of the probabilities for all possible values for the event times.

$$p(n|r,t) = \prod_i \int_0^t dt_i \; p(0 < t_1 < t_2 < \cdots < t_n < t) \tag{3.6.7}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1 \tag{3.6.8}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_3} dt_2 t_2 \tag{3.6.9}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_4} dt_3 \frac{t_3^2}{2} \tag{3.6.10}$$

$$= \frac{(rt)^n}{n!} e^{-rt} \tag{3.6.11}$$

$$= \frac{(\nu)^n}{n!} e^{-\nu} \quad \text{Poisson Distribution} \tag{3.6.12}$$

where $\nu \equiv rt$. This distribution has the following mean and variance

$$E[n] = \nu \tag{3.6.13}$$

$$Var[n] = \nu \tag{3.6.14}$$

The standard example of something that is Poisson distributed is the number of radio active decays within a fixed interval of time. If supernovae go off randomly the probability of seeing one during an hour of observations would be $r(1\,\text{hour})e^{-r(1\,\text{hour})}$ where r would be the total rate of supernovae in the monitored galaxies. Another example is the counts of something, say stars or galaxies, within a volume, or cell, that are uniformly distributed in space. In this case r is the average number density

of objects and t is the volume of the cell. It does not matter what the shape of the cell is. A common question is whether objects are uniformly distributed or clustered. This can be determined by comparing the number counts in cells to the predictions of a Poisson distribution. We will get back to this question later.

As a limit of the binomial distribution

Imagine a cube of space with volume, V, and a smaller cube within it with volume , v. Now imagine there are N uniformly distributed galaxies or stars in this volume. The number of galaxies in v will be n. n would be binomially distributed with the probability of any particular galaxy being in v equal to $p = \frac{v}{V}$.

Now let's take the limit of $N \to \infty$ and $p \to 0$ (or $V \to \infty$) while keeping the average density constant $\nu = N/V = Np$. Using Stirling's approximation one can show that $\frac{N!}{(N-n)!} \simeq N^n$ to lowest order.

$$\binom{N}{n} p^n (1-p)^{N-n} = \binom{N}{n} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \tag{3.6.15}$$

$$= \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \tag{3.6.16}$$

$$= \frac{N^n}{n!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \qquad \text{using } \frac{N!}{(N-n)!} \simeq N^n \tag{3.6.17}$$

$$\simeq \frac{\nu^n}{n!} e^{-\nu} \tag{3.6.18}$$

where I have used $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$. So the Poisson distribution is the binomial distribution in this limit.

A sometimes useful limit of the Poisson distribution when $\nu \gg 1$ is to treat n as continuous and replace $n!$ with the gamma function

$$p(n|\nu) \simeq \frac{\nu^n}{\Gamma(x+1)} e^{-\nu} \qquad \nu \gg 1 \tag{3.6.19}$$

Problem 8. *Consider a random uniform field of stars (or a gas of molecules) with number density η. Using the Poisson distribution find:*

1. *Find the distribution of the distances to the nearest star (molecule).*

2. *What is the average distance to the nearest star (molecule)?*

Problem 9. *Consider a particle traveling through a gas of atoms of density $\rho = m\eta$ at velocity v. If the cross sections for an interaction between the particle and an atom is σ what is the distribution of the distances the particle travels before it interacts with one of the atoms? What is the mean free path?*

3.7 Gaussian or normal

The Gaussian and the normal distribution are two names for the same thing. It is a very widely used probability distribution. The usual justification for this is the central limit theorem although it is also justified as the maximum entropy distribution for a fixed variance. We will get to these justifications later.

The pdf for the Gaussian distribution is

$$p(x|\sigma,\mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \equiv \mathcal{G}\left(x\,|\mu,\sigma\right) \tag{3.7.1}$$

The mean is μ and the variance is σ^2.

A note on notations: To signify that a variable x is normally distributed with a mean of μ and a standard deviation of σ one can write $x \sim \mathcal{N}(\mu,\sigma)$. Sometimes, in an abuse of notation, $\mathcal{N}(\mu,\sigma)$ can stand for the actual pdf (3.7.1). I will use $\mathcal{G}\left(x\,|\mu,\sigma\right)$ to signify this Gaussian function.

The *cumulative distribution function* is

$$F(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \tag{3.7.2}$$

with the **error function** defined as

$$\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2}du \tag{3.7.3}$$

Note that $\mathrm{erf}(-z) = -\mathrm{erf}(z)$ and $\mathrm{erf}(\infty) = 1$.

The *moment generating function* is

$$m_{x-\mu}(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} dx\; e^{tx}e^{-\frac{x^2}{2\sigma^2}} \tag{3.7.4}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} dx\;\exp\left[-\left(\frac{x}{\sqrt{2}\sigma} - \frac{t\sigma}{\sqrt{2}}\right)^2 + \frac{t^2\sigma^2}{2}\right] \tag{3.7.5}$$

$$= e^{\frac{1}{2}\sigma^2t^2} \tag{3.7.6}$$

The moments are

$$\mu_n = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} dx\ x^n e^{-\frac{x^2}{2\sigma^2}} = \begin{cases} \sigma^n(n-1)!! & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \tag{3.7.7}$$

where !! is the **double factorial**,

$$!!n = n \cdot (n-2) \cdot (n-4) \ldots 1 \tag{3.7.8}$$

The probability of x being within $n\sigma$ of the mean is

$$P(\mu - n\sigma \leq x \leq \mu + n\sigma) = 1 - F(\mu - n\sigma) - [1 - F(\mu + n\sigma)] \tag{3.7.9}$$

$$= \frac{1}{2}\left[\text{erf}\left(\frac{n}{\sqrt{2}}\right) - \text{erf}\left(-\frac{n}{\sqrt{2}}\right)\right] \tag{3.7.10}$$

$$= \text{erf}\left(\frac{n}{\sqrt{2}}\right) \tag{3.7.11}$$

some specific values for this are

$$P(-\sigma \leq x - \mu \leq \sigma) = 0.683 \tag{3.7.12}$$

$$P(-2\sigma \leq x - \mu \leq 2\sigma) = 0.954 \tag{3.7.13}$$

$$P(-3\sigma \leq x - \mu \leq 3\sigma) = 0.997 \tag{3.7.14}$$

$$P(-4\sigma \leq x - \mu \leq 4\sigma) = 0.999937 \tag{3.7.15}$$

Problem 10. *If $x \sim \mathcal{N}(0,1)$ what is the distribution of $y = 1/x^2$?*

Problem 11. *Show that the ratio of two normally distributed variables with the same mean is Cauchy distributed.*

3.8 Chebyshev inequality

There is an important and interesting bound on the amount of probability that lies within k times the variance of the mean that does not depend on normality or any other strong assumption about the distribution. If μ and σ are the mean and variance of the distribution the Chebyshev inequality is

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{3.8.1}$$

where

$$P(|X - \mu| \geq k\sigma) = 1 - \int_{\mu - k\sigma}^{\mu + k\sigma} dx\ p(x) \tag{3.8.2}$$

The only requirement on the distribution is that μ and σ exist. k does not need to be an integer.

We can compare this limit to the values for the normal distribution (3.7.12)–(3.7.15)

$$P(-\sigma \leq x - \mu \leq \sigma) \geq 0 \tag{3.8.3}$$
$$P(-2\sigma \leq x - \mu \leq 2\sigma) \geq 0.75 \tag{3.8.4}$$
$$P(-3\sigma \leq x - \mu \leq 3\sigma) \geq 0.888\ldots \tag{3.8.5}$$
$$P(-4\sigma \leq x - \mu \leq 4\sigma) \geq 0.96 \tag{3.8.6}$$

The proof is straightforward,

$$P(|X - \mu| \geq k\sigma) = \int_{-\infty}^{\infty} dx\ p(x)\Theta(|x - \mu| \geq k\sigma) \tag{3.8.7}$$
$$\leq \int_{-\infty}^{\infty} dx\ p(x) \left(\frac{(x-\mu)}{k\sigma}\right)^2 \qquad \left(\frac{(x-\mu)}{k\sigma}\right)^2 \geq \Theta(|x - \mu| \geq k\sigma)\ \forall x \tag{3.8.8}$$
$$\leq \frac{1}{k^2\sigma^2} \int_{-\infty}^{\infty} dx\ p(x)(x - \mu)^2 \tag{3.8.9}$$
$$= \frac{1}{k^2} \tag{3.8.10}$$

This clearly holds for discrete distributions as well.

The Chebyshev inequality allows you to put strict limits of the probability of outliers if the variance is known and implies a more general meaning for the variance of a distribution. It is often used in formal proofs and is a special case of **Markov's inequality** which is $P(|X| > a) \leq E[|X|]/a$ where a is any positive real number.

3.9 central limit theorem

The Gaussian distribution plays an important role in statistics. The distribution of surprisingly large number of phenomena are observed to be well represented by a Gaussian distribution. The traditional explanation for this is the central limit theorem. It holds that the sum of a large number of identically distributed independent random variables will be close to Gaussian distributed even if they are not individually Gaussian distributed. If the noise in a measurement can be considered the sum of many small unknown contributions than you would expect it to be Gaussian distributed.

Let's say we have N identically distributed variables x_i. We can define a set of standardized variables

$$z_i = \frac{x_i - \mu}{\sigma}. \tag{3.9.1}$$

With this scaling it is clear that $\langle z_i \rangle = 0$ and $\langle z_i^2 \rangle = 1$. The sum of these will be $Z = \sum_i z_i$. $\langle Z \rangle = 0$ and $\langle Z^2 \rangle = \sum_{ij} \langle z_i z_j \rangle = \sum_i \langle z_i^2 \rangle = N$ because each one is uncorrelated. So the standardized variable for the sum is

$$Y = \frac{1}{\sqrt{N}} Z = \frac{1}{\sqrt{N}} \sum_i z_i. \tag{3.9.2}$$

This will again have mean zero and variance 1. Now let's find the moment generating function for Y,

$$\begin{aligned}
m_Y(t) = &\langle \exp(tY) \rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} \sum_i z_i \right) \right\rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} z_i \right) \right\rangle^N && (3.9.3)\\
&= \left\langle 1 + \frac{t}{\sqrt{N}} z_i + \frac{t^2}{N} \frac{z_i^2}{2} + \frac{t^3}{N^{3/2}} \frac{z_i^3}{3!} + \dots \right\rangle^N && (3.9.4)\\
&= \left[1 + \frac{t}{\sqrt{N}} \langle z_i \rangle + \frac{t^2}{N} \frac{\langle z_i^2 \rangle}{2} + \frac{t^3}{N^{3/2}} \frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N && (3.9.5)\\
&= \left[1 + \frac{t^2}{2N} + \frac{t^3}{N^{3/2}} \frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N && (3.9.6)\\
&\simeq \lim_{N \to \infty} \left[1 + \frac{t^2}{2N} \right]^N && (3.9.7)\\
&= e^{\frac{t^2}{2}} && (3.9.8)
\end{aligned}$$

This is the moment generating function for a Gaussian as we saw earlier.

It is important to note that this theorem is strictly true only for a sum of an infinite number of variables with the same variance. You might not expect this to apply to our concept of noise coming from many small random contributions that are not all the same. If the variance of one of the variables where much larger than the others it would dominate the distribution of the sum for example. However the Gaussian distribution is widely and successfully used. We will later see another justification for it based on an entropy argument. It can also be shown that many distributions tend toward Gaussian in some limit that is commonly encountered.

The distribution of the sum of independent random variables

Let's do a practical experiment to see how quickly the sum of variables will converge to a Gaussian distribution as the number of variables increases. To do this we will

need the pdf of the sum of random variables. There is a way of doing this that is of general use. Let us take the sum of n random numbers to be $S = \sum_i x_i$. The pdf of variable x_i is $p_i(x_i)$, each one may be different. We can marginalize over all the variables and use a Dirac delta function to force the sum of them to be S

$$p(S) = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \; \delta(S - \sum_i x_i) \; p_1(x_1) \ldots p_n(x_n) \tag{3.9.9}$$

$$= \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \; \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} \exp\left[-ik(S - \sum_i x_i)\right] \; p_1(x_1) \ldots p_n(x_n) \tag{3.9.10}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \int_{-\infty}^{\infty} dx_i e^{+ikxi} p_i(x_i) \tag{3.9.11}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \tilde{p}_i(k) \tag{3.9.12}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \tilde{p}_S(k) \tag{3.9.13}$$

where $\tilde{p}_i(k)$ is the Fourier transform of $p_i(x_i)$. This means that $\prod_i \tilde{p}_i(k)$ is the Fourier transform of the pdf of S. As we already know, the Fourier transform of a probability distribution is call its **characteristic function**. This is true for discrete as well as continuous random variables. In the special case where the distributions are all the same this will be $[\tilde{p}(k)]^n$. Note that in Fourier space the normalization requirement is $\tilde{p}(0) = 1$.

Let's look at a uniform distribution between $-L/2$ and $L/2$. The characteristic function (Fourier transform) of this distribution is

$$\tilde{p}(k) = \frac{1}{L} \int_{-L/2}^{L/2} dx \; e^{+ikx} = \frac{2}{Lk} \sin\left(\frac{kL}{2}\right) = \mathrm{sinc}\left(\frac{kL}{2}\right). \tag{3.9.14}$$

So the pdf for the sum of n uniformly distributed variables, each over a range L/n is

$$p_n(S) = \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \mathrm{sinc}^n\left(\frac{kL}{2n}\right). \tag{3.9.15}$$

Figure 3.2 shows this case for some small values of n with $L = 2$. In this case each x_i has a maximum of 1 so S has a maximum of n. For this reason the tails of the distribution are cut off relative to the Gaussian which extends to infinity. Even so you can see that the distribution becomes remarkably Gaussian even for $n = 5$ or 6.

This exercise can be done numerically for any distribution. It is not necessary to have an analytic expression for the Fourier transform of $p_i(x_i)$. Any numerical DFT

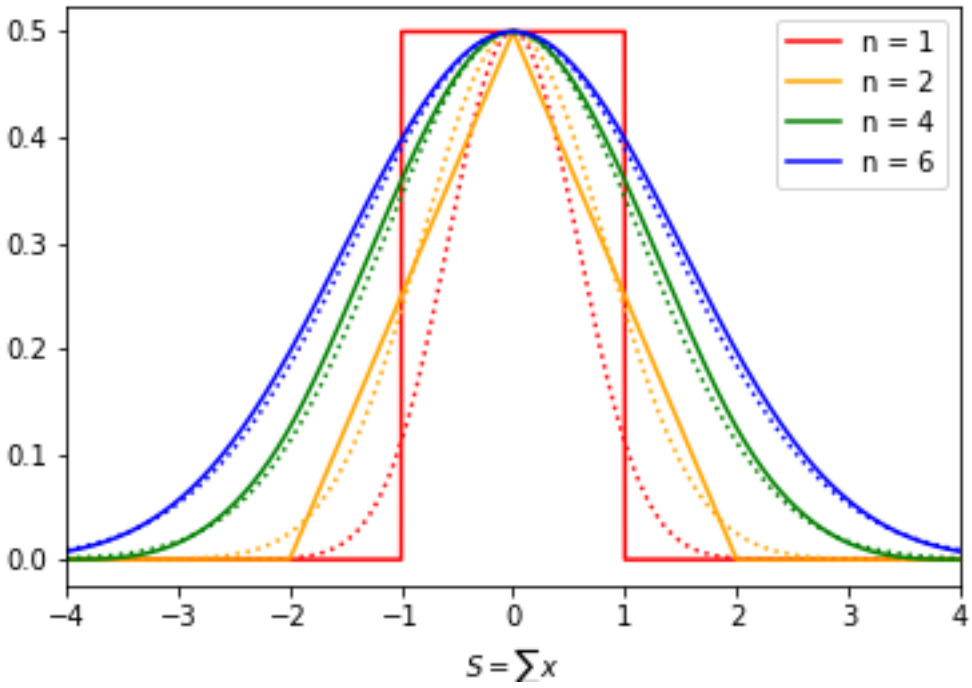

Figure 3.2: Probability distribution for the sum of n random variables that are uniformly distributed between -1 and 1. The normalizations have been changed so that there maximum is 0.5 in all cases. The dotted curves are for Gaussians with the same variance. You can see that the distribution converges to Gaussian remarkably quickly even for a very non Gaussian initial distribution.

(Discrete Fourier Transformation) and inverse DFT will do the same trick although care must be taken with the normalization convention that your software uses and a phase factor that comes in when n is even.

This technique for finding the distribution of the sum of variables can be used to study things like random walks and diffusion. The same idea is also used to derive halo mass functions in cosmology. In general, the **characteristic function** contains all the information contained in the pdf. It is often used in proofs and in calculating certain properties of a distribution.

Problem 12. *If x_1 and x_2 are independent Poisson distributed variables what is the distribution of $S = x_1 + x_2$? What is its mean and variance? Justify your results.*

3.10 connection between Poisson and Gaussian distributions

You can see from the figure 3.1 of the Poisson distribution that as the average gets larger the Poisson pdf gets more symmetric and looks more Gaussian. Let's make this connection more precise. The Poisson distribution is

$$p(n|\nu) = \frac{(\nu)^n}{n!}\, e^{-\nu} \tag{3.10.1}$$

Let's make the substitution $n = \nu(1+\delta)$ which also means $\delta = (n-\nu)/\nu$. Let's take the limit where $\nu \gg 1$ while $\delta \ll 1$ which also means $n \gg 1$. Let's again use the Stirling's approximation

$$n! \sim \sqrt{2\pi n} e^{-n} n^n \tag{3.10.2}$$

This is a more accurate form of the approximation than was used before.

Making this substitution we get the probability

$$p(n) = \frac{\nu^{\nu(1+\delta)} e^{-\nu}}{\sqrt{2\pi} e^{-\nu(1+\delta)} \left[\nu(1+\delta)\right]^{\nu(1+\delta)+1/2}} \tag{3.10.3}$$

$$= \frac{e^{\nu\delta}(1+\delta)^{-\nu(1+\delta)-1/2}}{\sqrt{2\pi\nu}} \tag{3.10.4}$$

Let's look at the lowest order terms of the log of the numerator

$$\ln\left[(1+\delta)^{-\nu(1+\delta)-1/2}\right] = -(\nu(1+\delta)+1/2)\ln(1+\delta) \tag{3.10.5}$$

$$= -\left(\nu+\nu\delta+1/2\right)\left(\delta-\frac{\delta^2}{2}+\ldots\right) \qquad \nu \gg 1 \tag{3.10.6}$$

$$\simeq -\left(\nu+\nu\delta\right)\left(\delta-\frac{\delta^2}{2}+\ldots\right) \tag{3.10.7}$$

$$\simeq -\nu\delta-\frac{\nu\delta^2}{2}+\ldots \tag{3.10.8}$$

Putting this back into the above

$$p(\delta) = p(n) \tag{3.10.9}$$

$$\simeq \sqrt{\frac{\nu}{2\pi}}\, e^{-\frac{\nu\delta^2}{2}}. \tag{3.10.10}$$

So if ν is large the excursion from the mean, δ, is Gaussian distributed with a variance of $1/\nu$. In practice this can be a good enough approximation for moderate values of ν, say greater than 20. The photon noise or **shot noise** in astronomical images is Poisson distributed, but if the photon count is high it is essentially Gaussian distributed.

3.11 lognormal

The lognormal distribution is simply the distribution where the log of the variable is normally distributed instead of the variable itself. This distribution is of particular interest in astronomy because photometric errors are often taken to be Gaussian in magnitudes which is the 2.5 times the log of the flux so the flux will be lognormally distributed. Since the inverse log of a real number cannot be negative the distribution is bounded from below by 0. The distribution is also used to model the distribution of matter in many contexts. Another interpretation is that while the Gaussian is the right distribution for a sum of many random variable, the lognormal is the right one for a product of many random variables.

The pdf comes from just changing variable from the Gaussian

$$p(y)dy = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right\}\frac{dy}{y} & , \quad y>0 \\ 0 & , \quad y\le 0 \end{cases} \tag{3.11.1}$$

Some of its properties are

$$E[y] = \exp(\mu + \frac{\sigma^2}{2}) \tag{3.11.2}$$

$$\text{median[y]} = \exp(\mu) \tag{3.11.3}$$

$$\text{mode[y]} = \exp(\mu - \sigma^2) \tag{3.11.4}$$

$$Var[y] = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2) \tag{3.11.5}$$

If $\mu = 0$ and $\sigma \ll 1$ the distribution is approximately Gaussian with a mean of 1 and a variance of σ^2. So if we take $y = 1 + \delta$ and $\mu = 0$ we have a model for fractional density fluctuations, δ, that will always be positive, will have a median of 0 and will tend to Gaussian when the variance is small. This is, for example, a good model for the Lyman-α absorption in quasar spectra. A multivariable version of this is possible by changing variable from the multivariate Gaussian distribution (section **??**). This is sometimes also used as a model for density fluctuations in the Universe.

Problem 13. *Find the cumulative distribution function for the lognormal distribution in terms of the error function ,* erf().

3.12 Power law distribution

In astronomy it is common to model the distribution of many things (star masses, galaxy luminosities, planet masses, temperatures, densities of clouds, etc.) as a power law. This distribution is also known as a **Pareto distribution**. The integral of a power law diverges either as $x \to 0$ or as $x \to \infty$ so some limits need to be fixed for the distribution to make sense. The normalized PDF is

$$p(x|x_{\min}, x_{\max}, \alpha) = x^\alpha \times \begin{cases} 0 & , \quad x < x_{\min} \\ (\alpha+1)\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]^{-1} & , \quad x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \ln\left(\frac{x_{\max}}{x_{\min}}\right)^{-1} & , \quad x_{\min} < x < x_{\max}, \alpha = -1 \\ 0 & , \quad x > x_{\max} \end{cases} \tag{3.12.1}$$

The cumulative distribution is easily worked out

$$F(x|x_{\min}, x_{\max}, \alpha) = \begin{cases} 0 & , \quad x < x_{\min} \\ \frac{\left[x^{\alpha+1} - x_{\min}^{\alpha+1}\right]}{\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]} & , \quad x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \frac{\ln\left(\frac{x}{x_{\min}}\right)}{\ln\left(\frac{x_{\max}}{x_{\min}}\right)} & , \quad x_{\min} < x < x_{\max}, \alpha = -1 \\ 1 & , \quad x > x_{\max} \end{cases} \tag{3.12.2}$$

Problem 14. *Calculate the mean and variance of the power-law distribution.*

3.13 multivariate distributions

A multivariate distribution is the probability distribution for the joint probability of two or more random variables. Let's number these variable x_1 through x_k. For discrete variable $p(x_1, x_2, \ldots, x_k)$ is the probability that the first variable has the value x_1 *and* the second variable has the value x_2, etc. There is the obvious extension to continous variables where $p(x_1, x_2, \ldots, x_k)dx_1dx_2 \ldots dx_k$ is the probability of all the variable simultaneously being within infinitesimal ranges near those values.

Now the expectation value implies a sum or integral over all the variables. For an arbitrary function $f(x_1, x_2, \ldots, x_k)$

$$E[f(x_1, x_2, \ldots, x_k)] = \int \cdots \int dx_1 \ldots dx_k \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \quad (3.13.1)$$

$$= \prod_{i=1}^{k} \int dx_i \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \quad (3.13.2)$$

This is also written $\langle f(x_1, x_2, \ldots, x_k) \rangle$ or $\overline{f(x_1, x_2, \ldots, x_k)}$. The probability distribution is normalized so $E[1] = 1$.

The average and variance of each variable is defined in the same way as for a distribution of one variable. In this case there is also the **covariance** between two variable

$$C_{ij} = Cov[x_i x_j] \equiv E[(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)] \quad (3.13.3)$$

If the covariance is greater than zero it means that they both tend to be high *and/or* low relative to their means simultaneously. If the covariance is negative one tends to be high while the other is low and vice versa.

C_{ij} is called the **covariance matrix**. You can see that by construction it is symmetric, $C_{ij} = C_{ji}$ and that the diagonal components $C_{ii} = E[(x_i - \bar{x}_i)^2]$ are positive which together mean its eigenvalues are positive or zero. Later we will talk about the covariance matrix of parameters and of data, two different covariance matrices which can be confusing. The covariance matrix is always positive definite (see appendix A.1). The inverse of the covariance matrix, $\boldsymbol{C}^{-1}$, is sometimes called the **precision matrix** .

Change the units for the variables will change the value of their covariance so to better measure the degree of correlation it is convenient to normalize the variance so that it is unitless,

$$\rho_{xy} \equiv \frac{C_{xy}}{\sigma_x \sigma_y} \quad (3.13.4)$$

This is called the **correlation coefficient** or **Pearson's correlation coefficient** *of the distribution.*

$Cov[xy]$ satisfies all the requirements of an inner (or "dot" or "scalar") product. One of the results of this is that covariance satisfies the **Cauchy–Schwarz inequality**

$$|Cov[xy]|^2 \leq Var[x]Var[y] \tag{3.13.5}$$

A result of this is that $-1 \leq \rho_{xy} \leq 1$.

Another important relation is

$$C_{xy} = E[xy] - \bar{x}\bar{y} \tag{3.13.6}$$

which is an extension to the relation we already saw for the variance (3.2.6).

Two variables, x and y, are said to be **correlated variables** if $Cov[xy] \neq 0$. Otherwise they are uncorrelated. Two variables that are **independent** variables are also uncorrelated, but uncorrelated variables are not necessarily independent. Variable with a negative covariance can be called **anticorrelated**.

Some other properties of $\boldsymbol{C}$ are

- $\boldsymbol{C}$ is symmetric, $C_{ij} = C_{ji}$.
- $C_{ii} \geq 0$ for all i
- the eigenvalues are ≥ 0
- $\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} \geq 0$ for all non zero $\boldsymbol{x}$, i.e. $\boldsymbol{C}$ is a **positive semi-definite matrix**
- $\boldsymbol{C} = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{M}^T$ where $\boldsymbol{\Lambda}$ is diagonal and $\boldsymbol{M}$ is an orthogonal matrix, $\boldsymbol{M}^{-1} = \boldsymbol{M}^T$

Problem 15. *If x is uniformly distributed between -1 and 1, i.e. $x \sim U(-1, 1)$ show that the variable $y = x^2$ is uncorrelated with x, but not independent of x.*

Problem 16. *Consider points on a plane that are uniformly distributed within a circle. Are the x and y coordinates correlated? Are they independent?*

The **multinomial distribution** has a covariance of

$$Cov[x_i x_j] = \begin{cases} Np_i(1-p_i) & i = j \\ -Np_i p_j & i \neq j \end{cases} \tag{3.13.7}$$

The negative value reflects the the property that if x_i is larger than its mean, for a fixed N, x_j is more likely to be below its mean and vice versa. If the units are not distributed exactly according to their means then getting more in one bin implies there are less in others.

Example :
Let us consider a pixelized image of the sky taken through a telescope. The signal coming from one pixel will be modeled as

$$f_i = W_{ij}(s_j + n_j) + b + N_i \tag{3.13.8}$$
$$\tag{3.13.9}$$

or in matrix notation

$$\boldsymbol{f} = \boldsymbol{W}(\boldsymbol{s} + \boldsymbol{n}) + \boldsymbol{b} + \boldsymbol{N} \tag{3.13.10}$$

$\boldsymbol{s}$ is the signal coming from the sky and $\boldsymbol{n}$ is some noise coming from the atmosphere or some other source outside of the telescope. $\boldsymbol{b}$ is a background that contributes equally to all pixels. We have tried to subtract it but there is some uncertainty in that subtraction. $\boldsymbol{N}$ is noise coming from inside the telescope and camera. This noise might or might not be correlated between pixels. $\boldsymbol{W}$ is a matrix which quantifies the point spread function (psf), the smearing of the image by the resolution of the telescope. Generally $\boldsymbol{W}$ will not be square so that there are more pixel for $\boldsymbol{s}$ and $\boldsymbol{n}$ than there are in the final image. This is to approximate a continuous sky.
If all the noises have zero mean then

$$\langle \boldsymbol{f} \rangle = \boldsymbol{W}\boldsymbol{s} \tag{3.13.11}$$

since we have subtracted the background. This is a blurred version of the true sky. The covariance is

$$\boldsymbol{C} = \left\langle (\boldsymbol{f} - \langle \boldsymbol{f} \rangle)(\boldsymbol{f} - \langle \boldsymbol{f} \rangle)^T \right\rangle \tag{3.13.12}$$
$$= \left\langle (\boldsymbol{W}\boldsymbol{n} + \boldsymbol{b} + \boldsymbol{N})(\boldsymbol{W}\boldsymbol{n} + \boldsymbol{b} + \boldsymbol{N})^T \right\rangle \tag{3.13.13}$$
$$= \boldsymbol{W}\left\langle \boldsymbol{n}\boldsymbol{n}^T \right\rangle \boldsymbol{W}^T + \left\langle \boldsymbol{b}\boldsymbol{b}^T \right\rangle + \left\langle \boldsymbol{N}\boldsymbol{N}^T \right\rangle \tag{3.13.14}$$

All the cross terms are zero because the noises are independent and the averages of the noises are zero.
If $\boldsymbol{n}$ and $\boldsymbol{N}$ are uncorrelated between pixels this reduces to

$$\boldsymbol{C} = \sigma_n^2 \, \boldsymbol{W}\boldsymbol{W}^T + \sigma_b^2 \, \mathbf{1} + \sigma_N^2 \, \boldsymbol{I} \tag{3.13.15}$$

where $\mathbf{1}$ is the matrix with 1s in every entry. You can see that event when the noise is not correlated the psf will correlate the noise in nearby pixels and the background will correlate all the pixels with each other.

3.13.1 Principle components

Principle components play an important role in many data analysis problems as well as in machine learning and data compression. The principle components are the components of the random vector $\boldsymbol{x}$ after it has been transformed, or rotated, into the basis of the eigenvectors of the covariance $\boldsymbol{C}$. Explicitly:

$$\boldsymbol{y} = \boldsymbol{U}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{3.13.16}$$

The principle components are uncorrelated, $\langle \boldsymbol{y}\boldsymbol{y}^T \rangle = \boldsymbol{\Lambda}$, i.e. $\langle y_i y_j \rangle = 0$ for $i \neq j$. Graphically these principle components are orthogonal vectors in the space in which $\boldsymbol{x}$ lives. In the two dimensional case, they will tend to lineup with the directions in which the distribution has the most variance and the least variance. The variance of the components represent how spread out or centrally concentrated the distribution is in that direction.

It is always possible to find a set of principle components as long as the distributions second moments exist. If one or more of $\boldsymbol{C}$'s eigenvalues are zero, in directions of the corresponding eigenvectors the distribution has no variance so the distribution is constrained to a lower dimensional space. If two or more of the eigenvalues are equal, the principle components are not unique, but otherwise they are.

PCA is useful for several general reasons. One is that some principles components may have much less noise in them than others. Rather than use $\boldsymbol{x}$ in your analysis you may just use the principle components with the lease amount of noise in them. There is the added bonus they are uncorrelated so their covariance is diagonal which makes many calculations easier. This is a kind of data compression where the noise parts of the data are discarded.

The distribution of $\boldsymbol{x}$ need not be just from noise however. You might have a data set that has a distribution of intrinsic properties that you are interested in. These could the the velocity dispersion, luminosity and size of galaxies or they could be number of times a person visits different websites. The PCs with a large variance represent linear combinations of these properties which are widely varying in the population. PCs with small variance are linear combinations that are closer to constant. In this case the PCs with large variance might be used to differentiate between objects. They are the directions of most variation so individual objects might be well characterized by just a few high variance PCs rather than needing to fully specify its position in the original space. In this way we can obtain a different sort of data compression and in some cases, when the covariance is estimated from the data, discover some connections between variables which might not seem related at first.

Example :
Consider two pixels from the last example that are far enough apart that the psf does not correlate them. Their covariance can be represented as

$$\boldsymbol{C} = \begin{pmatrix} \sigma_N^2 + \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_N^2 + \sigma_b^2 \end{pmatrix} \tag{3.13.17}$$

The eigenvectors are $(1,1)/\sqrt{2}$ and $(1,-1)/\sqrt{2}$ so the principle components would be $(f_1+f_2)/\sqrt{2}$ and $(f_1-f_2)/\sqrt{2}$. The variance of these are $\sigma_N^2+2\sigma_b^2$ and σ_N^2.
The whole image, or any part of it, will have its own PCs that will take into account the psf. Some of these PCs could have much lower variance than others.

3.14 multivariate gaussian

The mutivariant Gaussian or normal distribution is by far the most often used multivariate distribution. It is a good approximation to many natural phenomena and is often used even when it is not. It is also often useful when trying to understand some statistical argument or principle to put in a multivariate Gaussian because often an analytic result can be obtained with it while it cannot in general. For these reasons it is essential for any good student of statistics to have a good intuitive understanding of and the ability to easily manipulate the multivariate normal distribution. I will go through some of its important properties and examples.

At this point it will be useful to use matrix notation. The n random variables will be grouped into a vector $\boldsymbol{x}$. The pdf of the multivariate Gaussian is a generalization of the one dimensional Gaussian pdf.

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{C}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{3.14.1}$$

$$\equiv \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu},\boldsymbol{C}\right) \tag{3.14.2}$$

where $\boldsymbol{C}$ is a n-by-n matrix and $\boldsymbol{\mu}$ is an n dimensional vector of parameters. $|\boldsymbol{C}|$ is the determinant of $\boldsymbol{C}$. This will define the function $\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu},\boldsymbol{C}\right)$. To signify that $\boldsymbol{x}$ is distributed in this way we write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{C})$ just like for the one dimensional case. $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

Theorem 3.14.1 *The means of the multivariate Gaussian are*

$$E[x_i] = \mu_i \quad or \quad E[\boldsymbol{x}] = \boldsymbol{\mu} \tag{3.14.3}$$

Theorem 3.14.2 *And the covariances of the mutivariate Gaussian are*

$$Cov[x_i x_j] = E[(x_i - \mu_i)(x_j - \mu_j)] = C_{ij} \quad or \quad Cov[\boldsymbol{x}\boldsymbol{x}] = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] = \boldsymbol{C} \tag{3.14.4}$$

Theorem 3.14.3 *Any linear transformation of Gaussian distributed variables are also Gaussian distributed.*

So $\boldsymbol{C}$ is the correlation matrix as the choice of notation suggests. For the **special case of a diagonal covariance matrix**, the diagonal elements are the σ^2's. The covariance matrix will take the form

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 & \dots \\ 0 & \sigma_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.14.5}$$

In this case there are no correlations between different variables.

PROOF OF MEAN: (theorem 3.14.1)
Let's calculate the means first

$$E[x_i] = \int_{-\infty}^{\infty} dx_i \dots \int_{-\infty}^{\infty} dx_i \dots \int_{-\infty}^{\infty} dx_n \; x_i \; p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.14.6}$$

$$\tag{3.14.7}$$

We can change variable to a set where $\boldsymbol{x}' = \boldsymbol{x} - \mu$ and all the others are unchanged. This will make $\boldsymbol{\mu}$ get substituted for $\boldsymbol{\mu}'$ which is the zero vector $mu_i' = 0$,

$$E[x_i] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_i' \dots \int_{-\infty}^{\infty} dx_n \; (\mu_i + x_i') \; p(\boldsymbol{x}'|\boldsymbol{\mu}', \boldsymbol{C}) \tag{3.14.8}$$

$$= \mu_i \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_i' \dots \int_{-\infty}^{\infty} dx_n \; p(\boldsymbol{x}'|\boldsymbol{\mu}', \boldsymbol{C}) + \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_i' \dots \int_{-\infty}^{\infty} dx_n \; x_i' \; p(\boldsymbol{x}'|\boldsymbol{\mu}', \boldsymbol{C}) \tag{3.14.9}$$

The first set of integrals must be 1 because the pdf is normalized. The second set must be zero because $p(\boldsymbol{x}'|0, \boldsymbol{C})$ is symmetric ($p(-\boldsymbol{x}'|0, \boldsymbol{C}) = p(\boldsymbol{x}'|0, \boldsymbol{C})$) and x_i' is antisymmetric.

PROOF OF VARIANCE: (theorem 3.14.3)

$$Cov[\boldsymbol{x}, \boldsymbol{x}] = \int_{-\infty}^{\infty} d^n x \ (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T \ p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.14.10}$$

$$= \int_{-\infty}^{\infty} d^n z \ \boldsymbol{z}^T \boldsymbol{z} \ p(\boldsymbol{z}|0, \boldsymbol{C}) \qquad \boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu} \tag{3.14.11}$$

Because $\boldsymbol{C}$ is a symmetric, positive definite matrix there exists a **eigendecomposition**

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{-1} \tag{3.14.12}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix whose elements are the eigenvalues and M is an **orthogonal matrix** which means that

$$\boldsymbol{U}^T = \boldsymbol{U}^{-1} \tag{3.14.13}$$

$$|\boldsymbol{U}| \equiv \det(\boldsymbol{U}) = 1 \tag{3.14.14}$$

The columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{C}$. Note also

$$\boldsymbol{\Sigma} = \boldsymbol{U}^T \boldsymbol{C} \boldsymbol{U} \tag{3.14.15}$$

Using this we can change variables into $\boldsymbol{y} = \boldsymbol{U}^{-1}\boldsymbol{z}$,

$$e^{\frac{1}{2}\boldsymbol{z}^T\boldsymbol{C}^{-1}\boldsymbol{z}} \ d^n z = e^{\frac{1}{2}\boldsymbol{z}^T\boldsymbol{U}^{-1}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{z}} \ d^n z = e^{\frac{1}{2}(\boldsymbol{U}^T\boldsymbol{z})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}^T\boldsymbol{z})} \ d^n z = e^{\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}} \ |\boldsymbol{U}| d^n y = e^{\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}} \ d^n y \tag{3.14.16}$$

So the components of $\boldsymbol{y}$ are all independent and we can use the one dimensional Gaussian distribution to calculate the mean of each component

$$Cov[\boldsymbol{x}, \boldsymbol{x}] = Cov[\boldsymbol{z}, \boldsymbol{z}] \tag{3.14.17}$$

$$= \int_{-\infty}^{\infty} d^n z \ \boldsymbol{z}\boldsymbol{z}^T \ p(\boldsymbol{z}|0, \boldsymbol{C}) \tag{3.14.18}$$

$$= \int_{-\infty}^{\infty} d^n y \ (\boldsymbol{U}\boldsymbol{y})(\boldsymbol{U}\boldsymbol{y})^T \ p(\boldsymbol{y}|0, \boldsymbol{\Sigma}) \tag{3.14.19}$$

$$= \int_{-\infty}^{\infty} d^n y \ \boldsymbol{U}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{U}^T \ p(\boldsymbol{y}|0, \boldsymbol{\Sigma}) \tag{3.14.20}$$

$$= \boldsymbol{U} \int_{-\infty}^{\infty} d^n y \ \boldsymbol{y}\boldsymbol{y}^T \ p(\boldsymbol{y}|0, \boldsymbol{\Sigma})\boldsymbol{U}^T \tag{3.14.21}$$

$$= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T \tag{3.14.22}$$

$$= \boldsymbol{C} \tag{3.14.23}$$

If all the eigenvalues of $\boldsymbol{C}$ are nonzero then it is invertible. This will always be the case for a proper Gaussian distribution since otherwise it would not be normalizable. However these cases do sometimes come up for distributions in parameters space. More on this later.

The third order central moments and all other odd ordered central moments are zero. The fourth order central moments are given by

$$\langle x_i x_j x_k x_l \rangle = C_{ij}C_{kl} + C_{ik}C_{jl} + C_{il}C_{jk} \tag{3.14.24}$$

In general then even central moments are given by Isserlis's theorem :

$$\langle x_1 x_2 \dots x_n \rangle = \sum_P \prod_{\text{pairs ij in } P} C_{ij} \tag{3.14.25}$$

where the sum is over all distinct ways of breaking the n variables into pairs and the product is over those pairs. There will be $(k-1)!/(2^{k/2-1}(k/2-1)!$ terms for the kth order moments although if some of the variables are repeated some of the term will be the same. In the context of quantum field theory this is known as Wick's theorem .

Conditional Gaussian distribution

Let's break the parameters, $\boldsymbol{x}$, into two sets, $\boldsymbol{y}$ and $\boldsymbol{z}$. We will fix the parameters $\boldsymbol{z}$ and ask what the pdf for the parameters $\boldsymbol{y}$ is, $p(\boldsymbol{y}|\boldsymbol{z})$. If the covariance matrix is diagonal then $p(\boldsymbol{y}|\boldsymbol{z})$ is clearly Gaussian. When the covariance is not diagonal the distribution of $\boldsymbol{y}$ is still Gaussian distributed, but with a different covariance and mean.

Let's partition the covariance matrix into a part that involves only components of $\boldsymbol{y}$, $\mathbf{C}_{yy}$, a part that involves only components of $\boldsymbol{z}$, $\mathbf{C}_{zz}$ and a component that involves mixtures of the two, $\mathbf{C}_{xy}$.

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{zy} \\ \mathbf{C}_{zy}^T & \mathbf{C}_{zz} \end{bmatrix} \tag{3.14.26}$$

The conditional pdf is then

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}'_y, \boldsymbol{\Sigma}_{yy}\right) \qquad \begin{cases} \boldsymbol{\mu}'_y = \boldsymbol{\mu}_y + \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{yy} = \boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T \end{cases} \tag{3.14.27}$$

which means

$$p(\boldsymbol{y}|\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^{D_y}|\boldsymbol{\Sigma}_{yy}|}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}'_y)^T\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}'_y)} \tag{3.14.28}$$

$$\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}_y - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right)^T\left(\boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_y - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right)\right] \tag{3.14.29}$$

Marginalized Gaussian distribution

If we integrate over the parameters $\boldsymbol{z}$ we get the marginal distribution

$$p(\boldsymbol{y}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{x}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{y}, \boldsymbol{z}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}) \tag{3.14.30}$$

Using the same definitions (without proof) this is

$$p(\boldsymbol{y}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}_y, \boldsymbol{C}_{yy}\right) \tag{3.14.31}$$

So the correlation with $\boldsymbol{z}$ drop out.

The proof for the conditional and marginal distributions in the general case are rather long algebraically. I wont go through it, but one step in it is an identity that will be useful in manipulating covariance matrices. This is the matrix **completion of squares** formula

$$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \mathbf{b}^T\boldsymbol{x} = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\mathbf{b}\right)^T \boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\mathbf{b}\right) - \frac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b} \tag{3.14.32}$$

for a symmetric and invertible $\boldsymbol{A}$ which is the matrix equivalent of the scalar formula $ax^2 + bx = a(x + \frac{b}{2a})^2 - \frac{b^2}{4a}$.

Combining two multivariate Gaussians

$$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_1, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_2, \boldsymbol{C}_2\right) = \mathcal{G}\left(\boldsymbol{\mu}_1\,|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) \tag{3.14.33}$$

$$\boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2 \tag{3.14.34}$$

$$\boldsymbol{\mu}_c = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{C}_1\boldsymbol{\mu}_1 + \boldsymbol{C}_2\boldsymbol{\mu}_2\right) \tag{3.14.35}$$

In particular if

$$\boldsymbol{C}_1 = \sigma_1^2 \quad \text{and} \quad \boldsymbol{C}_2 = \sigma_2^2 \tag{3.14.36}$$

then

$$\begin{aligned} \boldsymbol{C}_1^{-1} &= \tfrac{1}{\sigma_1^2} \quad \text{and} \quad \boldsymbol{C}_2^{-1} = \tfrac{1}{\sigma_2^2} \\ \boldsymbol{\Sigma} &= \sigma_1^2 + \sigma_2^2 \\ \boldsymbol{\Sigma}^{-1} &= (\sigma_1^2 + \sigma_2^2)^{-1} \\ \boldsymbol{\mu}_c &= \tfrac{\mu_1\sigma_1^2 + \mu_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \end{aligned} \tag{3.14.37}$$

A particularly important application of this rule is for the distribution of the sum of two independent Gaussian distributed variables.

Theorem 3.14.4 *If* $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{C}_1)$ *and* $\boldsymbol{x}' \sim \mathcal{N}(0, \boldsymbol{C}_2)$ *and their sum is* $\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x}'$ *then* $\boldsymbol{s} \sim \mathcal{N}(0, \boldsymbol{C}_1 + \boldsymbol{C}_2)$.

Let's call them $\boldsymbol{x}$ and $\boldsymbol{x}'$ and their sum $\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x}'$.

$$
\begin{aligned}
p(\boldsymbol{s}) &= \int_{-\infty}^{\infty} d^n x \int_{-\infty}^{\infty} d^n x' \; p(\boldsymbol{s}, \boldsymbol{x}, \boldsymbol{x}') && (3.14.38)\\
&= \int_{-\infty}^{\infty} d^n x \int_{-\infty}^{\infty} d^n x' \; p(\boldsymbol{x}, \boldsymbol{x}') p(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{x}') && (3.14.39)\\
&= \int_{-\infty}^{\infty} d^n x \int_{-\infty}^{\infty} d^n x' \; p(\boldsymbol{x}, \boldsymbol{x}') \delta^D(\boldsymbol{s} - \boldsymbol{x} - \boldsymbol{x}') && (3.14.40)\\
&= \int_{-\infty}^{\infty} d^n x \; p(\boldsymbol{x}, \boldsymbol{s} - \boldsymbol{x}) && (3.14.41)\\
&= \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{x} \,|0, \boldsymbol{C}_1\right) \mathcal{G}\left(\boldsymbol{s} - \boldsymbol{x} \,|0, \boldsymbol{C}_2\right) && (3.14.42)\\
&= \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{x} \,|0, \boldsymbol{C}_1\right) \mathcal{G}\left(\boldsymbol{x} \,|\boldsymbol{s}, \boldsymbol{C}_2\right) && (3.14.43)\\
&= \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{x} \,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) \mathcal{G}\left(\boldsymbol{s} \,|0, \boldsymbol{\Sigma}\right) && (3.14.44)\\
&= \mathcal{G}\left(\boldsymbol{s} \,|0, \boldsymbol{\Sigma}\right) \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{x} \,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) && (3.14.45)\\
&= \mathcal{G}\left(\boldsymbol{s} \,|0, \boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2\right) && (3.14.46)
\end{aligned}
$$

3.15 χ^2 distribution

The χ^2 distribution is not a multivariate distribution, but is closely related to the multivariate Gaussian. Consider a multivariate Gaussian distribution with uncorrelated variable, or equivalently a diagonal covariance. Let's define a new variables

$$
z = \sum_i^n \frac{(x_i - \mu_i)^2}{\sigma_i^2}. \tag{3.15.1}
$$

z is often called χ^2. This can be confusing because the random variable is not χ, but $z = \chi^2$. We want to change variables from $x_1, x_2, \ldots$ to z. The Gaussian distribution

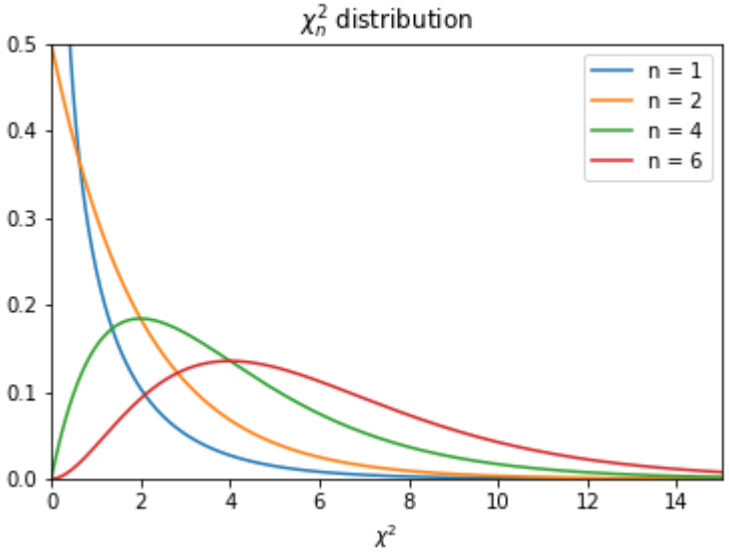

Figure 3.3: χ^2_n distribution for some different degrees of freedom, n.

is

$$p(x_1, x_2, \dots x_n)dx_1 \dots dx_n = \frac{1}{(2\pi)^{n/2}\prod_i \sigma_i} e^{-\frac{1}{2}\sum_i^n \frac{(x_i-\mu_i)^2}{\sigma_i^2}} dx_1 \dots dx_n \tag{3.15.2}$$

$$= \frac{1}{(2\pi)^{n/2}\prod_i \sigma_i} e^{-\frac{1}{2}z} dx_1 \dots dx_n \tag{3.15.3}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i'^2} dx_1' \dots dx_n' \qquad x' = \frac{x-\mu}{\sigma} \tag{3.15.4}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}z} dx_1' \dots dx_n' \tag{3.15.5}$$

Since the Cartesian distance in x'-space is $\sqrt{\sum_i x_i'^2}$, z can be seen as the square of the radial coordinate in n dimensional space

$$dx_1' \dots dx_n' = r^{n-1} dr d\theta_1 d\theta_3 \cdots = \frac{1}{2} z^{n/2-1} dz d^n\Omega \tag{3.15.6}$$

Because the pdf is a function of only the z coordinate we can integrate, marginalize, over the angular coordinates which will result in a n dependent normalization constant. The final pdf is

$$p(z = \chi^2|n) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{3.15.7}$$

where the **gamma function** is defined as

$$\Gamma(x) \equiv \int_0^\infty dt\ e^{-t}t^{x-1}. \tag{3.15.8}$$

This is called the "χ^2 distribution of n degrees of freedom". It will be very important for calculating the significance of Gaussian distributed data. The *mean* of this distribution is $E[x] = n$ and the variance $Var[x] = 2n$. For this reason the value of χ^2_n/n is often given and compared to 1. The *mode* is $x =$max$(n-2, 0)$ so $\chi^2_n/n = 1$ is not actually the most likely value. The *skewness* is $\sqrt{8/n}$ so as n increases the pdf becomes more symmetric. The pdf is plotted in figure 3.3.

The cumulative distribution function can be written down in terms of other special functions without much insight coming from it except in the special case of $n = 2$ where it is

$$F(x|2) = 1 - e^{-x/2} \tag{3.15.9}$$

The general case is of course available in any statistical software package.

Theorem 3.15.1 *If $x_1 \sim \chi^2_{n_1}$, $x_2 \sim \chi^2_{n_2}$ and $s = x_1 + x_2$ then $s \sim \chi^2_{n_1+n_2}$.*

This can be proven in a similar way to how it was shown that the some of squares of Gaussian distributed variables is $\sim \chi^2$. Or with the characteristic function for the χ^2 distribution which is

$$\phi(t) = E[e^{ikx}] = (1-2it)^{-n/2} \tag{3.15.10}$$

Problem 17. *Prove theorem 3.15.1.*

Problem 18. *The velocity distribution for particles in an ideal gas is a multivariate Gaussian in three dimensions, (v_1, v_2, v_3). Using the above find the distribution for the particles' kinetic energy.*

3.16 student's t-distribution

Yet another distribution that comes up often is the student's t-distribution (or just the t-distribution). We will see that this is used to test if the means of two distributions are the same when the variance in each is not known. The pdf is

$$p(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left[1 + \frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}} \tag{3.16.1}$$

This distribution has a mean and mode at zero. It is symmetric about this point. Variance is $\frac{\nu}{\nu-2}$ for $\nu > 2$. It resembles a Gaussian, but with more weight in the wings, see figure 3.4.

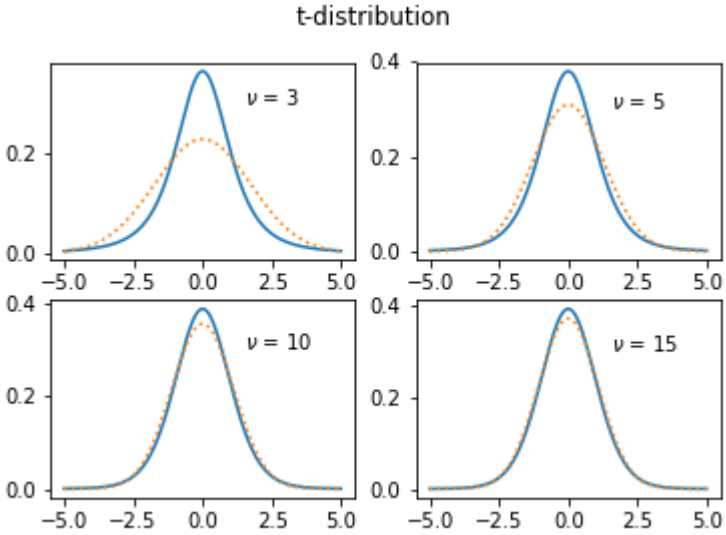

Figure 3.4: Student's t distribution for some different degrees of freedom, ν. The dotted curves are Gaussians with the same variances for comparison.

Chapter 4

Sampling

In the last section we dealt with probability distributions and random variables. The means and variances were the means of variances evaluated by summing (or integrating) over all possible values of the random variables. A random variable is a purely theoretical construction and real data consists of a finite set of observed values. These are *sampled* from the distribution or are a sample of the possible data sets. This is where we move from the purely mathematical subject of probability theory to the practical (and more subjective) field of statistics.

A **statistic** is simply any function of a sample or data points. The arithmetic mean and the sample variance are the simplest example of this. In the case of normally distributed data the probability distribution of these statistics among all possible data sets can be derived analytically. Which makes them an important example and, before computers where widely used one of the few practical statistics.

Fundamental to statistics and its connection to probability is the **law of large numbers**. This holds that for any function of random variables $f(x)$

$$\lim_{n->\infty} \left[\frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] = E_p \left[f(x) \right]. \tag{4.0.1}$$

where x_i are drawn from the distribution $p(x)$. From this follows many of the results of statistics.

To be a bit more precise, the sum above converges to the expectation value "in probability". This is somewhat technical, but the term appears a lot in the statistics literature so let me briefly explain. Let $l_n(\boldsymbol{x})$ be some function of n random variables (or an sequence of random variables). Let $P(|l_n - l| < \epsilon)$ be the probability that l_n is within a ball of radius ϵ centered on the value l. It is said that l_n converges *in probability* to l if for any positive real values ϵ and δ there exists an N such that for all $n \geq N$, $P(|l_n - l| < \epsilon) < \delta$. This means what you think it means, but in a

mathematically precise way. Converging in probability is often denoted with $\xrightarrow{p}$ in the literature. I will usually leave the p out and just use the arrow.

In this chapter we will look at some of the basic properties of a finite sample drawn from a distribution.

4.1 estimating the mean

Say we have a finite sample drawn from a distribution with pdf $p(x|\mu, \sigma)$ where μ is the mean and σ is the standard distribution. Let us say there are N samples denoted $x_1, \ldots x_N$ and they are all independent draws from the distribution.

The **arithmetic mean** or **sample mean** of this data is

$$\bar{x}_N \equiv \frac{1}{N} \sum_{i=0}^{N} x_i \tag{4.1.1}$$

which everyone knows. Confusingly this is usually called just the mean or average just like the mean or average of a distribution, $E[x]$, although it is usually clear from the context which one is meant. $E[x]$ is a sum or integral over all possible values of x weighted by the pdf and $\bar{x}_N$ is an unweighted sum over a finite sample of values.

We can take the expectation value of the arithmetic mean

$$\langle \bar{x}_N \rangle = \frac{1}{N} \sum_{i=0}^{N} \langle x_i \rangle \tag{4.1.2}$$

$$= \frac{1}{N} \sum_{i=0}^{N} \mu \tag{4.1.3}$$

$$= \mu \tag{4.1.4}$$

This means that the arithmetic mean of a sample is an estimate, or an *estimator*, of the mean of the distribution. This is the simplest example of an **unbiased estimator** (its average equals the quantity being estimated). It is not the only estimator of the mean and it is not always the best estimator of the mean.

For a finite sample the arithmetic mean will not always equal the mean of the distribution. One might want to know how good an estimate it is. One way to

quantify this is to calculate the variance of the arithmetic mean,

$$
\begin{aligned}
Var[\bar{x}_N] &= \left\langle [\bar{x}_N - \mu]^2 \right\rangle && (4.1.5)\\
&= \left\langle [\bar{x}_N]^2 \right\rangle - 2\mu\langle \bar{x}_N \rangle + \mu^2 && (4.1.6)\\
&= \left\langle [\bar{x}_N]^2 \right\rangle - \mu^2 && (4.1.7)\\
&= \left\langle \left[\frac{1}{N} \sum_{i=0}^{N} x_i \right]^2 \right\rangle - \mu^2 && (4.1.8)\\
&= \frac{1}{N^2} \sum_{i=0}^{N} \sum_{j=0}^{N} \langle x_i x_j \rangle - \mu^2 && (4.1.9)\\
&= \frac{1}{N^2} \left[\sum_{i=0}^{N} \left\langle x_i^2 \right\rangle + \sum_{i \neq j} \langle x_i x_j \rangle \right] - \mu^2 && (4.1.10)\\
&= \frac{1}{N^2} \left[\sum_{i=0}^{N} (\sigma^2 + \mu^2) + \sum_{i \neq j} \langle x_i \rangle \langle x_j \rangle \right] - \mu^2 && (4.1.11)\\
&= \frac{1}{N^2} \left[N(\sigma^2 + \mu^2) + N(N-1)\mu^2 \right] - \mu^2 && (4.1.12)\\
&= \frac{\sigma^2}{N} && (4.1.13)
\end{aligned}
$$

So you can see that the standard deviation of the mean will go down like $\propto 1/\sqrt{N}$ no matter what the underlying distribution is as long as the mean and variance exist. Of course to calculate this variance we need to know the underlying variance, σ^2, which we often do not know, and can even not exist.

So far we have not made any assumptions about how x is distributed except that the first 2 moments exist. Since the arithmetic mean is a linear function of the data, if the data is normally distributed the arithmetic mean will be normally distributed by theorem 3.14.4.

$$
\text{if} \quad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \sigma) \quad \text{then} \quad \bar{x}_N \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\sigma}{\sqrt{N}}\right) \tag{4.1.14}
$$

It often happens that one is making repeated measurements of something, say the luminosity of a star, and the variance of the noise is not the same for each measurement because the conditions change or you are combining data from different instruments that have different noise levels. Nevertheless the thing you want to know, the luminosity of the star, should be constant. The arithmetic mean (4.1.2) will on average equal μ, but what if one measurement has a lot of noise – σ_i is very large? This

data point will be a less good estimate of the mean than the other points. Including it in the sum might make the estimate worse rather than better!

Consider the estimator

$$\hat{\theta} = \sum_i w_i x_i \tag{4.1.15}$$

which we can call the **weighted mean**. Clearly the average of this, $\left\langle \hat{\theta} \right\rangle$ will equal μ if

$$\sum_i w_i = 1. \tag{4.1.16}$$

We have the freedom to choose these weights subject to this constraint. A good idea is to minimize the variance of the estimator. This will make it the simplest case of a **minimum variance estimator**. The variance of the estimator will be

$$\begin{aligned}
\sigma_\theta^2 &= \left\langle \hat{\theta}^2 \right\rangle - \mu^2 && (4.1.17)\\
&= \left\langle \left[\sum_i w_i x_i \right]^2 \right\rangle - \mu^2 && (4.1.18)\\
&= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 && (4.1.19)\\
&= \sum_i w_i^2 \langle x_i^2 \rangle + \sum_{i \neq j} w_i w_j \langle x_i \rangle \langle x_j \rangle - \mu^2 && (4.1.20)\\
&= \sum_i w_i^2 \left[\sigma_i^2 + \mu^2 \right] + \mu^2 \sum_{i \neq j} w_i w_j - \mu^2 && (4.1.21)
\end{aligned}$$

To minimize the variance we will use the technique of **Lagrange multipliers** which you should know from calculus. We minimize the function

$$F(\boldsymbol{w}) = \sigma_\theta^2(\boldsymbol{w}) + \lambda \left(1 - \sum_i w_i \right) \tag{4.1.22}$$

with respect to the weights. That is

$$\frac{\partial F}{\partial w_k} = \frac{\partial \sigma_\theta^2}{\partial w_k} - \lambda = 0 \tag{4.1.23}$$

The derivative of the variance is

$$\frac{\partial \sigma_\theta^2}{\partial w_k} = 2w_k \left[\sigma_k^2 + \mu^2\right] + 2\mu^2 \sum_{i \neq k} w_i \tag{4.1.24}$$

$$= 2w_k \left[\sigma_k^2 + \mu^2\right] + 2\mu^2 \left[\sum_{i=0}^{N} w_i - w_k\right] \tag{4.1.25}$$

$$= 2w_k \left[\sigma_k^2 + \mu^2\right] + 2\mu^2 \left[1 - w_k\right] \qquad \text{use constraint } \sum_i w_i = 1 \tag{4.1.26}$$

$$= 2w_k\sigma_k^2 + 2\mu^2 \tag{4.1.27}$$

putting this into (4.1.23) gives

$$w_k = \frac{\lambda - 2\mu^2}{2\sigma_k^2} \tag{4.1.28}$$

Plugging this into the constraint (4.1.16) and solving for

$$\lambda = 2\mu + 2\left[\sum_k \frac{1}{\sigma_k^2}\right]^{-1} \tag{4.1.29}$$

so

$$w_k = \left[\sum_i \frac{1}{\sigma_i^2}\right]^{-1} \frac{1}{\sigma_k^2} \tag{4.1.30}$$

So the estimator (4.1.15), the one with the minimum variance, is

$$\hat{\theta} = \frac{1}{\left[\sum_i \frac{1}{\sigma_i^2}\right]} \sum_i \frac{x_i}{\sigma_i^2}. \tag{4.1.31}$$

This is often called **inverse noise weighting**. You can see that a data point with a large σ_i^2 will be down weighted with respect to points that have small σ_i^2.

This can be generalized to the case where the data points are correlated as well, but I will leave that for later when we look at estimators and parameter estimation more generally.

4.2 estimating the variance

Let us go back to the case of N data points sampled from the same distribution. We might want to know the variance of the distribution. This could be the variance from

noise so we can measure how well our apparatus is working or it could be that we are interested in the variance of the "signal" itself that is not constant. For example say we want to characterize ocean waves from discrete measurements of the height of the water's surface. The variance in the height might be a good quantity to measure.

Known mean: If the mean of the underling distribution is known we can estimate the variance of that distribution with

$$S_N^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 \tag{4.2.1}$$

You can easily show that $\langle S_N^2 \rangle = \sigma^2$.

Unknown mean: In most cases one does not know the average ahead of time. In this case the best estimator for a Gaussian distributed x is

$$S_N^2 = \frac{1}{N-1}\sum_i \left(x_i - \bar{x}_N\right)^2 . \tag{4.2.2}$$

Why is there an $N-1$ instead of an N in the denominator? Let's look at the average of it

$$\begin{aligned}
\left\langle S_N^2 \right\rangle &= \frac{1}{N-1}\sum_i \left\langle (x_i - \bar{x}_N)^2 \right\rangle && (4.2.3)\\
&= \frac{1}{N-1}\left[\sum_i \left\langle x_i^2 \right\rangle - 2\left\langle \sum_i x_i \bar{x}_N \right\rangle + \sum_i \left\langle (\bar{x}_N)^2 \right\rangle\right] && (4.2.4)\\
&= \frac{1}{N-1}\left[\sum_i (\sigma^2 + \mu^2) - 2N\left\langle (\bar{x}_N)^2 \right\rangle + N\left\langle (\bar{x}_N)^2 \right\rangle\right] && (4.2.5)\\
&= \frac{1}{N-1}\left[\sum_i (\sigma^2 + \mu^2) - N\left\langle (\bar{x}_N)^2 \right\rangle\right] && (4.2.6)\\
&= \frac{1}{N-1}\left[N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)\right] \qquad \text{using (4.1.13)} && (4.2.7)\\
&= \sigma^2 && (4.2.8)
\end{aligned}$$

So this estimator is unbiased. Note that this does not require that the x's be normally distributed. If there were an N in the denominator of (4.2.2) then $\langle s_N^2 \rangle = (N-1)\sigma/N$ which means it would be **biased**, but since the bias gets smaller as N increases it would be a simple example of an **asymptotically unbiased estimator**.

Theorem 4.2.1 *If $x_i \sim \mathcal{N}(\mu, \sigma)$ and S_N is given by (4.2.2) then $z = \frac{(N-1)S_N^2}{\sigma^2}$ is χ^2_{N-1} distributed.*

half proof:

$$\frac{(N-1)S_N^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_i (x_i - \bar{x})^2 \tag{4.2.9}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i - \mu) - (\bar{x} - \mu)\right]^2 \tag{4.2.10}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\right] \tag{4.2.11}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i - \mu)^2\right] - 2N(\bar{x} - \mu)(\bar{x} - \mu) + N(\bar{x} - \mu)^2 \tag{4.2.12}$$

$$= \sum_i \frac{(x_i - \mu)^2}{\sigma^2} - \frac{N(\bar{x} - \mu)^2}{\sigma^2} \tag{4.2.13}$$

This is the difference of two χ^2 distributed quantities. $(\bar{x}-\mu)^2/(\sigma^2/N) \sim \chi_1^2$ because, as we already saw, $\bar{x}$ is normally distributed. $\sum_i (x_i - \mu)^2 \sim \chi_N^2$ because it is the sum of the squares of N normally distributed numbers. By theorem 3.15.1 the sum of a χ_m^2 distributed variable and a χ_n^2 distributed variable is $\sim \chi_{m+n}^2$ So $\frac{(N-1)S_N^2}{\sigma^2} \sim \chi^2 N - 1$. This assumes that $(\bar{x} - \mu)^2$ and S_N^2 are independent which is not obvious. We will return to this detail in section 8.3 where we will show that this is in fact true.

From what we know about the χ^2, distribution, this means that our statistic S_N^2 has the following properties from

$$\begin{aligned} &\left\langle \tfrac{(N-1)}{\sigma^2} S_N^2 \right\rangle = N - 1 && \Rightarrow \quad \langle S_N^2 \rangle = \sigma^2 \\ &Var\left[\tfrac{(N-1)}{\sigma^2} S_N^2\right] = \left\langle \left(\tfrac{(N-1)}{\sigma^2} S_N^2\right)^2 \right\rangle - \left\langle \tfrac{(N-1)}{\sigma^2} S_N^2 \right\rangle^2 = 2(N-1) && \Rightarrow \quad Var\left[S_N^2\right] = \tfrac{2\sigma^4}{(N-1)} \end{aligned} \tag{4.2.14}$$

So the standard deviation of our estimated variance again goes down like $\sim 1/\sqrt{N}$ for large N. We can also find the probability that S_N^2 will be within some range using the cumulative distribution for a χ^2 distribution

$$P\left(\frac{\sigma^2}{(N-1)} z_1 < S_N^2 < \frac{\sigma^2}{(N-1)} z_2\right) = F_{\chi^2_{N-1}}(z_2) - F_{\chi^2_{N-1}}(z_1) \tag{4.2.15}$$

Measuring the variance of a signal is closely related to measuring the correlation function or the power spectrum of a signal. We will return to that problem later.

4.3 estimating the mean when the variance is unknown

We have learned that $\bar{x}$ is $\mathcal{N}(\mu, \sigma/\sqrt{n})$ distributed if the x_i's are normally distributed. So if we have a measurement and we know the noise, σ, we can put an error on our estimate of the mean $\pm\frac{\sigma}{\sqrt{n}}$. But often we do not know the σ's. We can estimate it with S_n^2, but this estimate is based on the same data as the estimate of $\bar{x}$ and so $\bar{x}$ will *not* be $\mathcal{N}(\mu, S_n/\sqrt{n})$ distributed.

Theorem 4.3.1 *If $x_i \sim \mathcal{N}(\mu, \sigma)$ then*

$$t = (\bar{x} - \mu)\sqrt{\frac{n}{S_n^2}} \tag{4.3.1}$$

is student-t distributed with $n-1$ degrees of freedom.

The t-distribution was introduced in section 3.16.

So if we wanted to measure the average level of some chemical in people's blood, for example, we might model the underlying distribution, human variation plus measurement error, to be Gaussian. We do not know the variance among people or perhaps the error in our chemical testing equipment. We estimate the mean with the arithmetic mean, $\bar{x}$, and we can calculate the probability of this estimate being within $\pm\delta x$ as

$$p(\mu - \delta x < \bar{x} < \mu + \delta x) = \int_{-\delta x\sqrt{\frac{n}{S_n^2}}}^{+\delta x\sqrt{\frac{n}{S_n^2}}} dt\ p_t(t|\nu = n-1) \tag{4.3.2}$$

$$= \sqrt{\frac{n}{S_n^2}} \int_{-\delta x}^{+\delta x} dx'\ p_t\left(x'\sqrt{\frac{n}{S_n^2}}\middle|\ \nu = n-1\right) \tag{4.3.3}$$

where $p_t(t|\nu)$ is given in section 3.16. Note that we calculate the probability that $\bar{x}$, a statistic of random data, will be within some range of μ, an unknown parameter. This is an example of frequentist hypothesis testing. We will return to this kind of problem later and examine it in detail.

Problem 19. *You are measuring the energies of photons. Your detector saturates at some maximum energy $E_{\max}$ so that any photons with energies higher than this register as $E = E_{\max}$. If the probability distribution for photons is $f(E)$ what will be the mean of the observed energies? What if you are unable to detect photons above $E = E_{\max}$ at all?*

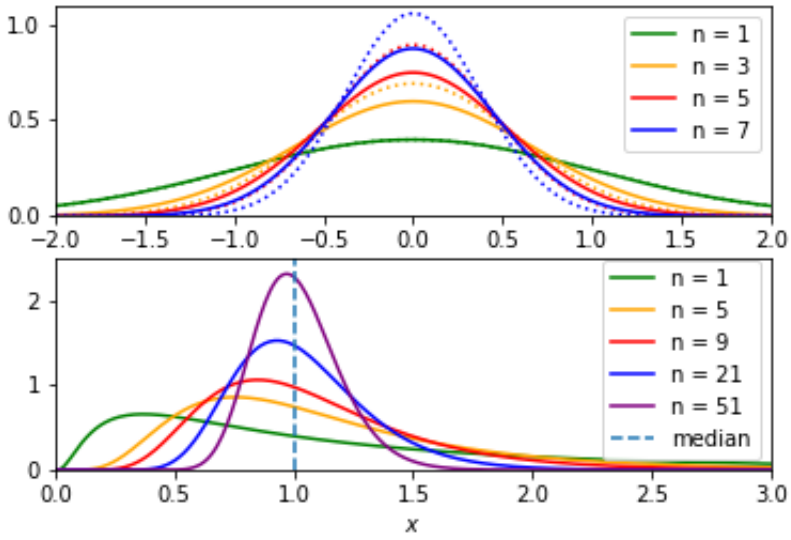

Figure 4.1: The probability of the sample median for normal (above) and lognormal (below) distributions. The $n = 1$ case is the original distribution. The dotted curves in the normal case are the distributions of the sample means based on the same n's.

4.4 median

It is often useful to estimate the median of a distribution. It can be a better representative value of a distribution than the mean when the distribution is highly skewed or there are a few large extreme outliers. A common example of this is the median income of a population. A small number of people with very high incomes can have a large effect on the mean income, but the median is a more **robust** representative value for a typical person in that population. Also, the median can often be more accurately estimated from a small number of observations than the mean. This is particularly true for a distribution with extended tails like a power-law or Lorentzian where the mean might not even be defined. Running median filtering is also a common way to subtract a background in say a spectrum and usually preforms better than a running mean filter.

Consider the median of a sample. Let us assume there are an odd number of observation so the median is well defined. For the median to have value $x_{\rm med}$ one observation must be between $x_{\rm med}$ and $x_{\rm med} + dx$. The probability of this is $p(x)dx$. In addition there must be $(N-1)/2$ observed smaller (and larger) values out of the remaining $N-1$ values. The probability of an observation being below $x_{\rm med}$ is the cumulative probability function $F(x_{\rm med})$. The probability of n independent observations out of

$N-1$ having being $< x_{\rm med}$ is the binomial distribution $P_{\rm binom}(n|N-1, p = F(x_{\rm med}))$. The probability of both of these things happening is the product of their probabilities (product rule for independent events). Any of the N values could be the median so there is a factor of N. The final pdf for the median is

$$p_m(x_{\rm med}|N) = Np(x_{\rm med})P_{\rm binom}\left(\frac{(N-1)}{2}\middle| F(x_{\rm med}), N-1\right) \tag{4.4.1}$$

$$= N\binom{N-1}{\frac{N-1}{2}}p(x_{\rm med})F(x_{\rm med})^{\frac{N-1}{2}}\left[1-F(x_{\rm med})\right]^{\frac{N-1}{2}} \tag{4.4.2}$$

$$= N\binom{2n}{n}p(x_{\rm med})F(x_{\rm med})^{n}\left[1-F(x_{\rm med})\right]^{n} \tag{4.4.3}$$

where $N = 2n+1$.

Let us find the mode of this distribution. The log of the probability is

$$\ln p_m(x) = \ln p(x) + n\ln F(x) + n\ln[1-F(x)] + \mathcal{C}. \tag{4.4.4}$$

Taking the derivative of this and setting it to zero gives

$$\frac{1}{p(\hat{x})}\frac{\partial p(\hat{x})}{\partial x} + n\frac{p(\hat{x})}{F(\hat{x})} - n\frac{p(\hat{x})}{1-F(\hat{x})} = 0 \tag{4.4.5}$$

using the fact the the derivative of the cumulative distribution is the pdf. For large N and thus n we can ignore the first term and

$$F(\hat{x}) = \frac{1}{2}. \tag{4.4.6}$$

as we would expect. For smaller N there will be a bias if $\frac{\partial p(\hat{x})}{\partial x} \neq 0$.

Let's expand the log-pdf for the median, (4.4.4) around the mode $\hat{x}$,. First

$$\ln\left[F(x)\right] \simeq \ln\left[F(\hat{x})\right] + \frac{1}{F(\hat{x})}F'(\hat{x})(x-\hat{x}) + \frac{1}{2}\left(\frac{1}{F(\hat{x})}F''(\hat{x}) - \frac{1}{F(\hat{x})^2}(F'(\hat{x}))^2\right)(x-\hat{x})^2 + \ldots \tag{4.4.7}$$

$$= -\ln[2] + 2p(\hat{x})(x-\hat{x}) + \left(p'(\hat{x}) - 2(p(\hat{x}))^2\right)(x-\hat{x})^2 + \ldots \tag{4.4.8}$$

and

$$\ln\left[1-F(x)\right] \simeq -\ln[2] + 2p(\hat{x})(x-\hat{x}) - \left(p'(\hat{x}) + 2(p(\hat{x}))^2\right)(x-\hat{x})^2 + \ldots \tag{4.4.9}$$

so

$$\ln p_m(x) \simeq \ln p_m(\hat{x}) - 4n(p(\hat{x}))^2(x-\hat{x})^2 + \mathcal{C} \tag{4.4.10}$$

A Gaussian approximation to $p_m(x)$ valid when N is large will then be

$$p_m(x) = \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(x-\hat{x})^2}{2\sigma_m^2}} \tag{4.4.11}$$

where the variance about $\hat{x}$ is given by

$$Var[x_{\text{med}}] = \sigma_m^2 \simeq \frac{1}{8np(x_{\text{med}})^2} \simeq \frac{1}{4Np(x_{\text{med}})^2} \quad N \gg 1 \tag{4.4.12}$$

For $x \sim \mathcal{N}(\mu, \sigma)$ the sample mean has a smaller variance than the sample median by a factor of $\frac{2}{\pi}$. A more careful calculation for small N gives a factor of $\sim \frac{2}{\pi}\frac{(N+2)}{N}$ between them. For a distribution with larger tails then Gaussian and with small sample sizes the median will have a smaller variance than the mean.

4.5 extreme values

The distribution of the sample maximum (or minimum or the n-th largest value) can be found in the same way as the the median

$$\begin{aligned} p_{\max}(x|N) &= Np(x)P_{\text{binom}}\left(N-1|F(x), N-1\right) && (4.5.1)\\ &= Np(x)P_{\text{binom}}\left(0|1-F(x), N-1\right) && (4.5.2)\\ &= Np(x)F(x)^{N-1} && (4.5.3) \end{aligned}$$

Problem 20. *Say there are n dust particles in a spherical balloon. What is the distribution of the distance between the skin of the balloon and the nearest dust particle?*

4.6 quantile estimation

The q-**quantiles** of a distribution are the set of values that divide the full range into q regions of equal probability. They are the generalization of the median which would be the 2-quantile. The nth q-quantile is at the point where $F(x) = n/q$. There are several slightly different ways to estimate this from a sample, but they all agree for large N and generally follow this approach. **Rank** the data (order them by value from least to greatest) and then take the data point whose rank is closest to $r = nN/q + 1/2$ to be an estimate of the nth q-quantile. This $1/2$ makes the ranks for the median ($q = 2, n = 1$) work out to the sample median we used before. There are other choices

which have over properties (see wikapedia). If r is an integer then we can work out pdf in the same way as before.

$$p(x_n|N) = Np(x_n)P_{\text{binom}}\left(r-1|F(x_n), N-1\right) \tag{4.6.1}$$

$$p(x_n|N) = Np(x_n)P_{\text{binom}}\left(\frac{nN}{q} - \frac{1}{2}\middle| F(x_n), N-1\right) \tag{4.6.2}$$

As we will see, when doing Monte Carlo calculations you might only have access to a sample taken from a distribution that you cannot write down analytically. It is often useful to estimate the quantile range the distribution or estimate a range that contains some fixed probability, say 68% or 95%. One might use (4.6.1) with an estimate of the true pdf to judge how well the range can be estimated.

Chapter 5

The Bayesian method

The Bayesian approach to inference gives us a general framework for constraining models for physical processes and for models that describe the probabilistic distribution of the data. It does this by attempting to calculate the probability of a model or specific values for model parameters given the data and any prior knowledge. The Bayesian interpretation of probability allows us to assign a probability to the possibility of a model being the true one *relative to the other models considered.* In contrast, the frequentist approach, that we will look at later, prohibits assigning probability to the models; only data is probabilistic.

5.1 Posterior, likelihood, prior and evidence

All Bayesian analyses begin with Bayes's theorem. We saw this theorem in section 1.5 as a basic property of conditional probabilities. Let me point out that the theorem itself is a mathematical relation and thus it is valid no matter what your interpretation of probability is or what your approach to statistical inference is. The difference between frequentist and Bayesian statistics fundamentally lies in to what probabilities are assigned.

Let $\boldsymbol{D}$ be some amount of data. Let M_i be a model that attempts to explain this data. It is a member of a set of models $\{M_1, M_2 \dots\}$. These models might be totally different with different parameters (say General Relativity, Newtonian Gravity and MOND) or they might differ by only the values of a model's parameters (the planet has unknown mass m). Let's let I represent everything else in the Universe that we will take to be fixed or irrelevant to our experiment (existence of the apparatus, the day of the week, the phase of the moon on a distant planet). We apply Bayes's

theorem to this situation

$$P(M_i|\boldsymbol{D}, I) = \frac{P(\boldsymbol{D}|M_i, I)P(M_i|I)}{P(\boldsymbol{D}|I)} \tag{5.1.1}$$

$$= \frac{P(\boldsymbol{D}|M_i, I)P(M_i|I)}{\sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)} \tag{5.1.2}$$

The second line follows from $P(\boldsymbol{D}|I) = \sum_i P(\boldsymbol{D}, M_i|I) = \sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)$ which is the probability that the data will occur assuming the correct model is one of the M_s's. I include I here only to emphasis that every probability has some implicit assumptions. Some of these assumptions could be incorporated into the model, but if they have no effect on the outcome of the experiment or they where never changed when the experiments where conducted they can be considered conditionals for all the probabilities. In the future the I will be considered implicit and not included.

In this context, each of the factors in Bayes's theorem have special names:

- $P(M_i|\boldsymbol{D})$ is called the **posterior probability** for model M_i given the data. This is the goal of Bayesian inference although one often summarizes this result by finding the average, mode, covariance or credibility regions.

- $P(\boldsymbol{D}|M_i)$ is called the **likelihood**. It is the probability of getting the observed data given the model M_i. It is often denoted $\mathcal{L}(\boldsymbol{D}|M_i)$. This is the same probability as is used in frequentist methods. Often this is a Gaussian, but not always. It includes the model that relates the parameters to the data and the description of the noise.

- $P(M_i)$ is called the **prior**. It is the probability of the model prior to the data $\boldsymbol{D}$ being considered. This might take into account some previous experiment with data $\boldsymbol{D}'$ in which case it would be the posterior of that experiment $P(M_i|\boldsymbol{D}')$. It might also take into account that some models, or range of parameters, are not possible in which case $P(M_i) = 0$ for some i. For example, the mass of a planet cannot be negative or Ω_{matter} cannot be greater than one. The prior is often denoted by $\pi(M_i)$ in the literature.

- $P(\boldsymbol{D}) = \sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)$ is called the **evidence**. Note that the evidence is not a function of M_i although it is implicitly dependent on the set of all models considered. Since the data does not change, the evidence will be a constant for a fixed set of models. We will sometimes denote the evidence as $\mathcal{E}(\boldsymbol{D})$.

5.2 Updating the Information

Strict interpretation would hold that $P(M_i|\boldsymbol{D}, I)$ is the prior conditional probability. It requires an additional step to interpret it as a "new" , or posterior, probability for the model. It could be written $P_{\text{new}}(M_i)$. This step is called **Bayes' rule** although it was first stated by Laplace. This process can be viewed as updating our knowledge of the model after we take into account new data or information.

This process can be thought of as a kind of chain where every bit of new information, data, updates our knowledge progressively. Imagine there are two experiments that constrain the same model. The data sets are $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. The posteriors for the two experiments are

$$p(M|\boldsymbol{D}_1) = \frac{p(M)p(\boldsymbol{D}_1|M)}{p(\boldsymbol{D}_1)} \quad p(M|\boldsymbol{D}_2) = \frac{p(M)p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_2)} \tag{5.2.1}$$

Now let's look at the posterior for both data sets,

$$p(M|\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{p(M)p(\boldsymbol{D}_1, \boldsymbol{D}_2|M)}{p(\boldsymbol{D}_1, \boldsymbol{D}_2)} \tag{5.2.2}$$

$$= \frac{p(M)p(\boldsymbol{D}_1|M)p(\boldsymbol{D}_2|\boldsymbol{D}_1, M)}{p(\boldsymbol{D}_1, \boldsymbol{D}_2)} \qquad \text{product rule} \tag{5.2.3}$$

Now if the data sets are statistically independent for the two experiments (experiment two was not influenced by the results of experiment one) then $p(\boldsymbol{D}_1, \boldsymbol{D}_2) = p(\boldsymbol{D}_1)p(\boldsymbol{D}_2)$ and $p(\boldsymbol{D}_2|\boldsymbol{D}_1, M) = p(\boldsymbol{D}_2|M)$ so

$$p(M|\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{p(M)p(\boldsymbol{D}_1|M)p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_1)p(\boldsymbol{D}_2)} \tag{5.2.4}$$

$$= \left[\frac{p(M)p(\boldsymbol{D}_1|M)}{p(\boldsymbol{D}_1)}\right] \frac{p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_2)} \tag{5.2.5}$$

$$= p(M|\boldsymbol{D}_1)\frac{p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_2)} \qquad \text{using 5.2.1} \tag{5.2.6}$$

So the posterior of experiment 1 can be used as a prior for experiment 2. Or it can be the other way around. The order in which the experiments where done should not matter.

Note that although experiments are usually taken to be independent they often are not. Some experiments are done because a previous experiment showed promising results or some experiments are extended in duration based on early results. This can give rise to a form of **confirmation bias**.

5.3 Parameter estimation

The most common use for Bayesian inference is parameter estimation or inference. In this case we have a model that describes the data that is a function of parameters $\theta_1, \theta_2, \ldots$. The different models discussed above are actually the same model with different values. We will assume that these parameters take on a continuous range of values, although this is not necessary. The sum in the evidence then becomes an integral and the posterior is

$$P(\theta_1, \theta_2, \ldots | \boldsymbol{D}) = \frac{\mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots)p(\theta_1, \theta_2, \ldots)}{\left[\int d^n\theta \ \mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots)p(\theta_1, \theta_2, \ldots)\right]} \tag{5.3.1}$$

The posterior expresses the probability of a set of parameter values being correct *given that the model is the correct one.*

There is an objection you might have to this. In the continuous case, the probability of the data having some specific value is technically zero; $\mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots)$ is a density, $\mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots)d\boldsymbol{D}$ is a probability. If we keep the $d\boldsymbol{D}$'s you will see that they cancel out of the posterior. This might not be a satisfying justification from a formal prospective, but I will accept it here and not include the $d\boldsymbol{D}$'s.

example: Poisson radiation

Let's say you have a sample of water from a swamp next to a nuclear power plant. We want to know the level of radioactive contamination in this water. Let there be $N(t)$ unstable nuclei in our sample. The rate of decay is $\frac{dN}{dt} = \lambda N(t)$ where λ is the decay constant. Let's say we know what element we are dealing with and previous studies have measured the decay constant to a high enough accuracy that we can consider it a known constant. The average rate of decay products going into a Geiger counter is then $r = \Omega\lambda N(t)$ assuming one product per decay. Ω is the solid angle covered by the Geiger counter from the prospective of the sample which we will also assume is well enough measured that it can be considered known. So if we can measure r we can easily find $N(t)$. We will measure the number of counts in the Geiger counter over a period of time that is small compared to $1/\lambda$ so that we can consider the change in $N(t)$ to be much smaller than $N(t)$ (Uranium 235 has a decay constant of 3.12×10^{-17} s^{-1} or $1/\lambda = 1.02$ Gyr so this isn't bad approximation in many cases).

Since each nucleus has a constant probability of decay the number of counts, n, will be Poisson distributed (see section 3.6).

$$p(n|r) = \frac{(r\delta t)^n}{n!} e^{-r\delta t} \tag{5.3.2}$$

where δt is the time over which the measurement is done. In this case n is the data and r is the parameter we would like to measure. This Poisson distribution is the

likelihood. We take the prior on the rate to be uniform between 0 and some large number r_{max}. We will see that the result will not depend on the value of r_{max} as long as it is much larger than the actual rate,

$$p(r) = \frac{\Theta(0 < r < r_{max})}{r_{max}} \tag{5.3.3}$$

We know that $p(n|r)$ is normalized to one for its sum over n from 0 to ∞, but to normalize the posterior by calculating the evidence we need to integrate $p(n|r)p(r)$ over r.

$$\mathcal{E}(n) = \int_{-\infty}^{\infty} dr\ p(n|r)p(r) = \frac{1}{r_{max}} \int_0^{r_{max}} dr\ \frac{(r\delta t)^n}{n!}\, e^{-r\delta t} \tag{5.3.4}$$

$$= \frac{\delta t^{-1}}{n! r_{max}} \int_0^{\delta t r_{max}} dx\ x^n\, e^{-x} \qquad x = r\delta t \tag{5.3.5}$$

$$\simeq \frac{\delta t^{-1}}{n! r_{max}} \int_0^{\infty} dx\ x^n\, e^{-x} \qquad r_{max} \gg 1/\delta t \tag{5.3.6}$$

$$= \frac{\delta t^{-1}}{n! r_{max}} \Gamma\left(n+1\right) \tag{5.3.7}$$

$$= \frac{1}{\delta t r_{max}} \qquad \text{because } \Gamma\left(n+1\right) = n! \tag{5.3.8}$$

So combining (5.3.2), (5.3.3), and (5.3.8) the posterior for the rate is

$$p(r|n) = \frac{\delta t}{n!}(\delta t r)^n e^{-r\delta t} \tag{5.3.9}$$

The normalization of the prior, r_{max}, drops out. This posterior is shown in figure 5.1 for some choices of δt and n.

The average of this distribution is

$$\langle r \rangle = \int_o^{\infty} dr\ r p(r|n) = \frac{\delta t}{n!} \int_o^{\infty} dr\ r(\delta t r)^n e^{-r\delta t} = \frac{1}{\delta t n!} \int_o^{\infty} dx\ x^{n+1} e^{-x} \tag{5.3.10}$$

$$= \frac{(n+1)!}{\delta t n!} = \frac{(n+1)}{\delta t} \tag{5.3.11}$$

and the variance is

$$Var\left[r\right] = \frac{(n+1)}{\delta t^2} \tag{5.3.12}$$

One might have expected that the rate should be $\sim n/\delta t$ and that the standard deviation should go like $\propto \sqrt{n}$. Why these extra 1s? We will see later that this small difference in expectation value for small n is related to our choice of prior.

There is nothing particularly special about the average of the posterior. The mode of the distribution can be found by finding the maximum of the log-posterior

$$\frac{\partial}{\partial r}\ln p(r|n) = \frac{\partial}{\partial r}\left([\ln(r\delta t) - r\delta t - \ln(\delta t/n!)\right] \tag{5.3.13}$$

$$= \frac{n}{r} - \delta t \tag{5.3.14}$$

so the most likely value is what we might have expected,

$$r_{mode} = \frac{n}{\delta t}. \tag{5.3.15}$$

This could be called the **maximum posterior estimate** (MPE or sometime MAP) for r which in the case of a uniform prior is also the **maximum likelihood estimator** (MLE).

example: estimating mean

Let's say we have a very simple model for the alcohol content of wine coming out of a winery. The model is that it is constant. We will call the concentration θ. We know that our measurement apparatus has a Gaussian distributed error of σ when measuring the concentration. Say we measure one bottle and get d for the concentration. This kind of model is often written

$$d_i = \theta + n_i, \tag{5.3.16}$$

the data is some fixed value plus a noise component. The likelihood will be

$$\mathcal{L}(d|\theta) = \mathcal{G}\left(d\,|\theta,\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(d-\theta)^2}{2\sigma^2}}. \tag{5.3.17}$$

Now we need a prior for θ. It is common to use a uniform prior in this kind of problem. The argument for this being that without any measurements no particular concentration should be considered more probable than any other. So the prior will be

$$p(\theta) = \begin{cases} \frac{1}{\theta_{\max}-\theta_{\min}} & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.3.18}$$

$$= \mathcal{C}\Theta(\theta_{\min} < \theta < \theta_{\max}) \qquad \mathcal{C} \equiv \frac{1}{\theta_{\max} - \theta_{\min}} \tag{5.3.19}$$

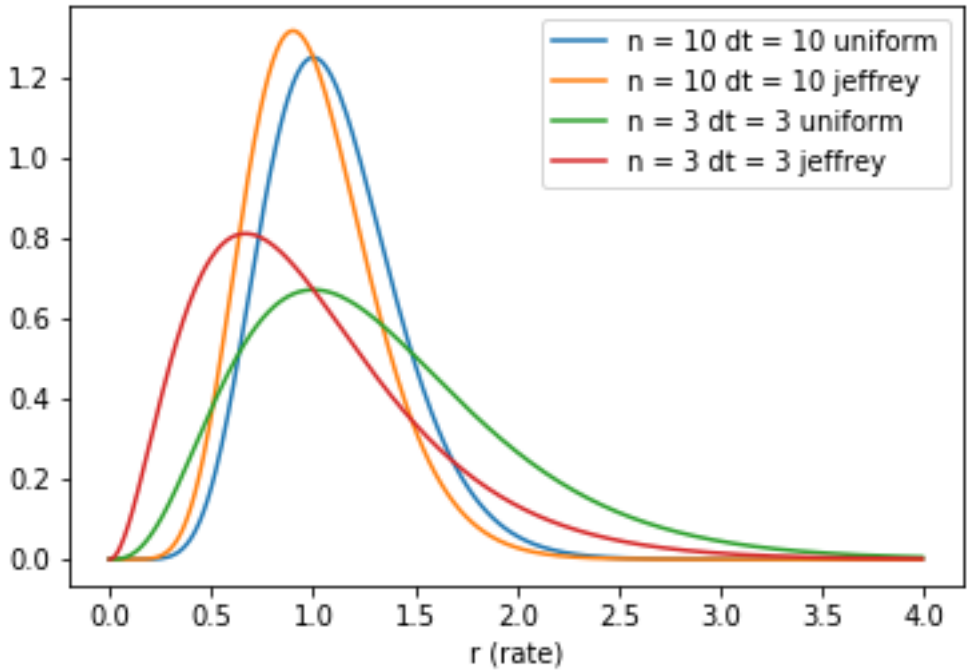

Figure 5.1: Posteriors for the rate r given n counts in δt time units using uniform and Jeffrey priors on the rate.

You might be concerned that the parameters $\theta_{\min}$ and $\theta_{\max}$ might effect the posterior, but we don't know their values. Note that, like in the previous Poisson example, if the likelihood constrains θ to a region that is much smaller than the range allowed by $p(\theta)$ then it will not make any difference. Note also that the normalization of both the likelihood and the prior appear in both the numerator and denominator of the posterior so they drop out. If we take the range of the prior to be much larger than σ, the uniform prior will drop out and not appear.

So in that case the posterior is equal to the likelihood, $\mathcal{G}\left(d\,|\theta,\sigma\right)$ which obviously has a mode at $\theta = d$ and the average is $\langle\theta\rangle = d$.

Now let's consider a slightly more complicated case. We measure N bottles of wine coming out of the factory getting $d_1, d_2 \ldots d_n$ measurements, all with the same σ. Since these are statistically independent measurements the likelihood will be

$$\mathcal{L}(\boldsymbol{d}|\theta) = \mathcal{G}\left(d_1\,|\theta,\sigma\right) \times \mathcal{G}\left(d_2\,|\theta,\sigma\right) \times \ldots \tag{5.3.20}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2}\sum_i \frac{(d_i-\theta)^2}{\sigma^2}\right) \tag{5.3.21}$$

which will also be the the posterior for a uniform prior. Making some changes,

$$\mathcal{L}(\boldsymbol{d}|\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_i \left(d_i^2 - 2d_i\theta + \theta^2\right)\right) \tag{5.3.22}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_i d_i^2 - 2\sum_i d_i\theta + n\theta^2\right]\right) \tag{5.3.23}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left[n\overline{d^2} + n(\theta-\bar{d})^2 - n(\bar{d})^2\right]\right) \tag{5.3.24}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{n}{2\sigma^2}\left[\overline{d^2} - (\bar{d})^2\right]\right) \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{d})^2\right) \tag{5.3.25}$$

where

$$\overline{d} \equiv \frac{1}{n}\sum_i d_i \qquad \overline{d^2} \equiv \frac{1}{n}\sum_i d_i^2. \tag{5.3.26}$$

To find the evidence we need to integrate this over θ.

$$\mathcal{E}(\boldsymbol{d}) = \frac{1}{(2\pi)^{(n-1)/2}\sigma^{n-1}\sqrt{n}} \exp\left(-\frac{n}{2\sigma^2}\left[\overline{d^2} - (\bar{d})^2\right]\right) \tag{5.3.27}$$

All the constant factors will drop out of the posterior. The only part that is dependent on θ is proportional to a Gaussian. Since we already know the normalization

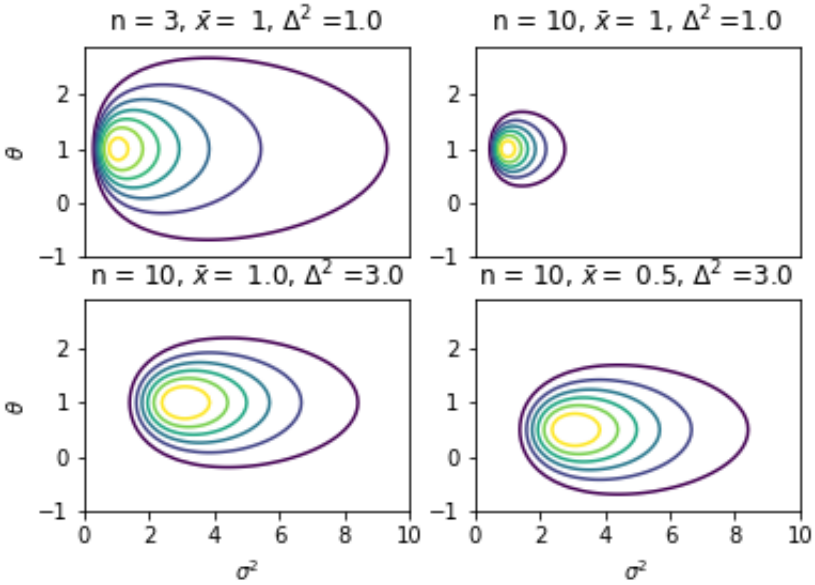

Figure 5.2: The posterior distributions for the mean and variance based on a sample of Gaussian distributed measurements with the number, sample mean and sample variance given above each one.

of a Gaussian we don't even need to do the integration in this case. The posterior is

$$P(\theta|\boldsymbol{d}) = \sqrt{\frac{n}{2\pi\sigma^2}}\exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{d})^2\right) = \mathcal{G}\left(\theta\,|\bar{d},\sigma^2/n\right). \tag{5.3.28}$$

In section 4.1 we found that the sample mean of Gaussian random variables is Gaussian distributed with a variance of σ^2/n. We see here that this is also true for the posterior distribution of the estimated mean. The mean is $\langle\theta\rangle = \bar{d}$. No surprise here.

example: estimating mean and variance

Let's make it a little more complicated. It does not seem reasonable that the alcohol content is exactly constant in every bottle of wine so we should allow for it to change randomly with an unknown variance. We still have a normally distributed error in the measurements with standard deviation σ_n. In addition we will assume the distribution of the alcohol content among bottles is normally distributed with a mean of θ and a variance of σ_a. We would like to know the variance so that in the future we can adjust the process to reduce the variance so that the product is more uniform. Some customers have been complaining.

Each data point is some constant plus (or minus) some random value plus random noise:

$$d_i = \theta + x_i + n_i \tag{5.3.29}$$

We can think of the likelihood as the probability that the actual alcohol content is $\theta + x$ and then the probability of the alcohol level $\theta + x$ being measured as d. We are not interested in the alcohol content of individual bottles so we sum, or integrate, over all possible values of x_i's to eliminate them from the likelihood

$$\mathcal{L}(\boldsymbol{d}|\theta, \sigma_n^2, \sigma_a^2) = \int_{-\infty}^{\infty} d^n x \; P(\boldsymbol{d}, \boldsymbol{x}|\theta, \sigma_a^2) \tag{5.3.30}$$

$$= \int_{-\infty}^{\infty} d^n x \; \left[\mathcal{G}\left(d_1 \left| x_1, \sigma_n^2\right.\right)\mathcal{G}\left(d_2 \left| x_2, \sigma_n^2\right.\right)\dots\right]\left[\mathcal{G}\left(x_1 \left| \theta, \sigma_a^2\right.\right)\mathcal{G}\left(x_2 \left| \theta, \sigma_a^2\right.\right)\dots\right] \tag{5.3.31}$$

$$= \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{d} \left| \boldsymbol{x}, \sigma_n^2\right.\right)\mathcal{G}\left(\boldsymbol{x} \left| \theta, \sigma_a^2\right.\right) \tag{5.3.32}$$

$$= \mathcal{G}\left(\boldsymbol{d} \left| \theta, \sigma_n^2 + \sigma_a^2\right.\right) \tag{5.3.33}$$

where we are using the results of section 3.14 to combine Gaussian pdfs. This is of course a consequence of the sum of normally distributed numbers being normally distributed. This is the same likelihood as we got in the first example except σ^2 is replaced with $\sigma_n^2 + \sigma_a^2$,

$$\mathcal{L}(\boldsymbol{d}|\theta, \sigma_n^2, \sigma_a^2) = \frac{1}{\sqrt{(2\pi)^{n/2}(\sigma_n^2 + \sigma_a^2)^n}} \exp\left(-\frac{n\left[\overline{d^2} - (\bar{d})^2\right]}{2(\sigma_n^2 + \sigma_a^2)}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2(\sigma_n^2 + \sigma_a^2)}\right) \tag{5.3.34}$$

To make things simpler let's make the following substitutions

$$\Delta^2 \equiv \overline{d^2} - (\bar{d})^2 \tag{5.3.35}$$

$$\sigma^2 \equiv \sigma_n^2 + \sigma_a^2 \tag{5.3.36}$$

You can see that σ_n and σ_a enter into the likelihood only in the combination $\sigma_n^2 + \sigma_a^2$. As a result you cannot constrain them separately unless the priors differentiates between them. This is possible. For example some previous calibration tests could put constraints on σ_n.

We will take the case where there are no previous constraints on either of the σ's. We can then use σ^2 as a parameter instead of σ_a^2. The likelihood is now

$$\mathcal{L}(\boldsymbol{d}|\theta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right) \tag{5.3.37}$$

We will assume a uniform prior for both θ and σ^2 (we will talk later about using a Jeffreys prior for σ^2). Further more the variance cannot be less than zero

$$P(\theta, \sigma^2) = \frac{\Theta(\theta_{\max} < \theta < \theta_{\min})}{(\theta_{\max} - \theta_{\min})} \frac{\Theta(0 < \sigma^2 < \sigma^2_{\max})}{\sigma^2_{\max}} \tag{5.3.38}$$

$$= \mathcal{C}\Theta(\theta_{\max} < \theta < \theta_{\min})\Theta(0 < \sigma^2 < \sigma^2_{\max}) \tag{5.3.39}$$

where $\mathcal{C}$ is going to represent the normalization constant.

Now we need to find the evidence by integrating the likelihood over the parameters.

$$\mathcal{E}(\boldsymbol{d}) = \mathcal{C} \int_0^{\sigma^2_{\max}} d\sigma^2 \int_{\theta_{\min}}^{\theta_{\max}} d\theta \; \mathcal{L}(\boldsymbol{d}|\theta, \sigma^2) \tag{5.3.40}$$

$$\simeq \mathcal{C} \int_0^{\sigma^2_{\max}} d\sigma^2 \int_{-\infty}^{\infty} d\theta \; \mathcal{L}(\boldsymbol{d}|\theta, \sigma^2) \tag{5.3.41}$$

$$= \frac{\mathcal{C}}{(2\pi)^{n/2}} \int_0^{\sigma^2_{\max}} d\sigma^2 \int_{\theta_{\min}}^{\theta_{\max}} d\theta \; \frac{1}{\sigma^n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right) \tag{5.3.42}$$

$$= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}} \int_0^{\sigma^2_{\max}} d\sigma^2 \; \frac{1}{\sigma^{n-1}} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \tag{5.3.43}$$

In doing this we have taken the the range of the θ integral to go to infinity. This is justifiable if $|\theta_{\max}| = |\theta_{\min}| \gg \sigma$. We don't know this ahead of time, but it can be justified in retrospect once constraints on σ are found. This can be considered a technical flaw that we will get back to later.

Now let's make the change of variables to

$$y = \sqrt{\frac{n\Delta^2}{2\sigma^2}} \quad \text{so} \quad d\sigma^2 = \frac{n\Delta^2}{y^3} dy \tag{5.3.44}$$

$$\mathcal{E}(\boldsymbol{d}) = \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}} \left(\frac{n\Delta^2}{2}\right)^{\frac{3-n}{2}} \int_{\sqrt{\frac{n\Delta^2}{2\sigma^2_{\max}}}}^{\infty} dy \; y^{n-4} e^{-y^2} \tag{5.3.45}$$

$$\simeq \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}} \left(\frac{n\Delta^2}{2}\right)^{\frac{3-n}{2}} \int_0^{\infty} dy \; y^{n-4} e^{-y^2} \tag{5.3.46}$$

$$= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}} \left(\frac{n\Delta^2}{2}\right)^{\frac{3-n}{2}} \Gamma\left(\frac{n-3}{2}\right) \tag{5.3.47}$$

Here we assumed that $\sigma^2_{\max} \gg n\Delta^2$ in the integration limits.

Now we can construct the posterior. The constant $\mathcal{C}$ in the prior and the evidence will cancel. We can then take the limits to go to infinity or at least so large that there

is no need to put the $\Theta()$ parts of the prior in the posterior because the likelihood will constrain the parameters to be much less than this value. The posterior is

$$P(\theta,\sigma^2|\boldsymbol{d}) = \frac{1}{\sqrt{2\pi}\Gamma\left(\frac{n-3}{2}\right)}\left(\frac{\Delta^2}{2}\right)^{\frac{n-3}{2}}\left(\frac{n}{\sigma^2}\right)^{\frac{n}{2}}\frac{1}{n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.3.48}$$

This posterior is plotted in figure 5.2 for some values of n, $\bar{d}$ and Δ^2.

The mod of the posterior can be found by setting its derivatives with respect to the parameters to zero. It is often more convenient to take the log of the posterior first. Since the log is a monitonic function its maximum will be at the same place.

$$\ln P(\theta,\sigma^2|\boldsymbol{d}) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2\sigma^2}\left[\Delta^2 + (\theta-\bar{d})^2\right] + \text{constant terms} \tag{5.3.49}$$

$$\frac{\partial}{\partial\theta}\ln P(\theta,\sigma^2|\boldsymbol{d}) = -\frac{n}{\sigma^2}(\theta-\bar{d}) \tag{5.3.50}$$

$$\frac{\partial}{\partial\sigma^2}\ln P(\theta,\sigma^2|\boldsymbol{d}) = \frac{n}{2\sigma^2}\left(-1+\frac{\Delta^2}{\sigma^2}+\frac{(\theta-\bar{d})^2}{\sigma^2}\right) \tag{5.3.51}$$

These are simultaneously zero at $\theta=\bar{d}$, $\sigma^2=\Delta^2=\overline{d^2}-\bar{d}^2$. These are almost, but not quite what we would have gotten with the arithmetic mean and variance we saw before in chapter 4. Specifically the $(N-1)^{-1}$ factor that we saw was needed to make the estimator unbiased has a been replaced with N^{-1}.

I chose to use σ^2 as a parameter, but I could just as well have chosen σ or $\sqrt{\sigma}$ as a parameter instead. The likelihoods would all be the same, but the evidence would be different since it would be an integral over a different variable. Since, by the chain rule,

$$\frac{\partial}{\partial\sigma^2}\ln P(\theta,\sigma^2|\boldsymbol{d}) = \frac{1}{2\sigma}\frac{\partial}{\partial\sigma}\ln P(\theta,\sigma|\boldsymbol{d}) = \frac{1}{4\sigma^3}\frac{\partial}{\partial\sigma^{1/2}}\ln P(\theta,\sigma^{1/2}|\boldsymbol{d}) \tag{5.3.52}$$

they will all be zero at the same spot the maximum of the posterior will give the same value. However the mean parameter values will not be the same, $\langle\sigma^2\rangle \neq \langle\sigma\rangle^2$.

Problem 21. *You have a series of uncorrelated measurements t_i. You have reason to believe that they are exponentially distributed, i.e.*

$$p(t) = \begin{cases} \frac{1}{\tau}\,e^{-\frac{t}{\tau}} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{5.3.53}$$

1. *What is the posterior for τ with a uniform prior?*
2. *What is the maximum posterior τ?*

5.4 Marginalization

The situation often comes up where there are parameters in the physical or statistical model that we are not interested in. For example we may not know what the variance is, but we are only interested in the mean. Or we may want to make a statement about the constraints on one or two parameters that is independent of what value all the other parameters have. In the Bayesian context these parameters that we are not interested in are called **nuisance parameters**. To remove them from the posterior we marginalize over them.

Let's say parameters $\alpha_1, \alpha_2, \ldots$ are the parameters we are interested in and parameters $\beta_1, \beta_2, \ldots$ are the ones we aren't interested.

$$\begin{aligned} P(\alpha_1, \alpha_2, \ldots | \boldsymbol{D}) &= \int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots \; P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots | \boldsymbol{D}) \\ &= \frac{\int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots \; P(\boldsymbol{D} | \alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots) P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots)}{\mathcal{E}(\boldsymbol{D})} \end{aligned} \tag{5.4.1}$$

example: the mean without the variance

As a simple example let's say we have the posterior (5.3.48). We are interested in the parameter θ, but we are not interested in the "noise" parameter σ^2. Let's marginalize over σ^2 so we have the distribution of θ alone.

We can ignore all the factors that don't have θ or σ^2 in them for the moment because they are just a normalization and we can recover the normalization at the end by integrating over θ. Let's make the substitution $A = n\Delta^2 + n(\theta - \bar{d})^2$ in which

case the relevant parts of the posterior are

$$P(\theta|\Delta^2,\bar{d}) = \int_0^\infty d\sigma^2\ P(\theta,\sigma^2|\Delta^2,\bar{d}) \tag{5.4.2}$$

$$\propto \int_0^\infty d\sigma^2\ \frac{e^{-\frac{A}{2\sigma^2}}}{\sigma^n} \tag{5.4.3}$$

$$\propto -2\int_\infty^0 dx\ x^{n-3}e^{-\frac{A}{2}x^2} \qquad x=\frac{1}{\sigma} \tag{5.4.4}$$

$$\propto 2^{\frac{n-3}{2}}A^{-\left(\frac{n-2}{2}\right)}\Gamma\left(\frac{n-2}{2}\right) \qquad \text{integral in Appendix A.4} \tag{5.4.5}$$

$$\propto \left[\Delta^2+\left(\theta-\bar{d}\right)^2\right]^{\frac{2-n}{2}} \tag{5.4.6}$$

$$\propto \left[1+\frac{\left(\theta-\bar{d}\right)^2}{\Delta^2}\right]^{\frac{2-n}{2}} \tag{5.4.7}$$

If we compare this to the t-distribution ((**??**) in section 3.16) we recognize that $x = |\theta-\bar{d}|\sqrt{n-3}/\Delta$ is a t-distribution with $\nu = n-3$ degrees of freedom. We can recover the normalization constant be comparing this to the standard form

$$P(\theta|\Delta^2,\bar{d}) = \frac{\Gamma\left(\frac{n-2}{2}\right)}{\sqrt{(n-3)\pi}\,\Gamma\left(\frac{n-3}{2}\right)}\left[1+\frac{(\theta-\bar{d})^2}{\Delta^2}\right]^{\frac{2-n}{2}} \tag{5.4.8}$$

Because this is symmetric about $\bar{d}$ the mean is $\langle\theta\rangle = \bar{d}$ and so is the mode. Since the variance of a t-distribution is $\frac{\nu}{\nu-2}$ and $\nu = n-3$ in this case

$$\left\langle x^2\right\rangle = (n-3)\frac{\left\langle(\theta-\bar{d})^2\right\rangle}{\Delta^2} = \frac{\nu}{\nu-2} = \frac{n-3}{n-5} \tag{5.4.9}$$

so

$$\left\langle(\theta-\bar{d})^2\right\rangle = \frac{\Delta^2}{(n-5)}. \tag{5.4.10}$$

From this example we learn that if we model a series of observations to be independent and normally distributed with the same mean and variance and we give the mean and variance uniform priors the posterior distribution of the model mean, θ, (not to be confused conceptually with the sample mean $\bar{d}$) will be t-distributed. As was discussed in section 4.3, the distribution of $t = (\bar{x}-\mu)\sqrt{n/S^2}$ is t-distributed with $\nu = n-1$ degrees of freedom. The number of degrees of freedom are different!

5.5 Choice of prior

As its name suggests, the prior expresses the information one had about the parameters before using the current data to constrain them. This information might come from a previous experiment or observation in which case the prior would be the posterior of that experiment. The prior can also express the theoretically allowed range of a parameter. For example if mass or flux is a parameter the prior should be zero for negative values. Usually there is some boundaries one can put on the value of a parameter at least on theoretical grounds - the mass of a planet cannot be greater than a solar mass.

For the Bayesian parameter estimation problem the actual prior bounds on a parameters are often unimportant. This is because the likelihood will be so small at the boundaries of parameter space that they will not effect the integral in the evidence and the posterior will be zero at these points. In other cases posterior might be significant at the theoretically imposed boundaries to parameter space. For example constraints on cosmological parameters from galaxy surveys or type Ia SNe often do not by themselves rule out the possibility that the average density of the Universe is negative ($\Omega_{matter} < 0$).

A **uniform prior** is often used in Bayesian analysis. This is the prior that is constant over a region of parameter space and zero outside of it. It is unnecessary to specify the limits when likelihood is zero at the boundaries because the normalization appears both in the prior and in the evidence so it cancels out of the posterior.

You might be tempted to always use a uniform prior when not using the results of previous experiment as a prior. It has the appearance of being unprejudiced in the sense that it will not favor one parameter value over another without the data supporting it. This appearance is deceptive however. The prior imposes a metric on parameter space - the prior probability for a parameter being in an infinitesimal region is $p(\alpha)d\alpha$. What is a uniform prior for one set of parameters will not be a uniform prior for another set even though they might describe the same model. For example a uniform prior for σ^2 in the above example is equivalent to prior proportional to σ on the parameter σ because $d\sigma^2 = 2\sigma d\sigma$. With this in mind the uniform prior does not seem so nonprejudicial. It picks out one parameterization which might be an arbitrary choice. Another example of this is the choice of whether to use frequency or wavelength (or period) in some problems. There is no natural reason to choose either one.

Another widely used prior is called **Jeffreys prior**. It is the prior

$$p(\alpha) = \frac{1}{\ln(\alpha_{max}/\alpha_{min})} \begin{cases} 1/\alpha & \alpha_{min} < \alpha < \alpha_{max} \\ 0 & \text{otherwise} \end{cases} \tag{5.5.1}$$

This prior gives equal weight to equal logorithmic ranges of α ($d\ln\alpha = d\alpha/\alpha$). If

parameter, α, is replaced with parameter $\beta = b\alpha^\gamma$ for any constants b and γ this prior will not change posterior since

$$\frac{d\beta}{\beta} = d\ln(b\alpha^\gamma) = \gamma d\ln\alpha \propto \frac{d\alpha}{\alpha} \tag{5.5.2}$$

The constant of proportionality γ will cancel out because it appears in the evidence as well. Jeffreys prior is often used for a "*scale*" parameter as apposed to "*location*" parameters. The difference between these types of parameters is not always clear to me, but it is clear that a scale parameter cannot be less than zero. Examples are energy or mass. A location parameter can be shifted by a constant without fundamentally changing the problem. Examples are the position of a planet or velocity of a galaxy. In the case of complete prior ignorance a uniform prior should be used for location parameters.

The normalization and value of Jeffreys prior is infinite if the range is extended to $0 < \alpha < \infty$. Similarly, the normalization of the uniform prior is formally zero for the range $-\infty < \alpha < \infty$. These ranges are routinely used when the posterior (likelihood times prior) has a well defined integral over these ranges. These are examples of **improper prior** distributions that are not valid distributions by themselves, but make sense in a posterior.

Many researchers feel that the arbitrariness inherent in choosing a prior is a serious flaw in the Bayesian approach. This criticism, I think, only makes sense when the prior is not expressing the results of some previous experiment. Frequentist methods do not have a general way of including prior information which is an important advantage to the Bayesian method. In general, if the data strongly constrains the parameters beyond what was previously known the choice of prior should not strongly affect the resulting posterior; the likelihood will do the constraining by itself. We will compare and contrast these methods further later.

example: Jeffreys prior

Going back the alcohol in wine example, we can now recognize σ^2 as a scale parameter and θ as a location parameter. Previously we used a uniform prior for σ^2. Let's see how things change if we use Jeffreys prior for σ^2. The posterior (5.3.48) will change to

$$P(\theta, \sigma^2|\boldsymbol{d}) \propto \left(\frac{1}{\sigma^2}\right)\frac{1}{\sigma^n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.5.3}$$

where the σ^{-2} factor is from the prior. By integrating this we can determine the normalization

$$P(\theta,\sigma^2|\boldsymbol{d}) = \frac{n^{n/2}}{\sqrt{2^n\pi}\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{\Delta^2}{2}\right)^{\frac{n-1}{2}}\frac{1}{\sigma^{n+2}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.5.4}$$

We can then marginalize over σ^2 as before to get the marginalized distribution for θ

$$P(\theta|\boldsymbol{d}) = \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\frac{1}{\Delta}\left[1+\frac{(\theta-\bar{d})^2}{\Delta^2}\right]^{-\frac{n}{2}} \tag{5.5.5}$$

This is again a t-distribution, but now it is of $\nu = n-1$ degrees of freedom. Recall that with the uniform prior we got a t-distribution of $n-3$ degrees of freedom. From section 4.3 we know that the $t = (\bar{x}-\mu)\sqrt{n/s^2}$ is t-distributed with $\nu = n-1$ degrees of freedom. So in a way this prior agrees with the frequentist result although keep in mind that these are really different quantities we are talking about. t is a function of the data and θ is a model parameter so there is no fundamental reason why there distributions should be the same.

Note also that as n gets bigger the difference between $n-1$ and $n-3$ gets less significant and the difference between the posterior distributions for uniform and Jeffreys become insignificant. It is only when the likelihood is a weak constraint on the parameters relative to the prior that the prior will have a strong effect on the posterior.

example: radiation with Jeffreys prior

Going back to the example given in section 5.3 where we found the posterior for a rate of radioactive decay. We might now recognize the rate as a scale parameter and prefer to us Jeffreys prior rather than the uniform prior we used before. The posterior, after renormalizing, will go from (5.3.9) to

$$p(r|n) = \frac{\delta t}{(n-1)!}(\delta t r)^{n-1}e^{-r\delta t} \tag{5.5.6}$$

The mean and variance of this distribution are more in agreement with frequentist expectations

$$\langle r\rangle = \frac{n}{\delta t} \qquad Var\left[r\right] = \frac{n}{\delta t^2} \tag{5.5.7}$$

Again in the limit of large n the posteriors are the same for the two choices of prior. Interestingly the maximum posterior value in this case is not $\frac{n}{\delta t}$ but

$$r_{mode} = \frac{n-1}{\delta t} \tag{5.5.8}$$

Figure 5.1 shows the posteriors for the rate r with some different values for n, δt and for the uniform and Jeffrey priors. You can see how as n increases the choice of prior becomes less important.

5.6 Bayesian Relativity

An important point about Bayesian parameter estimation is that **Bayesian analysis is always relative.** You always compare one model to another or a class of others. A corollary to this is : **You will always get an answer even when the model is completely wrong.** The posterior for a model that fits the data very badly will often look just fine. There will usually be a set parameters that fit the data best, but that does not mean they fit the data well. Although Bayesian model selection, covered next, purports to go some way toward solving this, it is still relative. Frequentist hypothesis testing which we will cover in a later chapter does a much more satisfying job of answering the question of whether the model is really a good fit to the data in a more general sense.

5.7 Calculating the evidence

It is often difficult or impossible to obtain an analytic expression for the evidence, the normalization of the posterior. In practice it is usually calculated numerically by integrating the likelihood times the prior over the parameter space. This is usually a simple task if there are only 1, 2 or 3 parameters. One can simply grid the parameter space or use a standard integration routine.

Note that if one is doing parameter estimation one only needs the posterior and any factors in the prior and likelihood that do not depend on the parameters will cancel out. For this reason it is often not necessary to normalize these probabilities individually, just the product of them. This can save some work, especially when the likelihood or prior are something strange that you don't know the normalization of.

When the dimension of the parameter space is larger, $\gtrsim 3$, numerical integration can be much more difficult. We will return to this problem later when we talk about Monte Carlo techniques for Bayesian analysis.

5.8 Example: luminosity function

Let's consider the problem of measuring the luminosity function from a data set of stars (or galaxies or AGN or supernovae, etc.) luminosities. This is the same problem as finding the spectral energy distribution (SED) when individual photons or particles are measured. This might be the case of astronomical measurements in the x-ray, γ-ray or high energy cosmic rays. It would also be the case for a particle physics experiment where particle energies are detected. I will call them magnitudes, but everything would be the same if they were energies or something else.

First we need to parameterize the luminosity function. Initially we will consider the simple case of a power-law

$$f(m) = \frac{(1+\alpha)}{\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right]}\, m^{\alpha} \tag{5.8.1}$$

where α will be the parameter we want to know. In this case the normalization depends on the parameter α. The luminosity function could be more complicated like a "broken power-law" where there would be a "break" at some luminosity and second power-law below or above this break. For now we will keep it simple.

no noise

Let's say there is no noise in the measurement of each individual luminosity, or that it is so small that it will not be important ("small" will be made more precise in the next section). The luminosity function (or the SED) is interpreted as proportional to the probability of a object having the magnitude m so the likelihood will be

$$\mathcal{L}(\{m\}|\alpha) = \prod_i f(m_i) = \frac{(1+\alpha)^N}{\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right]^N} \prod_i^N m_i^{\alpha} = \frac{(1+\alpha)^N}{\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right]^N} \left(\prod_i^N m_i\right)^{\alpha} \tag{5.8.2}$$

Note that the normalization of the luminosity function has α in it and we want to find α, so we cannot drop the normalization because it will not drop out of the posterior. Let's use a uniform prior on α. To avoid numerical problems it is useful to calculate the log of the posterior

$$\ln P(\alpha|\{m_i\}) = N\ln(1+\alpha) - N\ln\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right] + \alpha\ln\left(\prod_i^N m_i\right) - \ln\mathcal{E}(\{m_i\}) \tag{5.8.3}$$

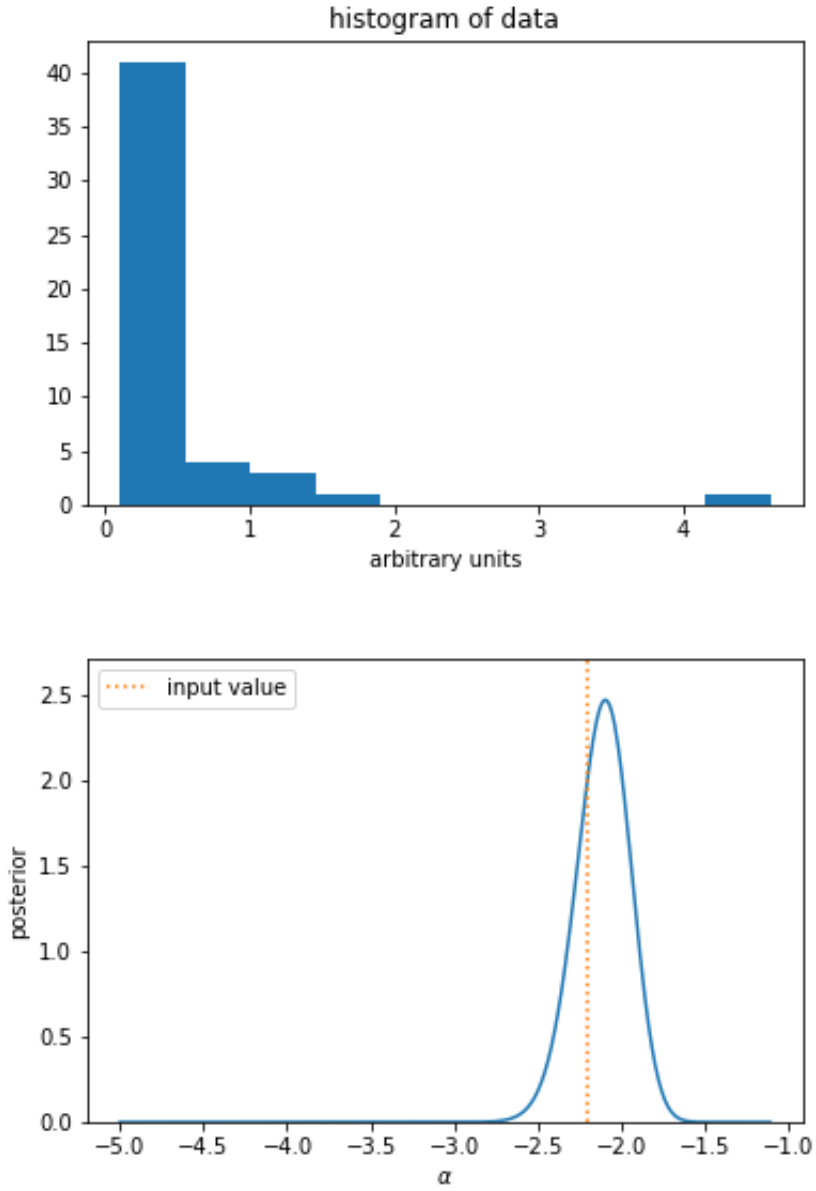

Figure 5.3: Bayesian constraint on the slope of the luminiosity function or SED. A histogram of the luminosities is given above. Below is the posterior for the slope α. There are 50 data points that were generated from a power-law with $\alpha = -2.2$.

It is not necessary to calculate the evidence analytically. We can calculate the first three terms for a range of α then take exp() of it and then normalize it numerically to 1 by adding it up. This is shown in figure 5.3 for a simulated data set.

Note that the range of m, m_{min} to m_{max} does not drop out. m_{max} should be as high as is detectable in the observations. We can take it to be infinite if appropriate. m_{min} is the minimum luminosity that is detectable. There may be objects that are not detected, but we can't see them so they aren't included. The likelihood takes into account not only what objects are measured, but also the regions where no objects are measured. Extending the limits into a region where they cannot be measured would result in an incorrect constraint.

Problem 22. *Assume that α is less than -1 and take m_{max} to be infinite. Find the maximum likelihood estimate for α.*

with noise

So far we have taken the measurements of the luminosities to be perfect. Now we will add some noise in our measurement. Let's look at the measurement of a single star first. The joint probability that a star will have magnitude m and an observed magnitude of m_o is

$$p(m, m_o) = p(m)p(m_o|m) \tag{5.8.4}$$

where $p(m_o|m)$ is the probability of measuring a star of magnitude m to have a magnitude m_o. Let's take this to be a Gaussian error

$$p(m_o|m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m_o-m)^2}{2\sigma^2}}. \tag{5.8.5}$$

Let's just rename $p(m)$ as $f(m)$ to prevent some confusion. $f(m)$ is the intrinsic luminosity function. We can approximate this joint probability by expanding the log

of $f(x)$

$$p(m_o, m) \propto f(m) e^{-\frac{(m_o-m)^2}{2\sigma^2}} \tag{5.8.6}$$

$$= \exp\left[\ln(f(m)) - \frac{(m_o-m)^2}{2\sigma^2}\right] \tag{5.8.7}$$

$$= \exp\left[\ln(f(m_o)) + \frac{\partial \ln f}{\partial m}(m-m_o) + \frac{\partial^2 \ln f}{\partial m^2}(m-m_o)^2 - \frac{(m_o-m)^2}{2\sigma^2} + \ldots\right] \tag{5.8.8}$$

$$= f(m_o) \exp\left[\frac{\partial \ln f}{\partial m}(m-m_o) + \frac{\partial^2 \ln f}{\partial m^2}(m-m_o)^2 - \frac{(m_o-m)^2}{2\sigma^2} + \ldots\right] \tag{5.8.9}$$

$$= f(m_o) \exp\left[\alpha\frac{(m-m_o)}{m_o} - \beta\frac{(m-m_o)^2}{m_o^2} - \frac{(m_o-m)^2}{2\sigma^2} + \ldots\right] \tag{5.8.10}$$

where

$$\alpha(m_o) \equiv \left.\frac{\partial \ln f}{\partial \ln m}\right|_{m=m_o} \qquad \beta(m_o) \equiv \left(\frac{\partial \ln f}{\partial \ln m} - \frac{\partial^2 \ln f}{\partial \ln m^2}\right)_{m=m_o} \tag{5.8.11}$$

Note that if the intrinsic luminosity function is a power law then $\beta = \alpha$.

Now let's find the maximum of the posterior for the true magnitude of the star. This is the most likely value for m given our data m_o. Since $p(m|m_o) = p(m)p(m_o|m)/p(m_o) = p(m_o, m)/p(m_o)$,

$$\frac{\partial}{\partial m} \ln p(m|m_o) = \frac{\partial}{\partial m} \ln p(m, m_o) - \frac{\partial}{\partial m} \ln p(m_o) \tag{5.8.12}$$

$$\simeq \frac{\alpha}{m_o} - \frac{2\beta}{m_o^2}(m-m_o) - \frac{1}{\sigma^2}(m-m_o) \tag{5.8.13}$$

So the maximum posterior is

$$\hat{m} \simeq m_o + \frac{\alpha\sigma^2 m_o}{(m_o^2 + 2\beta\sigma^2)} \tag{5.8.14}$$

So the most probable real magnitude is not the measured value m_o! In astronomy this is called **Eddington bias** although Eddington did not derive it in this Bayesian context and it has probably been derived by many people in many different fields. (Also, some might say bias is not a Bayesian concept.)

Leaving the $\Theta(m_{min} < m_o < m_{max})$ factor out, the observed luminosity function

will be

$$p(m_o) = \int_{-\infty}^{\infty} dm\ p(m, m_o) \tag{5.8.15}$$

$$= \int_{-\infty}^{\infty} dm\ p(m)p(m_o|m) \tag{5.8.16}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dm\ f(m)e^{-\frac{(m_o-m)^2}{2\sigma^2}} \tag{5.8.17}$$

$$= \frac{f(m_o)}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dm\ \exp\left[\alpha\frac{(m-m_o)}{m_o} - \beta\frac{(m-m_o)^2}{m_o^2} - \frac{(m_o-m)^2}{2\sigma^2}\right] \tag{5.8.18}$$

$$= \frac{f(m_o)}{\sqrt{2\pi}\sigma} \exp\left[\frac{\alpha^2\sigma^2}{2(m_o^2+2\beta\sigma^2)}\right] \int_{-\infty}^{\infty} dm\ \exp\left[-\frac{m_o^2+2\beta\sigma^2}{2\sigma^2 m_o^2}\left(m - m_o - \frac{\alpha m_o\sigma^2}{m_o^2+2\beta\sigma^2}\right)^2\right] \tag{5.8.19}$$

$$= \frac{f(m_o)}{\sqrt{1+\frac{2\beta\sigma^2}{m_o^2}}} \exp\left[\frac{\alpha^2\sigma^2}{2(m_o^2+2\beta\sigma^2)}\right] \tag{5.8.20}$$

which is not the true luminosity function. This luminosity function can be used in place of intrinsic luminosity function that was used in the previous section when noise in the measurements is significant. $f(m_o)$ may have some internal parameters besides the power-law slope. You can see that when the noise is very small $\sigma \sim 0$ this becomes the intrinsic luminosity function as it should.

This same treatment can be applied to the energy spectrum of detected particles or photons or to the brightnesses of stars instead of the magnitudes. In these cases there is a lower bound on the intrinsic value of zero. The integrals above will go from 0 to ∞ rather than $-\infty$ to ∞ and the result will have some erf() functions in it, but will be essentially the same.

intrinsic magnitudes

So far in this example we have dealt with apparent magnitudes, or energies etc., and wanted a model for their distribution. Now say we are studying the brightness of galaxies or stars and we have distance information for each object. We want to find the absolute magnitude or luminosity distribution, $\phi(L|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are some parameters. The probability of simultaneously having an intrinsic luminosity, L, an observed brightness l, a true radial distance R and an observed radial distance r can

be broken apart into different factors using the product rule:

$$\begin{aligned} p(L,l,R,r|\boldsymbol{\theta}) &= p(L|\boldsymbol{\theta})p(l,R,r|L,\boldsymbol{\theta}) && (5.8.21)\\ &= p(L|\boldsymbol{\theta})p(R|L,\boldsymbol{\theta})p(l,r|R,L,\boldsymbol{\theta}) && (5.8.22)\\ &= p(L|\boldsymbol{\theta})p(R|L,\boldsymbol{\theta})p(l|R,L,\boldsymbol{\theta})p(r|l,R,L,\boldsymbol{\theta}) && (5.8.23)\\ &= p(L|\boldsymbol{\theta})p(R|L)p(l|R,L)p(r|l,R,L) && (5.8.24) \end{aligned}$$

In the last line we were able to remove the $\boldsymbol{\theta}$ dependence because if L is specified the parameters of the luminosity function are irrelevant so the probabilities cannot depend on them. We can recognize $p(L|\boldsymbol{\theta})$ as being the intrinsic luminosity function $\phi(L|\boldsymbol{\theta})$.

$p(R|L)$ is the probability that the object will be at radius R given no other information than the luminosity. If we assume that a priori the probability of a galaxy is uniform in space then this will be proportional to the partial derivative of the volume with respect to R,

$$p(R|L)dR = \frac{\partial V}{\partial R}\frac{dR}{V} \tag{5.8.25}$$

For flat static space this is $p(R|L) = 3(R/R_{max})^2 R_{max}^{-1}$, but at cosmological distances we would have to differentiate between luminosity distance and angular size distance. Depending on which R is used the volume element might be a different function of it. We will assume here that this is not the case.

The distribution $p(r|l,R,L)$ represents the error in the measurement of the distance r. This might depend on the brightness, l because the measurement might be noisier for low brightness objects. It might also depend on some intrinsic property of the source that is related to L. But we will assume it has none of these dependencies so $p(r|l,R,L) \simeq p(r|R)$.

$p(l|R,L)$ contains the error in the measurement of the brightness and the relationship between R, L and the brightness. If R is the luminosity distance then we might expect

$$p(l|R,L) = p\left(l - \frac{L}{4\pi R^2}\middle|\,\sigma_l^2\right) \tag{5.8.26}$$

where σ_l^2 is a parameter quantifying the noise level, perhaps the variance of Gaussian noise.

The likelihood for one object is

$$\mathcal{L}(\boldsymbol{\theta}|l,r) = p(l,r|\boldsymbol{\theta}) = \int dL \int dR\ p(L,l,R,r|\boldsymbol{\theta}) \tag{5.8.27}$$

$$= \int dL\ \phi(L|\boldsymbol{\theta}) \int dR\ p(R|L)p(l|R,L)p(r|l,R,L) \tag{5.8.28}$$

$$\simeq \int dL\ \phi(L|\boldsymbol{\theta}) \int dR\ p(R)p\left(l - \frac{L}{4\pi R^2}\middle|\sigma_l^2\right) p(r|R) \tag{5.8.29}$$

$$\simeq 3\int \frac{dR}{R_{max}}\ \left(\frac{R}{R_{max}}\right)^2 p(r|R) \int dL\ \phi(L|\boldsymbol{\theta})p\left(l - \frac{L}{4\pi R^2}\middle|\sigma_l^2\right). \tag{5.8.30}$$

The posterior for the data set is then

$$p(\boldsymbol{\theta}|l,r) = \frac{\prod_i \mathcal{L}(\boldsymbol{\theta}|l_i,r_i)\pi(\boldsymbol{\theta})}{\int d\boldsymbol{\theta}\ [\prod_i \mathcal{L}(\boldsymbol{\theta}|l_i,r_i)\pi(\boldsymbol{\theta})]}. \tag{5.8.31}$$

Note that if the noise in the distance measurement were zero, i.e. $p(r|R) \simeq \delta^D(r-R)$ then R^2 part in (5.8.30) would just be a pre-factor that is not dependent on the parameters $\boldsymbol{\theta}$. This pre-factor would also be in the evidence and so would drop out of the posterior so we would be back to essentially the same case as before, noise in l convolved with the luminosity function. When there is noise in the distance measurement this extra factor expresses the fact that you are more likely to underestimate the radial distance than overestimate it even if the distribution of your errors in r are symmetric around R. There are more objects with larger R so more of them "scatter" down into a fixed r bin. This is another form of Eddington bias.

Note that the distribution of errors need not be the same for every object. It might change with type of object, position on the sky or the time at which the measurement was taken.

selection effects

A selection is a cut or diminishment in the probability of observing objects within a range of observable values usually caused by instrumental or observational effects. This can be quantified with a **selection function** which is the probability of observing an event given that it did occur. The selection is a function of observables and not model parameters.

In our example we will denote the selection function by $S(l,r)$. The likelihood will need to be modified because simply multiplying it by $S(l,r)$ would result in a

likelihood that is not normalized properly. The properly normalized likelihood is

$$\mathcal{L}_s(\boldsymbol{\theta}|l,r) = p(l,r|\boldsymbol{\theta},S) = \frac{p(l,r|\boldsymbol{\theta})S(l,r)}{\int dl \int dr\ p(l,r|\boldsymbol{\theta})S(l,r)} \tag{5.8.32}$$

This is now a different function of the parameters.

In the case of a simple cut or **truncation**, the selection function is 1 for observable cases and zero for non-observable cases. In this case the numerator of (5.8.32) will not be modified because there can not be any observed cases that violate the selection. But the denominator will be changed, the limits of integration will be different, so the posterior will be different. This fact is easily missed.

Consider the case where there is no error in the distance measurement and the brightness measurement. We can simplify the pre-selection likelihood (5.8.30)

$$p(l,r|\boldsymbol{\theta}) = \int \frac{dR}{R_{max}}\ 3\left(\frac{R}{R_{max}}\right)^2 \delta^D(r-R) \int dL\ \phi(L|\boldsymbol{\theta})\delta^D\left(l - \frac{L}{4\pi R^2}\right) \tag{5.8.33}$$

$$= \frac{3}{R_{max}}\left(\frac{r}{R_{max}}\right)^2 \int dL\ \phi(L|\boldsymbol{\theta})\delta^D\left(l - \frac{L}{4\pi r^2}\right) \tag{5.8.34}$$

$$= \frac{3}{R_{max}}\left(\frac{r}{R_{max}}\right)^2 \int dL\ \phi(4\pi r^2 l|\boldsymbol{\theta})\delta^D\left(l - \frac{L}{4\pi r^2}\right) \tag{5.8.35}$$

$$= \left(\frac{12\pi}{R_{max}^3}\right) r^4 \phi(4\pi r^2 l|\boldsymbol{\theta}) \tag{5.8.36}$$

In the last set I changed variables to $x = L/(4\pi r^2)$ and then integrated over the delta function. We integrate this with the selection function to find the normalization

$$\int_0^{R_{max}} dr \int_0^\infty dl\ r^4\phi(4\pi r^2 l|\boldsymbol{\theta})S(l) = \int_0^{R_{max}} dr \int_{l_{min}}^\infty dl\ r^4\phi(4\pi r^2 l|\boldsymbol{\theta}) \tag{5.8.37}$$

$$= \int_0^{R_{max}} dr\ r^2 \int_{4\pi r^2 l_{min}}^\infty dL\ \phi(L|\boldsymbol{\theta}) \tag{5.8.38}$$

$$= \int_0^{R_{max}} dr\ r^2 \int_0^\infty dL\ \phi(L|\boldsymbol{\theta})\Theta(L > 4\pi r^2 l_{min}) \tag{5.8.39}$$

$$= \int_0^\infty dL\ \phi(L|\boldsymbol{\theta}) \int_0^{\sqrt{L/(4\pi l_{min})}} dr\ r^2 \tag{5.8.40}$$

$$= \frac{1}{3(4\pi l_{max})^{3/2}} \int_0^\infty dL\ L^{3/2}\phi(L|\boldsymbol{\theta}) \tag{5.8.41}$$

So the likelihood is

$$\mathcal{L}(\{r_i l_i\}|\boldsymbol{\theta}) = \left(3(4\pi l_{max})^{3/2}\right)^n \frac{\prod_i r_i^4 \phi(4\pi r_i^2 l_i|\boldsymbol{\theta})}{\left[\int_0^\infty dL\ L^{3/2}\phi(L|\boldsymbol{\theta})\right]^n} \tag{5.8.42}$$

The constants and r_i^2 factors will cancel out of the posterior giving, with uniform priors,

$$p(\boldsymbol{\theta}|\{r_i, l_i\}) = \mathcal{C}\frac{\prod_i \phi(4\pi r_i^2 l_i|\boldsymbol{\theta})}{\left[\int_0^\infty dL\ L^{3/2}\phi(L|\boldsymbol{\theta})\right]^n} \tag{5.8.43}$$

where the normalization constant is

$$\mathcal{C}^{-1} = \int d\boldsymbol{\theta}\ \frac{\prod_i \phi(4\pi r_i^2 l_i|\boldsymbol{\theta})}{\left[\int_0^\infty dL\ L^{3/2}\phi(L|\boldsymbol{\theta})\right]^n} \tag{5.8.44}$$

The difference in the posterior from what it would be were there no noise and no selection function (the numerator) is a manifestation of **Malmquist bias** - high luminosity objects are sampled from a larger volume when there is a magnitude limited selection. Interestingly the actual brightness limit does not appear in the posterior.

The most commonly used model for the luminosity function of galaxies is the **Schechter function**,

$$\phi(L|\alpha, M_*)dL = \phi_* \left(\frac{L}{L_*}\right)^\alpha e^{L/L_*}\frac{dL}{L_*} \tag{5.8.45}$$

with $\alpha \sim -1.25$. In this case normalizing this function is problematic because it diverges at $L = 0$, but the normalization in (5.8.43), with the brightness limit, is well defined. In actual cases the selection might be quite a bit more complicated than a simple brightness or magnitude cut. The selection might also depend on surface brightness and color for example. Also within the sampled volume the galaxies might have different redshifts so a single observed band will correspond to different intrinsic wavelengths so k-corrections must be taken into effect which is generally dependent on the intrinsic luminosity.

Problem 23. *You have some data from a solar neutrino detector. The detector can detect electron neutrinos, ν_e, muon neutrinos, ν_μ and tau neutrinos, ν_τ, but can not distinguish between ν_μ and ν_τ. You have counts of each over several years - n_{ν_e} and $n_{-\nu_e} = n_{\nu_\mu} + n_{\nu_\tau}$. The efficiency of detection is different for each kind of neutrino because of the different interaction cross-sections and different ways of identifying them. You have calculated that you expect to detect fractions of the fluxes S_e and S_{-e} relative to the true fluxes. What is the posterior for the probability that a neutrino*

coming from the sun will be each of the flavors? Or, in other words, what is the posterior for the relative flux of the neutrino flavors?

Problem 24. *You are given the job of measuring the mass function of galaxy clusters, $f(M|\boldsymbol{\theta})$. There is a selection effect that prevents us from detecting any clusters with masses below $M_{\rm min}$. The mass function is modeled with the function*

$$f(M|\alpha, M_*) \propto \left(\frac{M}{M_*}\right)^{\alpha} e^{-M/M_*} \tag{5.8.46}$$

Make the approximation that the individual masses are very well measured. Find the likelihood function in this case for n measured clusters.

censoring

Related to selection or truncation is the concept of **censoring**. This is the case where only an upper or lower limit for the measured quantity is known. This is a common case in astronomy. For example, one might have a population of galaxies that are selected in the visible, but you are interested in the distribution of radio emission from them. Some galaxies might have only an upper limit on their radio flux. This is different from selection because these objects are known to exist and are included in the data set. Note that this is not the same as missing data, i.e. cases where the galaxy was not observed with a radio telescope.

Censored data should not be ignored and can be incorporated into the likelihood function in a simple way. If the proposed distribution of radio fluxes is $p(f|\boldsymbol{\theta})$ then the probability of the flux being below the threshold f_{max} is the cumulative distribution

$$F(f_{max}|\boldsymbol{\theta}) = \int_{-\infty}^{f_{max}} df\ p(f|\boldsymbol{\theta}) \tag{5.8.47}$$

so this is the factor representing an upper limit that should be used in the likelihood. The limit f_{max} does not need to be the same for each object. It could depend on observation time or weather conditions for example.

In this way detections and non-detections provide information. The likelihood is constrained even when there are no detections at all. The treatment of censored data is often called **survival analysis**. This refers to its use in epidemiology where some individuals survive the trial period and thus there is a lower limit on their lifespan. Similar situations arise in astronomy when a finite observation period might be smaller than the period of a variable star, the orbital period of a planet or the time-delay of a gravitational lens.

Problem 25. *Repeat the calculations of section 5.8 but for an energy spectrum with a lower bound of zero. Find the observed spectrum and the posterior for an intrinsic power-law spectrum.*

Problem 26. *Normally in astronomy the normalization of the luminosity function is given in objects per volume, for example galaxies per* Mpc^3*, or per area on the sky. How would you include in the posterior found in section 5.8 a constraint on the normalization of the luminosity function?*

Chapter 6

Linear models, least-squares and regression

Here we will look at a particular kind of model for the data that is very common, a linear model. By "linear" it is meant that the expression for the measured quantity is linear in the parameters of the model not the data itself. Fitting such a model is sometimes referred to as **linear regression** for historical reasons. This type of model can be applied to a large class of problems and because of its simplicity some quite general solutions can be derived.

A linear model for the data point, d_i, is of the form

$$d_i = \sum_\alpha M_{i\alpha}\theta_\alpha + n_i \qquad \text{or} \qquad \boldsymbol{d} = \boldsymbol{M\theta} + \boldsymbol{n} \tag{6.0.1}$$

where $\boldsymbol{n}$ is the noise, $\boldsymbol{\theta}$ are the parameters of the model and $\boldsymbol{M}$ is a fixed matrix. The simplest case is fitting a line to data, but linear models cover a much broader class of problems. $M_{i\alpha}$ could be a point spread function (psf) and $\boldsymbol{\theta}$ an image to be reconstructed. Or the parameters could be the coefficients of the Fourier modes that describe the data in which case the discrete Fourier transform would be contained in $\boldsymbol{M}$. Related to this, $\boldsymbol{d}$ could be the data from a radio telescope in visibility-space and $\boldsymbol{\theta}$ the image in angular (or configuration) space. It is also true that some nonlinear models can be transformed into linear ones by transforming the data. For example, $d_i = Ax_i^\theta$ implies $\ln(d_i) = B + \ln(x_i)\theta$ so $y_i = \ln(d_i)$ has a linear relationship with θ, although in these cases the noise in $\ln(d_i)$ will not be additive if it was for d_i. Even in these cases some insight into the problem can be gained from linear modeling.

6.1 linear model fitting with a Gaussian likelihood

The simplest case of linear model fitting is fitting a line to data that has one **independent variable** or **predictor variable** and one **dependent variable**. The independent variable predicts the value of the dependent variable and has a small enough error that it can be considered perfectly measured. For example the independent variable might be the time and the dependent variable be sea levels or temperature.

Somewhat counter intuitively, I find this subject easiest to understand if you start from the most general problem and then look at special cases rather than the other way around. Most textbooks start with fitting a line to data with uncorrelated errors. I find that the algebra tends to obscure the meaning and that in practice you are unlikely to ever use the formulas for the simpler cases yourself since curve fitting programs are readily available. For this reason I start with the general case.

A linear model is of the form

$$y = \sum_{\alpha=0}^{M} \theta_\alpha f_\alpha(\boldsymbol{x}) \tag{6.1.1}$$

where y is the dependent variable and x is the independent variable. It is linear in the parameters $\boldsymbol{\theta}$. The simplest case is the average (only θ_0 and $f_0(x) = 1$), the next simplest is a line ($f_0(x) = 1$ and $f_1(x) = x$). Every measured (or selected) value x_i in our data set has a measured value y_i. The prediction of the model can be written

$$y_i = M_{i\alpha}\theta_\alpha \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{M\theta} \tag{6.1.2}$$

where the matrix $\boldsymbol{M}$ contains the values of the functions $f_\alpha(\boldsymbol{x})$ at each point $\boldsymbol{x}_i$

$$\boldsymbol{M} = \begin{pmatrix} f_1(\boldsymbol{x}_1) & f_2(\boldsymbol{x}_1) & \cdots \\ f_1(\boldsymbol{x}_2) & f_2(\boldsymbol{x}_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{6.1.3}$$

We will assume the errors in the data points $\boldsymbol{y}$ are Gaussian. From our discussion of the multivariant Gaussian we know the the log-likelihood is of the form

$$\ln \mathcal{L} = -\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{M\theta})^T \boldsymbol{C}^{-1} (\boldsymbol{y} - \boldsymbol{M\theta}) + \ln\left((2\pi)^N |\boldsymbol{C}|\right)\right] \tag{6.1.4}$$

If we take uniform priors on the parameters then the posterior will be proportional to the likelihood.

Let's first find the maximum of the likelihood (and posterior) with respect to the parameters. The parameter values at this point are called the **Maximum Likelihood Estimator** or MLE of the parameters. I will denote this point as $\hat{\boldsymbol{\theta}}$. It helps to go into Einstein notation for taking the derivative of the likelihood

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_\alpha} = -\frac{1}{2}\frac{\partial}{\partial \theta_\alpha}\left[(y_i - M_{i\beta}\theta_\beta)C^{-1}_{ij}(y_j - M_{j\gamma}\theta_\gamma) + \ln\left((2\pi)^N|\boldsymbol{C}|\right)\right] \tag{6.1.5}$$

$$= \frac{1}{2}\left[M_{i\alpha}C^{-1}_{ij}(y_j - M_{j\gamma}\theta_\gamma) + (y_i - M_{i\beta}\theta_\beta)C^{-1}_{ij}M_{j\alpha}\right] \tag{6.1.6}$$

$$= M_{i\alpha}C^{-1}_{ij}y_j - M_{i\alpha}C^{-1}_{ij}M_{j\gamma}\theta_\gamma \qquad \boldsymbol{C} \text{ is symmetric} \tag{6.1.7}$$

$$= \boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} - \boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\boldsymbol{\theta} \tag{6.1.8}$$

Setting this to zero gives the MLE

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{6.1.9}$$

You might be tempted to say $\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1} = \boldsymbol{M}^{-1}\boldsymbol{C}\left(\boldsymbol{M}^T\right)^{-1}$ and then cancel all the matricies out and get $\hat{\boldsymbol{\theta}} = \boldsymbol{M}^{-1}\boldsymbol{y}$. This generally is not possible however. Usually the number of parameters is small (2 for a line) and the number of data points is much larger. In this case the matrix $\boldsymbol{M}$ clearly cannot be inverted, it has more rows than columns. There are also cases where the number of parameters might be larger than the number of data points. For example an image reconstruction problem might have this property.

A linear problem that has more data points than parameters is considered **overdetermined** in which case $\boldsymbol{M}$ will be taller than it is wide and a unique best-fit solution for the parameters can be found. An **underdetermined** problem has more parameters than (relevant) data points will not have a unique best-fit solution. There will be a range in parameter space that fits the data equally well.

If the number of data points is equal to the number of parameters and $\boldsymbol{M}$ is invertible then the curve will pass through each data point. The model can be used to interpolate between points in this case. A square $\boldsymbol{M}$ might still not be invertible either because there are data points with the same x and different y or because the functions $f_\alpha(x)$ do not allow for enough freedom to reach every point. The result of this would be that the columns (and rows) of $\boldsymbol{M}$ are not linearly independent. Polynomials ($y = \theta_o + \theta_1 x + \theta_2 x^2 + \dots$) will provide enough freedom and are often used for interpolation in this way.

We are not content with just the maximum likelihood estimate of the parameters $\boldsymbol{\theta}$. In this case we can find the complete posterior for them. If we look at (6.1.4) you will see that since the parameters come into the model linearly they come into the $\ln \mathcal{L}$

only up to quadratic order. Any quadratic can be put into the form $(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{A}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})+c$ where c does not contain $\boldsymbol{\theta}$. This can be shown by "completing the squares" as we saw in our section on the multivariant Gaussian (sections 3.14). So **the posterior for the linear model parameters is Gaussian**. We can find the inverse of the covariance, the precision matrix, by taking derivatives of the log-likelihood

$$-\frac{1}{2}A_{\alpha\beta} = \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \tag{6.1.10}$$

This is easily done with equation (6.1.7). The posterior is then

$$p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) = \frac{\sqrt{|\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}|}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right] \tag{6.1.11}$$

and the covariance for the parameters is $\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}$. This is usually not diagonal even when $\boldsymbol{C}$ is diagonal. For this reason the parameters of a linear model will be correlated.

It is not necessary that there be only one independent (predictor) variable per data point. For example the independent variables might be time, temperature and pressure and the dependent variable might be the humidity or the rainfall in the next 24 hours. A linear model is then of the form

$$y = \sum_\alpha \theta_\alpha f_\alpha(x,z,w,\dots) \tag{6.1.12}$$

where $x, z, w, \dots$ are the independent variables. $M_{i\alpha} = f_\alpha(x_i, z_i, w_i, \dots)$. If all the f_α's are linear then this is fitting a hyperplane in parameter space.

It is also not nessary that there be only one dependent variable. If we have two, y and z, that are related linearly to the parameters by

$$y = \sum_\alpha \theta_\alpha f_\alpha(\boldsymbol{x}) \quad \text{and} \quad z = \sum_\alpha \theta_\alpha g_\alpha(\boldsymbol{x}) \tag{6.1.13}$$

we can rearrange this into a matrix form

$$\begin{pmatrix} y_1 \\ z_1 \\ y_2 \\ z_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{x}_1) & f_2(\boldsymbol{x}_1) & \cdots \\ g_1(\boldsymbol{x}_1) & g_2(\boldsymbol{x}_1) & \cdots \\ f_1(\boldsymbol{x}_2) & f_2(\boldsymbol{x}_2) & \cdots \\ g_1(\boldsymbol{x}_2) & g_2(\boldsymbol{x}_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \end{pmatrix} \tag{6.1.14}$$

which is the same form as before ($\boldsymbol{y}' = \boldsymbol{M}'\boldsymbol{\theta}$) so it can be solved in exactly the same way. Some of the f_α's and g_α could be zero so that the parameters could be related to only one dependent variable or not.

6.2 fitting a line

Let's look as the simplest nontrivial and the most common case - fitting a line to data with uncorrelated Gaussian errors in one variable. The model is

$$y = \theta_o + \theta_1 x \tag{6.2.1}$$

Translating this into the matrix form gives

$$\boldsymbol{M} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{pmatrix} \tag{6.2.2}$$

The inverse of the covariance matrix for uncorrelated errors with equal variances is

$$\boldsymbol{C}^{-1} = \frac{\boldsymbol{I}}{\sigma^2} \tag{6.2.3}$$

where $\boldsymbol{I}$ is the identity matrix.

The inverse covariance for the parameters is

$$\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M} = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 1 & 1 & \dots \\ x_1 & x_2 & x_3 & \dots \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{pmatrix} \tag{6.2.4}$$

$$= \frac{1}{\sigma^2}\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \tag{6.2.5}$$

$$= \frac{N}{\sigma^2}\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \tag{6.2.6}$$

We can easily invert this matrix to find the parameter covariance

$$\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1} = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \tag{6.2.7}$$

and the MLE (6.1.9) is

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{6.2.8}$$

$$= \frac{1}{N(\overline{x^2}-\overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 & \dots \\ x_1 & x_2 & x_3 & \dots \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{pmatrix} \tag{6.2.9}$$

$$= \frac{1}{N(\overline{x^2}-\overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} \tag{6.2.10}$$

$$= \frac{1}{(\overline{x^2}-\overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} \tag{6.2.11}$$

$$= \frac{1}{(\overline{x^2}-\overline{x}^2)}\begin{pmatrix} \overline{x^2}\overline{y} - \overline{x}\ \overline{xy} \\ \overline{xy} - \overline{x}\ \overline{y} \end{pmatrix} \tag{6.2.12}$$

And (6.1.11) is the posterior. So, in this case, we just need to calculate the sample averages $\overline{x}$, $\overline{y}$, $\overline{x^2}$ and $\overline{xy}$ to find the best fit line.

Of course in practice this fitting is usually done by a software library. The software will easily handle inhomogeneous noise and correlations between data points.

6.3 fitting a line when both variables are uncertain

It sometimes arises (particularly in astronomy) that the measurement of the independent variable has significant noise in it also. In this case the distinction between dependent and independent variables is not meaningful and we cannot use the solution found in the previous section.

Let us call the observed values for the variable $\boldsymbol{x}^o, \boldsymbol{y}^o$ and the "true" values $\boldsymbol{x}, \boldsymbol{y}$. The variance in their measurements will be σ_y^2 and σ_x^2. Our model requires that

$y_i = \theta_o + \theta_1 x_i$. The likelihood is

$$\mathcal{L}(\boldsymbol{x}^o, \boldsymbol{y}^o|\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{(2\pi\sigma_x\sigma_y)^N} \exp\left[-\frac{1}{2}\sum_i \frac{(y_i^o - \theta_0 - \theta_1 x_i)^2}{\sigma_y^2}\right] \exp\left[-\frac{1}{2}\sum_i \frac{(x_i^o - x_i)^2}{\sigma_x^2}\right] \tag{6.3.1}$$

$$= \prod_i \mathcal{G}\left(y_i^o \left|\theta_0 + \theta_1 x_i, \sigma_y^2\right.\right)\mathcal{G}\left(x_i^o \left|x_i, \sigma_x^2\right.\right) \tag{6.3.2}$$

$$= \prod_i \mathcal{G}\left(\theta_1 x_i \left|y_i^o - \theta_0, \sigma_y^2\right.\right)\mathcal{G}\left(x_i \left|x_i^o, \sigma_x^2\right.\right) \tag{6.3.3}$$

$$= \prod_i \frac{1}{\theta_1}\mathcal{G}\left(x_i \left|\frac{y_i^o - \theta_0}{\theta_1}, \frac{\sigma_y^2}{\theta_1^2}\right.\right)\mathcal{G}\left(x_i \left|x_i^o, \sigma_x^2\right.\right) \tag{6.3.4}$$

We can use the rule for combining multivariant Gaussians that was introduced in section 3.14 to rearrange this

$$\mathcal{L}(\boldsymbol{x}^o, \boldsymbol{y}^o|\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{\theta_1^N}\prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left|x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right)\mathcal{G}\left(x_i \left|\mu_c, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \tag{6.3.5}$$

where the exact value of μ_c will not be important except that it does not contain x_i.

We are interested in the posterior for the parameters θ_0 and θ_1 and not in the actual value of x in each case (the x_i's). So we marginalize over these values which in this case are parameters

$$P(\boldsymbol{\theta}|\boldsymbol{x}^o, \boldsymbol{y}^o) = \int d^n x \; P(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{x}^o, \boldsymbol{y}^o) \tag{6.3.6}$$

$$= \mathcal{C}\int d^n x \; \mathcal{L}(\boldsymbol{x}^o, \boldsymbol{y}^o|\boldsymbol{\theta}, \boldsymbol{x}) \tag{6.3.7}$$

$$= \frac{\mathcal{C}}{\theta_1^N}\prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left|x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1}\right.\right)\int dx_i \mathcal{G}\left(x_i \left|\mu_c, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \tag{6.3.8}$$

$$= \frac{\mathcal{C}}{\theta_1^N}\prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left|x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \qquad \boldsymbol{G} \text{ is normalized} \tag{6.3.9}$$

$$= \frac{\mathcal{C}}{\left(2\pi(\sigma_y^2 + \theta_1^2\sigma_x^2)\right)^{N/2}} \exp\left[-\frac{1}{2}\sum_i \frac{(y_i^o - \theta_0 - \theta_1 x_i^o)^2}{(\sigma_y^2 + \theta_1^2\sigma_x^2)}\right] \tag{6.3.10}$$

Where $\mathcal{C}$ is a normalization constant that needs to be found by integrating over the parameters. This is *not* Gaussian. Note that when $\sigma_y^2 \gg \theta_1^2\sigma_x^2$ the posterior approaches the solution found before and when $\sigma_y^2 \ll \theta_1^2\sigma_x^2$ the line is steeper and the noise in the x variable becomes more important.

Now we can find the MLE for the parameters by finding the maximum of the likelihood

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_0} = \sum_i \frac{(y_i^o - \hat{\theta}_0 - \hat{\theta}_1 x_i^o)}{(\sigma_y^2 + \hat{\theta}_1^2 \sigma_x^2)} = 0 \qquad \Rightarrow \qquad \overline{y} - \hat{\theta}_0 - \hat{\theta}_1 \overline{x} = 0 \tag{6.3.11}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_1} = -\frac{N\hat{\theta}_1 \sigma_x^2}{(\sigma_y^2 + \hat{\theta}_1^2 \sigma_x^2)} + \sum_i \frac{(y_i^o - \hat{\theta}_0 - \hat{\theta}_1 x_i^o) x_i^o}{(\sigma_y^2 + \hat{\theta}_1^2 \sigma_x^2)} = 0 \qquad \Rightarrow \quad \overline{yx} - \hat{\theta}_0 \overline{y} - \hat{\theta}_1 \overline{x^2} - \hat{\theta}_1 \sigma_x^2 = 0 \tag{6.3.12}$$

Solving these equations gives

$$\hat{\theta}_0 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{(\sigma_x^2 + \overline{x^2} - \overline{x}^2)} \tag{6.3.13}$$

$$\hat{\theta}_1 = \frac{\overline{y}\ \overline{x^2} + \overline{y}\ \sigma_x^2 - \overline{xy}\ \overline{x}}{(\sigma_x^2 + \overline{x^2} - \overline{x}^2)} \tag{6.3.14}$$

You can see that if $\sigma_x^2 = 0$ the former solution (6.2.12) is recovered.

orthogonal least-squares

Another way of fitting a line to data points when it is not clear that there are independent variables is to find the line that minimizes the sum of the squares of the minimum distances between the points and the line. This is the length of the line segment perpendicular the the line that passes through the point. This is equivalent to the above solution only when the noise in all variables are equal and constant. If the noise is unknown this is a simple way of fitting a model.

6.4 regression with censored data

Sometimes one is faced with a regression problem where some of the data points are upper limits on the dependent variable. As discussed in section 5.8, these points should be taken into account in the likelihood with the cumulative distribution up to

the upper limit. For the case of the a linear model and Gaussian errors this is

$$F\left(y_{\text{upper}}|x\right) = \int_{-\infty}^{y_{\text{upper}}} dy' \; p\left(y' \middle| \sum_{\alpha} \theta_\alpha f_\alpha(x), \sigma\right) \tag{6.4.1}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y_{\text{upper}}} dy' \; \exp\left[-\frac{1}{2\sigma^2}\left(y' - \sum_{\alpha} \theta_\alpha f_\alpha(x)\right)^2\right] \tag{6.4.2}$$

$$= \frac{1}{2}\left[1 + \text{erf}\left(\frac{y_{\text{upper}} - \sum_\alpha \theta_\alpha f_\alpha(x)}{\sigma}\right)\right] \tag{6.4.3}$$

The full likelihood will be the product of these factors for each upper limit and the regular likelihood (6.1.4) for the measured values. The maximum of this needs to be found numerically.

6.5 least-squares

So far in this chapter we have considered the data to be Gaussian distributed with known covariance matrix, $\boldsymbol{C}$. In that case we can find the posterior and maximum likelihood solution for a linear model. The same techniques are often used even when the covariance is not known. We can seek the solution that simply minimizes the square of the difference between the predicted and measured values for each of the data points. In other words minimize

$$M_{SE} = ||\boldsymbol{y} - \boldsymbol{M\theta}|\,|_2^2 \equiv \sum_i \left(y_i - \sum_\alpha M_{i\alpha}\theta_\alpha\right)^2 \tag{6.5.1}$$

In some contexts this is called the **mean squared error** or MSE. (It is conventional to define $||\boldsymbol{x}||_p \equiv \left(\sum_i x_i^p\right)^{1/p}$.) You can see from our previous discussion that this is the same thing as finding the MLE solution for the case where the data is Gaussian distributed and the covariance is constant and diagonal. The solution follows just as before only without the covariance

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{y} \tag{6.5.2}$$

This is the least-squares solution. Found without assuming anything about the distribution of the data, but that does not mean its the best solution in all cases. The matrix $\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T$ is sometimes called the **pseudoinverse** or **Moore-Penrose inverse** of the matrix $\boldsymbol{M}$ (replacing the transposes with the Hermitian transpose for complex matrices).

The minimum χ^2 problem of section 6.1 can be converted into a least-squares problem by **pre-whitening** the data. Whitening matrix $\boldsymbol{W}$ that has the property $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{C}^{-1}$. The data vector $\boldsymbol{x}$ can then be transformed by

$$\boldsymbol{w} = \boldsymbol{W}\boldsymbol{x}. \tag{6.5.3}$$

The new data vector will have a covariance matrix equal to the identity matrix because

$$\begin{aligned} < \boldsymbol{w}\boldsymbol{w}^T > &= \boldsymbol{W} < \boldsymbol{x}\boldsymbol{x}^T > \boldsymbol{W}^T && (6.5.4)\\ &= \boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^T && (6.5.5)\\ &= \boldsymbol{W}\left(\boldsymbol{W}^T\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T && (6.5.6)\\ &= \boldsymbol{W}\boldsymbol{W}^{-1}\left(\boldsymbol{W}^T\right)^{-1}\boldsymbol{W}^T && (6.5.7)\\ &= \boldsymbol{\mathcal{I}}. && (6.5.8) \end{aligned}$$

This is a generalization of the standardized variables we have seen before. See appendix A.5 for more details on this subject.

Because the components of the $\boldsymbol{w}$ data vector are uncorrelated, its χ^2 reduces to a sum of squares. For this reason, from a computational point of view, maximizing χ^2 and least-squares are essentially same problem.

rank & under/overdeterminedness

A concept of importance here is the **rank**, r, of the $\boldsymbol{M}$ matrix. The rank of a matrix is the number of linearly independent columns (or equivalently rows) of the matrix. Consider a matrix that has m rows by n columns. A matrix with *full rank* has rank equal to the minimum of m and n, i.e. the smallest dimension of the matrix. A square matrix ($m = n$) with full rank is invertible. A square matrix with less than full rank is singular. If $\boldsymbol{M}$ is an m-by-n matrix there are m parameters and n data points.

When $n < m$ the problem is underdetermined, more parameters than data in essence. In this case, the parameters cannot be uniquely determined, but some linear combinations of them might be. Those linear combinations span a subspace that is of dimension equal to the rank, r of the matrix $\boldsymbol{M}$. The remainder of parameter space, of dimension $\max(0, m - r)$, is spanned by linear combinations of the parameters which are not constrained by the data. This subspace is the **right null space** of the matrix $\boldsymbol{M}$. That is $\boldsymbol{M}\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta}$ in this subspace. For every solution that minimizes M_{SE} you can add any vector of parameters in this null subspace and it will not change M_{SE}. Thus the solution is *degenerate*, not unique. The real requirement for being underdetermined is $r < m$ which will always be the case for $n < m$, but also includes cases where the model contains some hidden degrees of freedom that do not change its prediction for the data. For example, if you want to make an image

reconstruction on a grid of pixels that are smaller than the psf or in a region where there is not data you will not be able to find a unique best-fit answer.

In the opposite case, $r < n$, which is always true if $m < n$, there is a unique best-fit set of parameters. The problem is overdetermined. Here there is a degeneracy in the data in that multiple data sets will give exactly the same best-fit parameters. These differ by vectors in the **left null space** of $\boldsymbol{M}$ ($\boldsymbol{x}^T\boldsymbol{M} = 0$). Perhaps the simplest example is where the only parameter is the mean, μ. Any change in the data that keeps the sample mean constant will not change the best-fit solution. There are $n-1$ linearly independent ways of doing that. In general, there are $\max(0, n-r)$ linearly independent vectors that can be added to the data without changing the best-fit parameters.

calculating the pseudoinverse

The pseudoinverse is usually found by single-valued decomposition or SVD (see appendix A.2). The SVD decomposition of $\boldsymbol{M}$ is $\boldsymbol{M} = \boldsymbol{SVD}^T$, where $\boldsymbol{V}$ is a diagonal matrix, but it is not square. This is a generalization of the eigen decomposition. The number of columns of $\boldsymbol{V}$ will be the number of parameters and the number of rows will be the number of data points. The pseudoinverse of $\boldsymbol{M}$ is then

$$\boldsymbol{M}^+ \equiv \left(\boldsymbol{M}^T\boldsymbol{M}\right)^T \boldsymbol{M}^T = \boldsymbol{DV}^+\boldsymbol{S}^T \tag{6.5.9}$$

where $\boldsymbol{V}^+$ is found by taking the reciprocal of the nonzero entries, i.e. $V^+_{ii} = 1/V_{ii}$ for $V_{ii} \neq 0$. The SVD decomposition can be calculated relatively quickly and accurately for a reasonable size matrix. When the the matrices get large this can become prohibitively expensive, computational time $\mathcal{O}\left[\min(m^2n, mn^2)\right]$. There exist various approximate and iterative methods for finding the inverse that are used when the dimensions are high.

Any good linear algebra software package will provide ways of determining the rank of a matrix. Usually the SVD decomposition routine provides this.

6.6 Bayesian Prediction

We have dealt with Bayesian inference. There is another class of statistical problems where one is interested in prediction. We will return to it in later chapters, but here we consider how it can be addressed within the Bayesian framework.

We can find the posterior for the parameters $p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X})$ where $\boldsymbol{Y}, \boldsymbol{X}$ is the training set with known independent and dependent variables. Now consider a new independent value $\boldsymbol{x}$. We can calculate the probability of a new dependent variable by

marginalizing the likelihood over the posterior,

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{Y},\boldsymbol{X}) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{Y},\boldsymbol{X}) \tag{6.6.1}$$

$$= \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{Y},\boldsymbol{X})p(\boldsymbol{\theta}|x,\boldsymbol{Y},\boldsymbol{X}) \tag{6.6.2}$$

$$= \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}) \tag{6.6.3}$$

$$= \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}). \tag{6.6.4}$$

The first step was the product rule. The second step follows from the requirement that the parameters do not depend on the new point x where we want to make a prediction and that the y value at this point does not depend on the data set, it only depends on the model and x. In the last step we recognized $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})$ as the likelihood for the new data point. You can see that this is basically taking the expectation value of the likelihood over the posterior for the parameters given the training data.

The above is a prediction for the observed value of $\boldsymbol{y}$ which includes noise so it could be biased. But given a set of parameters and the independent variable(s) the regression model predicts a unique dependent variable(s) so we can predict the "real" $\boldsymbol{y}$ without noise. In this case we can interpret $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})$ as a delta function that matches one unique $\boldsymbol{y}$ to each $\boldsymbol{x}$ given parameters $\boldsymbol{\theta}$. The prediction becomes

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{Y},\boldsymbol{X}) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \delta^D\left(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta})\right) p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}) \tag{6.6.5}$$

where $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta})$ is the "regression model". This allows you to predict a distribution for $\boldsymbol{y}$ given $\boldsymbol{x}$ (and the assumed regression model) if you can do the integral. This is however not the same thing as the distribution of y's that might be *observed* in the future which would included noise.

with a linear model and Gaussian noise

In the special case of a linear model and one independent variable $\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_i f^i(x)\boldsymbol{\theta}_i = \boldsymbol{f}_x \cdot \boldsymbol{\theta}$. This is the projection of the parameters onto one vector and thus the delta function in (6.6.5) restricts the integral to a hyperplane perpendicular to the vector $\boldsymbol{f}_x$.

We know from section 6.1 that if the errors are Gaussian and the prior is uniform (or Gaussian) the posterior for the parameters will be Gaussian. We also know from section 3.14 that a Gaussian marginalized over a hyperplane is also a Gaussian. It

follows that $p\left(y|x,X,Y\right)$ is a Gaussian distribution. All we need to do is find the mean and variance of it which is easily done

$$\begin{aligned}
\langle y\rangle &= \int_{-\infty}^{\infty} dy\ y\ p\left(y|\boldsymbol{x},\boldsymbol{Y},\boldsymbol{X}\right) && (6.6.6)\\
&= \int_{-\infty}^{\infty} dy\ y \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \delta^D\left(y-\boldsymbol{f}_x\cdot\boldsymbol{\theta}\right))\ p\left(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}\right) && (6.6.7)\\
&= \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \boldsymbol{f}_x\cdot\boldsymbol{\theta}\ p\left(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}\right) && (6.6.8)\\
&= \boldsymbol{f}_x\cdot\hat{\boldsymbol{\theta}} && (6.6.9)
\end{aligned}$$

So the mean (and mode) are what you might expect, the prediction using the best fit parameters. The variance is

$$\begin{aligned}
\left\langle\left(y-\langle y\rangle\right)^2\right\rangle &= \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \left[\boldsymbol{f}_x\cdot\boldsymbol{\theta}-\boldsymbol{f}_x\cdot\hat{\boldsymbol{\theta}}\right]^2 p\left(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}\right) && (6.6.10)\\
&= \sum_{ij} f^i(x)\boldsymbol{C}^{\theta}_{ij}f^j(x) && (6.6.11)\\
&= \boldsymbol{f}_x^T\left(\boldsymbol{f}_x^T\boldsymbol{C}^{-1}\boldsymbol{f}_x\right)^{-1}\boldsymbol{f}_x && (6.6.12)
\end{aligned}$$

where the expression (6.1.11) for the covariance of the parameters has been used in the last step. This variance properly takes into account the uncertainty in the model that is fit to the existing data.

The same approach can be applied to the case where there is a Gaussian prior on the parameters.

6.7 nonparametric regression and smoothing

Another approach to regression is to smooth the data instead of fitting all of it to a polynomial or other functions. In this case one is usually not interested in the parameters of the fit in themselves, just in finding a continuous function $\hat{f}(x)$ that predicts y for any value of x. This allows for more flexibility in the fit because no functional form is assumed, at least not over the full range of the independent variable.

One popular way of doing this is kernel smoothing. One chooses a kernel function K(x). It might be a top-hat function, a Gaussian, B-spline or something else. It is symmetric about $x=0$ and drops off rapidly for $|x|>1$. The estimated function

$\hat{f}(x)$ is then

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i \mathrm{K}\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^{n} \mathrm{K}\left(\frac{x-x_i}{h_x}\right)} \tag{6.7.1}$$

where h_x is a scale factor that needs to be chosen. This is also called the **Nadaraya-Watson estimator**. The error in this estimator at some evenly spaced points can be calculated by bootstrap (section 7.4.1) and then the scale h_x can be increased or decreased until the desired amount of variance in the fit is reached. A large h_x will be stiffer and thus its bias will be larger. A smaller h_x will be more flexible and have larger variance. The h_x can also be found by minimizing the MSE in k-fold cross-validation.

The average of this is an unbiased estimator of the kernel smoothed function

$$\left\langle \hat{f}(x) \right\rangle = \frac{\sum_{i=1}^{n} f(x_i) \mathrm{K}\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^{n} \mathrm{K}\left(\frac{x-x_i}{h_x}\right)}. \tag{6.7.2}$$

and the variance if the noise is uncorrelated between data points is

$$\sigma^2_{\hat{f}(x)} = \left\langle \hat{f}(x)^2 \right\rangle - \left\langle \hat{f}(x) \right\rangle^2 \tag{6.7.3}$$

$$= \frac{\sum_{i=1}^{n} \sigma_i^2 \, \mathrm{K}\left(\frac{x-x_i}{h_x}\right)^2}{\left[\sum_{i=1}^{n} \mathrm{K}\left(\frac{x-x_i}{h_x}\right)\right]^2}. \tag{6.7.4}$$

It is not required that there be only one independent variable. In other words, x could be a vector and $\hat{f}(\boldsymbol{x})$ could be a function of several variables. The kernel could be isotropic in these variables or there could be a different h_x for each dimension.

Kernel smoothing is a particular case of a wider class of regression methods where a curve is fit to a subset of the points that are near x instead of all the points at once. Spline fitting is another example.

Problem 27. *Calculate the covariance of the kernel regression function at two different independent variable values:*

$$\hat{C}_{xz} \equiv \left\langle \hat{f}(x)\hat{f}(z) \right\rangle - \left\langle \hat{f}(x) \right\rangle \left\langle \hat{f}(z) \right\rangle \tag{6.7.5}$$

Chapter 7

Supervised learning & resampling techniques

In this chapter we consider methods that require less knowledge of the data's distribution. Such methods attempt to draw conclusions by estimating the distribution of the data from the data itself. Such methods generally require a large amount of data and there conclusions are less conclusive than could be the case if the distribution were known. On the other hand, they assume less about the data and so have a very broad range of applicability.

7.1 supervised learning & regression

It might be known that the distribution of $\boldsymbol{y}$ is Gaussian but the covariance is not known or it might be that the distributions of $\boldsymbol{y}$ and $\boldsymbol{x}$ are not known at all in which case the least-squares solution is an educated guess. Without knowing the distribution of the y_i's and x_i's we can't say much about the posterior of the θ_α's. But in some problems we might not be too interested in the actual values of the parameters θ_α. The problem of finding these parameter values is called the inference problem. In many cases people are more concerned with prediction than inference, i.e. find a model that will predict the dependent variable given a set of independent variables. This type of problem is especially common in commercial settings and in the social and medical sciences. The independent variables might be age, income, and number of children and the dependent variable the amount of money they pay for a car. There might be no good argument that the errors or intrinsic distributions of the variables are of any particular form. Can we still make progress on this problem? If we have enough data the answer is a limited yes.

When constructing a linear prediction model the question immediately arises as

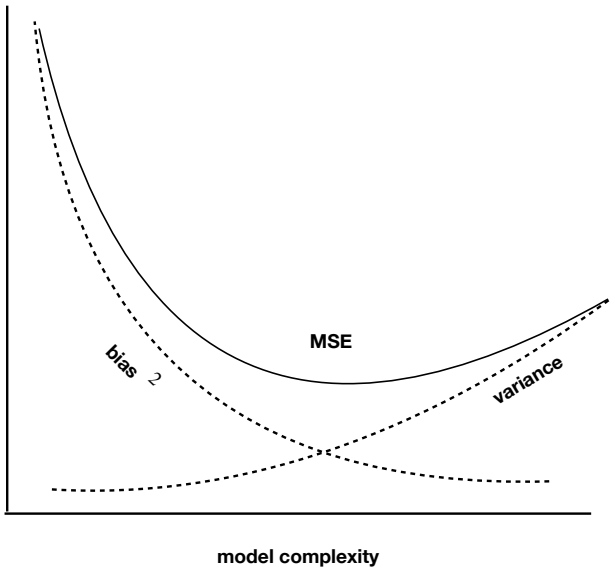

Figure 7.1: General behavior of the MSE and its bias and variance components as a function of the model complexity.

to how many model parameters should be used. Without a well motivated physical model there is no reason to limit the number on theoretical grounds and without knowing the distribution of the variables we cannot use Bayesian model selection or hypothesis testing (covered later) to limit the model space. If we use too many parameters our model will fit the data well in the sense that the mean squared error, M_{SE} (6.5.1) will be small for the data used in the fit, but any data that was not used to fit the model will not be predicted well. We will have over fitted the data. In supervised learning this is known as the **bias–variance trade off**.

The regression model should be viewed as a statistic just like any other. It is a function of a data set and returns a prediction for the dependent variables. Unlike some other statistics (mean, variance. median, MLE, etc.) it is a function of the independent variables so you might think of it as a continuous collection of statistics. Like other statistics we can consider its distribution, bias, variance, tec.

An instructive way to look at the over fitting problem is to decompose the MSE as follows: Let y be the measured dependent value, let $f(x)$ be the true relationship between the independent and dependent variables, n_y is the noise in the measurement of y and the $\hat{f}_{\{x_i\}}(x)$ is the model trained or fit to the data set $\{x_i\}$ that predicts y

given x. The MSE averaged over possible data sets is

$$\left\langle (y - \hat{f}_{\{x_i\}}(x))^2 \right\rangle = \left\langle \left(f(x) + n_y - \hat{f}_{\{x_i\}}(x) \right)^2 \right\rangle \tag{7.1.1}$$

$$= \left\langle f(x)^2 + n_y^2 + \hat{f}_{\{x_i\}}(x)^2 - 2f(x)\hat{f}_{\{x_i\}}(x) \right\rangle \qquad \langle n_y \rangle = 0 \tag{7.1.2}$$

$$= f(x)^2 + \sigma_y^2 + \left\langle \hat{f}_{\{x_i\}}(x)^2 \right\rangle - 2f(x)\left\langle \hat{f}_{\{x_i\}}(x) \right\rangle \tag{7.1.3}$$

$$= f(x)^2 + \sigma_y^2 + Var[\hat{f}_{\{x_i\}}(x)] + \left\langle \hat{f}_{\{x_i\}}(x) \right\rangle^2 - 2f(x)\left\langle \hat{f}_{\{x_i\}}(x) \right\rangle \tag{7.1.4}$$

$$= \sigma_y^2 + Var[\hat{f}_{\{x_i\}}(x)] + \left(f(x) - \left\langle \hat{f}_{\{x_i\}}(x) \right\rangle \right)^2 \tag{7.1.5}$$

$$= \sigma_y^2 + Var[\hat{f}_{\{x_i\}}(x)] + \text{Bias}\left[\hat{f}_{\{x_i\}}(x) \right]^2 \tag{7.1.6}$$

When the model is simple, like a line, the variance in the model prediction $Var[\hat{f}_{\{x_i\}}(x)]$ will be small. (The variance is over all possible training sets.) The bias will be large for a model that is too simple because it might not capture some of the real structure in $f(x)$. As the model becomes more complex the bias will go down, but the variance in the model will go up because spurious features caused by noise will be incorporated in the model and these features will change between data sets. The MSE will then ideally have a minimum somewhere between too simple and too complex. This decomposition and general behavior is valid for linear and nonlinear models. σ_y^2 is an unavoidable lower bound on the MSE because of noise.

cross-validation

A common, practical way to select a model and estimate its predictive error is called k-fold **cross-validation**. The data is split into k subsets. $k-1$ of the subsets are used to fit a model by LSQ or other method. This data set is called the *training* set. The remaining subset that was not used in the fit is called the *validation* set. The mean squared error, MSE, is calculated using this model and the validation set. This is repeated so that the each of the k subsets are used as a validation set once.

If we call the model fit to all but the jth set $\hat{Y}_{-j}(x)$ then the predicted error is

$$\text{MMSE} = \frac{1}{k}\sum_{j=1}^{k} \frac{1}{n_j} \sum_{i \in \{j\}} \left[y_i - \hat{Y}_{-j}(x_i) \right]^2 \tag{7.1.7}$$

where the inner sum is over the subset left out. The number of parameters can be increased until the MMSE reaches a minimum and starts to increasing due to over

fitting. Finally the model with the smallest MMSE is fit to all the data. The error can be estimate from an average of the This gives a better estimate of the expected error given a model and uses the training data more efficiently. A measure of the bias can be found by subtracting the MSE from the whole set from the MMSE.

A special case of cross-validation where the validation set is a single data point and the training set is all the others is called **jackknife** resampling; more on this in section 7.4.2.

This is our first encounter with the subject of **supervised learning** which is a topic in machine learning. The computer "learns" how to predict y from the x's. The independent variables are often call **feature variables** in this context. The machine is "intelligent" in that it can predict y's based on x values that it has never seen before. Thus if we replace the term "fitting" with "training", the linear model becomes the simplest form of artificial intelligence. More complicated nonlinear models like support vector machines (SVM) and artificial neural networks (ANN) perhaps fit this description better. They are trained, or "learn", in much the same way using cross-validation.

It is also possible to train a linear model with something other than the least-squares solution. For example the solution could minimize

$$||\boldsymbol{y} - \boldsymbol{M\theta}|\,|_1 \equiv \sum_i |y_i - \sum_\alpha M_{i\alpha}\theta_\alpha| \tag{7.1.8}$$

which is less sensitive to outliers. In general, the function that is minimized in order to find the best fit model is called the **loss function** (or sometimes the **cost function**).

7.2 R^2

The **coefficient of determination**, R^2, is sometimes used for model selection also. It compares the MSE to the variance of the data,

$$R^2 = 1 - \frac{\sum_i \left[y_i - \hat{Y}(x_i)\right]^2}{\sum_i (y_i - \overline{y})^2} \tag{7.2.1}$$

You can see that if the model is very good R^2 will be close to 1.

The R^2 statistic by itself tends to favor overly complex models. To correct for this the **adjusted** R_a^2 statistic

$$R_a^2 = 1 - \frac{n-1}{n-N_p} R^2 \tag{7.2.2}$$

is advocated where n is the number of data points and N_p is the number of parameters. The motivation for this statistic comes from normally distributed samples. It does not have as clear a justification in other cases although it is sometimes still used.

7.3 adding a prior

We might want to add a prior for our linear model's parameters. There are several reasons why we might do this. One is that we might be trying to reconstruct something with a lot of parameters, like an image, using relatively sparse data. In this case a prior or **regularization** can help make the parameters that are not well constrained behave nicely. It makes the model "stiffer" in the sense that it will not wiggle around fitting every stray data point. In some cases we might have a well justified prior. For example it is well justified to assume the Cosmic Microwave Background (CMB) is a Gaussian field while making a map of it. In other cases we might want to make our reconstruction smooth everywhere the likelihood is not telling us otherwise. For example in trying to deconvolve a blurred image we don't want to add features that are not supported by the data. In many fields the prior is called the **regularization function**.

This is related to the over fitting problem, model selection and feature selection. A prior can be used to force parameters that are not required for the fit to be small (or to be some other chosen value). In this way one does not have to pick which parameters to include. The prior will select which ones are useful for the fit. In this way a problem that are highly underdetermined can be forced to give a unique answer because the combination of the sum of the loss function and the regularization function has a unique minimum. The trade-off is that there are one or more parameters related to the "strength" of the regularization relative to the loss (or log likelihood) that need to be chosen. This is usually done by adjusting the strength to minimize the cross-validation MMSE.

A popular choice for regularization is to use a Gaussian likelihood and a Gaussian, uncorrelated prior on the parameters giving a posterior of the form

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\theta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\theta}) + \ln\left((2\pi)^N|\boldsymbol{C}|\right) - \lambda\,\boldsymbol{\theta}^T\boldsymbol{\theta}\right] \tag{7.3.1}$$

where λ is a free parameter that regulates the strength of the prior. And the maximum posterior solution will be

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{7.3.2}$$

which can be found as before. As before, for the least-squares solution we just remove $\boldsymbol{C}$. In that case you can think of λ as being in units of the variance in the data points. Using this prior while fitting a model is sometimes called **ridge regression**. The parameters that are not well supported by the likelihood will take a penalty for being large and thus will be suppressed. Instead of adding parameters to the model until cross-validation or model selection shows that it is no longer justified you can

instead reduce λ until it is no longer justified. This is particularly useful when the independent (a.k.a. predictor, a.k.a. feature) parameters are not ordered in some way so that it is not clear which ones are important (i.e. Are the number of books owned more or less important than the age in predicting income?). You can think of the prior as stiffening the model so that it doesn't loosely over fit all the data points.

An alternative to ridge regression which has some rather nice properties is **LASSO regression** (least absolute shrinkage and selection operator). This is equivalent to a Gaussian likelihood and a Laplacian (aka exponential) prior using ℓ_1 length

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{M\theta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{M\theta}) + \ln\left((2\pi)^N|\boldsymbol{C}|\right) - \lambda\,||\boldsymbol{\theta}||_1\right] \tag{7.3.3}$$

$||\boldsymbol{\theta}||_1 = \sum_i |\theta_i|$ There is no closed form solution for the maximum posterior, but there are many computer libraries that will find it for you numerically. The LASSO is mostly used in prediction and data compression. It has the property of forcing unimportant parameters to exactly zero rather than just to small values like for ridge regression. In this way it does a kind of automatic model selection by identifying which parameters can be discarded.

Problem 28. *The observed data $\boldsymbol{d}$ are related to some model parameters $\boldsymbol{\theta}$ by the vector relation*

$$\boldsymbol{d} = \boldsymbol{W\theta} + \boldsymbol{n} \tag{7.3.4}$$

where $\boldsymbol{n}$ is Gaussian distributed noise with covariance $\boldsymbol{N}_{ij} = \langle n_i n_j\rangle$ and $\boldsymbol{W}$ is a fixed matrix.

1. *What is the maximum likelihood solution for $\boldsymbol{\theta}$?*
2. *If you add a prior*

$$p(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{1/n}\sqrt{|\boldsymbol{A}|}}\; e^{-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{A}^{-1}\boldsymbol{\theta}} \tag{7.3.5}$$

what is the solution for $\boldsymbol{\theta}$ that maximizes the posterior? The result is known as a **Wiener filter**.

7.4 resampling techniques

The discussion of k-fold validation in the previous sections does relate to some other approximate techniques that are widely used in science. These methods seek to estimate the expectation value and variance of a statistic using the data itself without assuming a specific distribution for it. There are many such techniques, but the most widely used ones are bootstrap and jackknife resampling.

7.4.1 Bootstrap (nonparametric bootstrap) resampling

Let us say that we have n data points $\boldsymbol{x}_i$. The data here might be one number each trial or many. Consider the **bootstrap pdf**

$$f^{bs}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i) \tag{7.4.1}$$

where the δ is a Dirac delta function. This is the maximum likelihood estimate for the pdf of the data itself. It is an estimate for the pdf of the data. For a discrete distribution with finite outcomes it is clear that this will converge to the true distribution as $n \to \infty$. It is also true that in the continuous case that it will asymptotically converge to the real distribution.

Let's consider any statistic that is a function of these data point $t(x_1, x_2, \ldots x_n)$. Assuming that each data point is statistically independent, the expectation value of this statistic will be

$$E[t(x_1, x_2, \ldots x_n)] = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \; p(x_1)\ldots p(x_n) t(x_1, x_2, \ldots x_n). \tag{7.4.2}$$

Using the bootstrap estimation of the pdf (7.4.1) gives

$$E^{bs}[t(x_1, x_2, \ldots x_n)] = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \; f^{bs}(x_1)\ldots f^{bs}(x_n) t(x_1, x_2, \ldots x_n) \tag{7.4.3}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n} t\left(x_{i_1}, x_{i_2}, \ldots x_{i_n}\right) \tag{7.4.4}$$

These sums contain all possible combinations of the data in the "slots" of $t\left(x_{i_1}, x_{i_2}, \ldots x_{i_n}\right)$. All of these combinations except one, the original data, has repeated values in them.

The sums can be done in some simple cases. Let's consider the arithmetic mean,

$\overline{x} = \frac{1}{n}\sum_i x_i$

$$E^{bs}[\overline{x}] = \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n}\left(x_{i_1}+x_{i_2}+\ldots x_{i_n}\right) \tag{7.4.5}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n} x_{i_1} \qquad \text{all terms are the same} \tag{7.4.6}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n} x_{i_1}\left[\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n}\right] \tag{7.4.7}$$

$$= \frac{n^{n-1}}{n^n}\sum_{i_1=1}^{n} x_{i_1} \tag{7.4.8}$$

$$= \overline{x} \tag{7.4.9}$$

So the bootstrap mean for the sample mean is the same as the sample mean. Okay, now let's look at the bootstrap variance for the mean to estimate an error for it,

$$Var^{bs}[\overline{x}] = E^{bs}[\overline{x}^2] - \overline{x}^2 \tag{7.4.10}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n^2}\left(x_{i_1}+x_{i_2}+\ldots x_{i_n}\right)^2 - \overline{x}^2 \tag{7.4.11}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n^2}\left(x_{i_1}^2+x_{i_2}^2+\cdots+x_{i_1}x_{i_2}+x_{i_1}x_{i_3}\ldots\right) - \overline{x}^2 \tag{7.4.12}$$

$$= \frac{1}{n^n}\left[\frac{1}{n^2}\left(n^{n-1}\sum_{i_1=1}^{n}x_{i_1}^2 + n^{n-1}\sum_{i_2=1}^{n}x_{i_2}^2 + \cdots + n^{n-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}x_{i_1}x_{i_2} + n^{n-2}\sum_{i_1=1}^{n}\sum_{i_3=1}^{n}x_{i_1}x_{i_3}\ldots\right.\right. \tag{7.4.13}$$

$$= \left(\frac{1}{n^2}\sum_{i_1=1}^{n}x_{i_1}^2 + \frac{n(n-1)}{n^2}\overline{x}^2\right) - \overline{x}^2 \tag{7.4.14}$$

$$= \frac{1}{n}\left(\frac{1}{n}\sum_{i_1=1}^{n}x_{i_1}^2 - \overline{x}^2\right) \tag{7.4.15}$$

This is a asymptotically unbiased estimator of the variance. The ensemble average is

$$E\left[Var^{bs}[\overline{x}]\right] = \frac{\sigma^2}{n} \tag{7.4.16}$$

which is the same as for the true variance.

The average and mean are special cases. An estimate of the expectation value of any statistic can be found in this way. For example the least-squares estimate for a parameter is a statistic and we could estimate its expectation and variance in this way. In practice the sums in (7.4.4) are difficult to do explicitly because it has many term. There are

$$\binom{2n-1}{n} \tag{7.4.17}$$

distinct bootstrap samples which is 92,378 for $n = 10$ and $\simeq 4.53 \times 10^{58}$ for $n = 100$.[1] As a result the sums are nearly always estimated by Monte Carlo sampling which is quite easy to do in this case. The sums are over all combinations of the data values. We can choose n random values from the original data *with replacement* to get a new data set taken form the distribution $f^{bs}(x)$. We then calculate our statistic from this. We can do this as many times as we please to get a sample of values for our statistic.

Bootstrap resampling can also be used for a kind of poor man's model selection in the case of nested models. If an additional parameter is added to the model, we can calculate its variance among the bootstrap samples. If this variance is significantly smaller than the difference between the best fit value for the parameter and the value the parameter would have had in the simpler model then you can conclude that the new parameter is justified. For example, in fitting polynomials to data, if we add a new coefficient and the bootstrap variance for this coefficient indicates that it is consistent with zero then we would not except this new coefficient.

The justification for bootstrap sampling requires that the data points are independently distributed and that there are a lot of them. This might not be the case, but it is often used when there is some correlation between the data points if there is enough data that these correlations are not expected to influence the result. Sometimes consecutive data points or data points within some range in the independent variables are known to be correlated, but the data is spread over many of these ranges so that the correlation can be ignored. This needs to be investigated carefully however.

How many bootstrap samples should you take? As a rule of thumb it should be more than 1,000. You should take a look at the histogram of the sampled statistic to judge if the mean and variance are well defined. Feigelson & Babu (2012) recommend that the number of samples should be at least $n(\ln n)^2$. Note that it is important that all of the original n samples be independent.

A note on terminology: What I am calling bootstrap is sometimes called **non-parametric bootstrap**. **Parametric bootstrap** is where a specific model is chosen and random samples are drawn from it to find what the distribution of the statistic would be if this model where correct. I will generally call this second type **Monte**

[1] The number of bootstrap samples is n^n if you count permutations of the same set as different.

Carlo sampling. In science the term "bootstrap" usually refers to the nonparametric kind.

Problem 29. *Show that*

$$\binom{2n-1}{n} \tag{7.4.18}$$

is the number of distinct bootstrap samples. (Hint: see problem 7.)

7.4.2 Jackknife resampling

Another related technique is called the jackknife resampling. This is related to k-fold cross-validation (§7.1) where validation sets consists of one data point and the training sets are all but one, "leave-one-out" resampling. The use of this technique extends well beyond regression models as were presented previously. The term jackknife is more common when considering inference problems and cross-validation when considering prediction or machine learning problems.

Consider any statistic $t(x)$ that is being calculated from n data points. Let's take t_n to mean the expectation of the statistic for a sample of size n. In general this statistic will be biased relative to its value with an infinitely large data set. For a wide class of statistics we expect the leading order of the bias to be $\propto n^{-1}$ so we can write

$$t_n = t_\infty + \frac{t_b}{n} + \mathcal{O}(n^{-2}) \tag{7.4.19}$$

Applying this to the $n-1$ case gives

$$t_{n-1} = t_\infty + \frac{t_b}{n-1} + \mathcal{O}(n^{-2}) \tag{7.4.20}$$

Combining these we can eliminate the lowest order bias and solve for the asymptotic limit

$$\begin{aligned} t_\infty &= nt_n + (1-n)t_{n-1} + \mathcal{O}(n^{-2}) && (7.4.21)\\ &= t_n + (n-1)\left(t_n - t_{n-1}\right) + \mathcal{O}(n^{-2}) && (7.4.22) \end{aligned}$$

We have only one sample of size n, but we have n sub-samples of size $n-1$ and we can use them to estimate t_{n-1}. Take $t_{n-1}^{(i)}$ to be the statistic calculated from the data with the ith data point left out. The jackknife estimate for t_{n-1} is

$$\bar{t}_{n-1}^{J} \equiv \frac{1}{n}\sum_{i=1}^{n} t_{n-1}^{(i)} \tag{7.4.23}$$

From (7.4.22), the jackknife estimate for the bias in the statistic is

$$\begin{aligned} \text{bias}_t =& t_\infty - t_n && (7.4.24)\\ =& \simeq (n-1)\left[t_n - \bar{t}^J_{n-1}\right] && (7.4.25)\end{aligned}$$

Using this we get the jackknife estimate of the statistic t

$$\begin{aligned} t^J_n =& t_n + \text{bias}_t && (7.4.26)\\ =& nt_n + (1-n)\bar{t}^J_{n-1} && (7.4.27)\end{aligned}$$

which includes a correction for bias. We can also calculate the jackknife estimate of the variance for our statistic by applying the same logic,

$$Var^J[t_n] = \frac{n-1}{n}\sum_{i=1}^{n}\left(t^{(i)}_{n-1} - \bar{t}^J_{n-1}\right)^2 \tag{7.4.28}$$

This derivation is a bit dodgy actually because the expansion (7.4.19) is not unique. It can be shown in general that

$$\left\langle \sum_{i=1}^{n}\left(t^{(i)}_{n-1} - \bar{t}^J_{n-1}\right)^2 \right\rangle \leq Var[t_{n-1}] \tag{7.4.29}$$

and that the prefatory is a common scaling for statistics, i.e. $Var[t_n] = \frac{n-1}{n}Var[t_{n-1}] + \mathcal{O}(n^{-3})$ So the justification for the jackknife is really in that it is an educated guess and works well in many specific cases.

Problem 30. *If* $t_n = \frac{1}{n}\sum_i x_i = \overline{x}$ *show that* $t^J_n = \overline{x}$ *and*

$$Var^J[t_n] = \frac{s^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_i (x_i - \overline{x})^2 \tag{7.4.30}$$

i.e. an unbiased estimate of the variance.

Problem 31. *We know that* $t_n = \frac{1}{n}\sum_i (x_i - \overline{x})^2$ *is a biased estimator of the variance from previous chapters. Show that the jackknife estimate for* t_n *is* $t^J_n = \frac{1}{(n-1)}\sum_i (x_i - \overline{x})^2$ *which is not biased.*

These problems show that the jackknife estimator and variance estimate works in the simplest cases. In practice they are used when the statistic is more complicated than the mean or variance and the distribution of the data is unknown. The jackknife estimate and variance are found easily numerically.

Problem 32. *With a computer create some fake data by adding Gaussian noise to a points on a line. Fit a line to the fake data. Find the jackknife error and bias on the slope and intercept of the line.*

Problem 33. *Do the same as problem 32, but with bootstrap resampling instead of jackknife.*

7.5 Robustness & breakdown point

We have seen that some objectives in selecting an estimator might be for it to be unbiased or have a small variance. Another objective might be that it be insensitive to assumptions about the distribution of the data or that it be insensitive to contamination of the data with points that are outliers. Statistics with this property are called robust or **resistant statistics**. We have already seen that the sample median is more robust than the sample mean. The degree of robustness is measured with the **breakdown point**. Loosely, this is the fraction of the data that can be contaminated with data that is arbitrarily distributed before giving a wrong answer. The maximum breakdown point is 0.5 because at this point it would not be possible to differentiate between the contamination and the non-contamination. The sample median has a breakdown point of 0.5 because if $n/2-1$ of the points were arbitrarily large the median would still be within a cloud of the uncontaminated points. The sample mean has a breakdown point of 0 because a single arbitrarily large value will cause the mean to be arbitrarily large.

7.5.1 culling or trimming

One approach to making statistics more robust is to simply remove outliers. This can be made systematic by calculating the residuals for a model $|x_i| = |d_i - f_i(\boldsymbol{\theta})|$ where $f_i(\boldsymbol{\theta})$ is the model prediction for data point d_i and then discarding a fraction α of the data that has the largest residuals. The final statistic is calculated from this culled or trimmed data set. The **α-trimmed mean** is an example of this. It's breakdown point will be α rather than 0 for the sample mean. The **trimmed least squares** is another example. Here regression is done on the culled data set. These are examples of the general class of **L-estimators** which are linear combinations of order statistics.

For the trimmed least squares and other trimmed methods the culling or trimming depends on the initial model used to find the residues. If the data has a large degree of contamination you would expect them to be unreliable and this method can give spurious results. In general M-estimators (next section) are favored over these

trimming techniques although in meany cases they amount to the same thing and have the same dangers.

7.5.2 M-estimators

M-estimators are a generalization of the idea behind least-squared fitting and also a generalization of the maximum likelihood estimator. The method is generally using to make a fit more robust and thus less sensitive to outliers. The idea is to find the parameter values that minimize the function

$$\sum_i \rho(d_i, \boldsymbol{\theta}) \tag{7.5.1}$$

where $\rho(d_i, \boldsymbol{\theta})$ is called the **loss function**. For least squares $\rho(d_i, \boldsymbol{\theta}) = (d_i - f_i(\boldsymbol{\theta}))^2$ and for maximum likelihood $\rho(d_i, \boldsymbol{\theta}) = \ln[\mathcal{L}(d_i|\boldsymbol{\theta})]$. If one isn't so certain what the likelihood is or one is concerned that the data might be contaminated with spurious values, these particular choices might not be the best ones.

In choosing a loss function if seems advantageous that it have a minimum at $x = d_i - f_i(\boldsymbol{\theta}) = 0$. The quadratic form of the LS loss function at large values of x_i will make it sensitive to outliers. To reduce this one typically makes the function linear or constant at large $|x_i| = |d_i - f_i(\boldsymbol{\theta})|$. One option is just $\rho(x) = |x|$. This reduces the sensitivity to outliers, but it weights exact matches of the data and the model more strangely than is usually justified. In some cases it can also result in many equal minima (Think of a line fit to more than two data points.). Another possibility is the **Huber loss function**

$$\rho(x) = \begin{cases} \left(\frac{x}{s}\right)^2 & , \quad |x| < s \\ \frac{|x|}{s} & , \quad |x| > s \end{cases} \tag{7.5.2}$$

which is quadratic near its minimum. Still another choice is **Tukey's biweight function**

$$\rho(x) = \begin{cases} \frac{x^2}{2}\left(1 - \frac{x^2}{s^2} + \frac{x^4}{3s^4}\right) - \frac{s^2}{6} & , \quad x < s \\ 0 & , \quad x > s \end{cases} \tag{7.5.3}$$

These loss functions are shown in figure 7.2.

This approach involves introducing an extra scale s that is not known a priori. It essentially determines which data points will be considered outliers. Without some justification for this scale it can be equivalent to just throwing data out because you didn't like it which is not good practice so great care should be taken in applying this method in a scientific context. You could have a situation where you have some idea of what the variance of the events you are interested in is so that you can use an

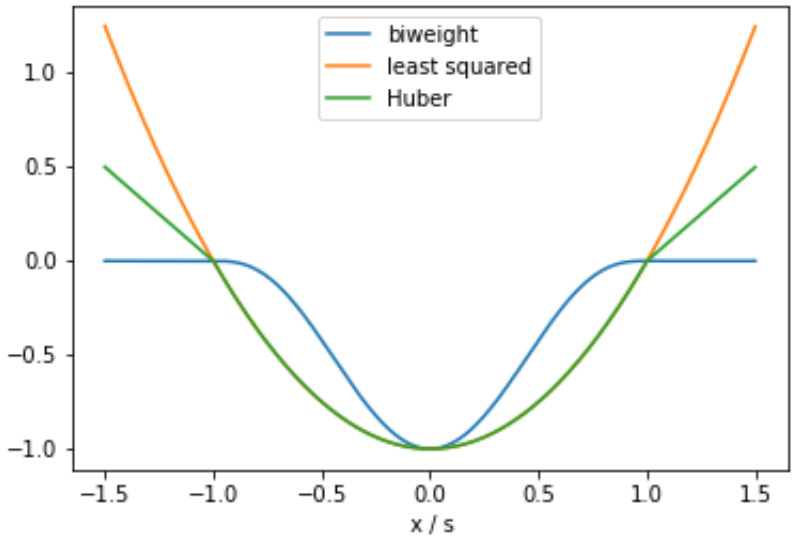

Figure 7.2: Some examples of loss functions.

M-estimator to essentially filter out background events. This will work as long as the number of background events is not so large that they dominates the sum of the loss functions. When this happens you are at the breakdown point of the M-estimator.

M-estimators[2] are only asymptotically normally distributed. The variance of the M-estimator can be estimated by bootstrapping the data. In other words, the estimator is calculated for each of many bootstrap resamplings of the data and the variance is taken.

[2]Besides M and L-estimators there are also R-estimators which are based on rank residuals. Jaeckel's rank regression is an example of this.

Chapter 8

Hypothesis testing & frequentist parameter fitting

Hypothesis testing is the frequentist version of Bayesian inference and model selection. In some cases it is easier to apply hypothesis testing and, as we have seen, Bayesian model selection has some undesirable characteristics, at least in some cases. There are many specific hypothesis tests that are commonly used in practice and so are essential to understand for any scientist.

Hypothesis testing takes a distinctly different approach to the question of whether a theory or hypothesis is supported by the data or not. The Bayesian method always compares the probability of competing models while hypothesis testing seeks to *disprove* a hypothesis by showing that the observed data would not be likely if the hypothesis were true. From the frequentist point of view there is no probability at all associated with parameters or models. (The mass of the electron is just what it is. If you run the experiment over again it would not change to some other value.) It is only the data that is probabilistic.

The basic steps in most applications of hypothesis test are as follows:

1. State the hypothesis as a well posed true or false question. The goal is the falsify this question. This is called the **null hypothesis** and denoted H_o.

2. Choose or invent a statistic (called a **goodness-of-fit statistic**) that is affected by the truth of the hypothesis.

3. Calculate the value of the statistic with the data.

4. Determine by analytic or numerical methods the probability distribution of the statistic given that the null hypothesis is true. Identify a direction or directions where the probability of getting that values for the statistic gets less and less

probable. This is usually as the statistic becomes very large absolutely or in magnitude.

5. With this distribution, calculate how probable it is for the statistic to be further in the direction of bad fits than the value calculated using the data. This is call the **p-value**.

6. If this probability is *sufficiently improbable* the hypothesis is ruled out. If it is not sufficiently improbable the hypothesis is *consistent* with this statistic.

To explain hypothesis testing let me tell a little fable. Someone brings you an unidentifiable animal. You say, "I think it is a dog." That is your hypothesis. You think about what a dog definitely has. Dogs have fur. That's your statistic. If the animal doesn't have fur you can say that the animal is not a dog. If it has fur you can say that this characteristic is consistent with it being a dog. You can't say it is a dog. There are other animals that have fur and there might be some other characteristics of this animal that are inconsistent with being a dog, say it has no claws. In most cases you can't even completely prove the hypothesis false, only unlikely. It might be a dog with a rare disease that made it lose it's fur or a rare genetically engineered dog that doesn't grow fur. Note that, in this case, there is no specific alternative hypothesis, it is either dog or not dog. Asking if it is a cat would be a different hypothesis and a different test. The existence of fur would not distinguish between dogs and cats. You would need a different statistic.

In most cases, we can never prove a hypothesis right. In fact, in some cases a statistical test might show consistency with a hypothesis that is clearly ruled out by another statistical test. The "null" in null hypothesis refers to the rejection process.

Errors or failures in hypothesis testing are by tradition categorized into two types:

- **Type I errors** - This is the case where the hypothesis is rejected, but is in fact true. You might call this a **false positive**.

- **Type II errors** - This is the case where the hypothesis is not rejected, but is in fact false. You might call this a **false negative**.

In the continuous case the probability of any particular data set, and thus a particular value for the statistic, is infinitesimally small so one must refer to a range in order to get a finite probability. The conclusion of the hypothesis test is then stated in two forms: "If the null hypothesis where true, the statistic would be larger (or smaller) than the measured value p fraction of the time." or "If the null hypothesis where true, the statistic would be further from its expectation value than the measured value p fraction of the time." By "time" I mean repeated trials under the condition that the null hypothesis is correct. The first case is called a **one-sided test** and the second a

two-sided test. Which one you use depends on the problem. p being the **p-value** also known as the **significance** of the test. The smaller it is the more evidence you have against the null hypothesis. $1-p$ is called the **confidence level** by which H_o is rejected.

The one and two-sided tests require that the statistic be one dimensional. This necessarily reduces the perhaps complicated distribution of the data to one number which is a simplification of the possible ways that the data can disagree with the hypothesis. A single statistic cannot test all aspects of a distribution. In choosing a statistic there is usually an implicit or explicit *alternative hypothesis*, H_1, that differs from the null hypothesis in some way and a good statistic is one that distinguishes between them well in that the probability distributions for the statistic given the two hypotheses do not overlap much. A statistic is thus tailored to test one aspect of the data's distribution.

8.1 mean of two populations are the same

A classic example of a frequentist hypothesis test is the test for the difference between the means of two populations. The statistic used is

$$Z = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.1.1}$$

The null hypothesis we will use is that the means of the populations are equal $\mu_1 = \mu_2$. (You could hypothesize that the difference of the means is some finite value or less than some value, etc.) If the data points $\{x_1\}$ and $\{x_2\}$ are normally distributed then Z is normally distributed because it is the some of normally distributed variables. A two-tailed test with a Gaussian distribution can be used to rule out this hypothesis. It is two tailed because we would usually consider any significant different in the means, no matter what the direction, to be in contradiction to the null hypotheses.

We might not know the measurement errors and need to estimate them from the data in which case the statistic

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{8.1.2}$$

is used. $S^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$ (see 4.2). With the same null hypothesis this statistic has very nearly a t-distribution with

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \tag{8.1.3}$$

degrees of freedom.

Problem 34. *Two measurements of the Hubble parameter, H_o, are done using different techniques. One reports $H_o = 70 \pm 5$ km/s/Mpc and the other $H_o = 80 \pm 7$ km/s/Mpc. Assuming these are marginalized Gaussian, 1 σ errors, how can you determine if these measurements are compatible with one another?*

What can you conclude if they are not consistent?

8.2 the variance of two populations are the same

We might also wonder if two populations have the same variance. You can test this with the statistic

$$f = \frac{S_1^2}{S_2^2} \tag{8.2.1}$$

This is of the form

$$\frac{X_\alpha^2/\alpha}{X_\beta^2/\beta} \tag{8.2.2}$$

where X_α^2 is a χ_α^2 distributed variable. In this case $\alpha = n_1 - 1$ and $\beta = n_2 - 1$. Such a ratio has a **F-distribution**, specifically F_{n_1-1,n_2-1}. The pdf is

$$p_F(f) = \frac{\alpha^{\alpha/2}\beta^{\beta/2} f^{\alpha/2-1}}{B(\alpha/2, \beta/2)(\alpha f + \beta)^{(\alpha+\beta)/2}} \tag{8.2.3}$$

where $B(\alpha, \beta)$ is the beta function. Thus this is called the **F-test** for the difference of two variances. This is a **student's t test** for the difference of two means. A two sided test is appropriate for the hypothesis test of $\sigma_1^2 = \sigma_2^2$. A high value of f will tend to be high if $\sigma_1^2 > \sigma_2^2$ so if the alternative hypothesis is specifically that $\sigma_1^2 > \sigma_2^2$ you could us an upper one tailed test.

Problem 35. *You work for a company that manufactures widgets. You have been having problems with your manufacturing process because of variations in the quality of the chemicals used. You buy these chemicals from two companies, A and B. You want to test if one of these companies is significantly more reliable than the other. You take samples from each company and have the concentration of chemical α measured.*

$$\rho_\alpha^A = [97, 90, 95, 90, 101, 99, 99, 107, 102, 95] \tag{8.2.4}$$

$$\rho_\alpha^B = [101, 94, 93, 96, 94, 97, 94, 98, 98, 90, 90, 95] \tag{8.2.5}$$

Determine if one has a significantly different variance than the other.

In practice the F-test may be unreliable if the samples are not Gaussian, i.e. it is sensitive to non-normality. As a result it might be more of a test of normality rather than the equality of the variances in some cases. Bartlett's test is an alternative test for the equality of variances that is also sensitive to non-normality, but somewhat less so. Levene's test is yet another test that is purported to be insensitive to non-normality. There are many specialized hypothesis tests in the literature (and on the internet) with known or approximatable p-values. If confronted with a specific problem it might pay to do some research for an appropriate test.

8.3 χ^2 test for the constancy of a signal

Let us consider a simple case that we have analyzed using Bayesian inference and see how it is analyzed using hypothesis testing. We (again) have n independent data points, x_i like in the wine example at the beginning of chapter 5.

Let us look at three important, but subtly different null hypothesizes pertaining to this data set that will come up later in a more complicated form.

Null Hypothesis I *The signal is constant, it's value is equal to μ and the errors are Gaussian distributed with the known variance σ.*

Here μ is some fixed value that is not derived from the data, maybe zero. The measurement errors are also fixed and known. The likelihood with this hypothesis is

$$p(x_i|\mu) = \frac{1}{(2\pi))^{n/2}\sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \tag{8.3.1}$$

Let us use the statistic

$$X^2(\boldsymbol{x},\mu) = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} \tag{8.3.2}$$

We know from section 3.15 that this statistic is χ^2 distributed with n degrees of freedom. We can calculate X^2 with our data and find the cumulative probability up to this value $F_{\chi^2_n}(X^2)$. If this is large then we can say the a mean of μ is ruled out at the $1 - F_{\chi^2_n}(X^2)$ confidence level.

Null Hypothesis II *The signal is constant and the errors are Gaussian distributed with the known variance σ.*

We have relaxed the requirement that the mean has some specific value. For this hypothesis we might use a statistic is very similar

$$X^2(\boldsymbol{x}, \overline{x}) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2} \tag{8.3.3}$$

with the sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ in place of an hypothesized mean. This statistic will not be χ^2_n distributed however. As we saw in section 4.2 the statistic is χ^2_{n-1} distributed. This is because one degree of freedom has been lost because of the constraint that $\overline{x}$ be the sample mean.

An instructive way to look at the degrees of freedom that will be useful later is as a projection of the data into subspaces. Consider each possible data set to be an n-dimensional vector. Consider decomposing the data vector as follows

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \overline{x} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_1 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix} \tag{8.3.4}$$

$$= (\boldsymbol{x} \cdot \hat{\boldsymbol{m}})\, \hat{\boldsymbol{m}} + [\boldsymbol{x} - (\boldsymbol{x} \cdot \hat{\boldsymbol{m}})\, \hat{\boldsymbol{m}}] \tag{8.3.5}$$

The vector on the left is Gaussian distributed with n degrees of freedom as in hypothesis I. The first vector on the right is the projection of the data vector onto the $\hat{\boldsymbol{m}} = (1, 1, \dots)/\sqrt{n}$ vector. I'll call this the "mean vector". It will be in a one dimensional subspace and Gaussian distributed with one degree of freedom (There will be a factor of n that comes from the magnitude of the vector and reduces the variance in $\overline{x}$ to σ^2/n). The remaining vector will be in an $n-1$ dimensional subspace.

Using (8.3.4) the magnitude of the data vector can be expressed as

$$\boldsymbol{x} \cdot \boldsymbol{x} = n\overline{x}^2 + (\boldsymbol{x} - \overline{x}) \cdot (\boldsymbol{x} - \overline{x}). \tag{8.3.6}$$

Using this we can find a relationship between the statistics $X^2(\boldsymbol{x}, \mu)$ and $X^2(\boldsymbol{x}, \overline{x})$,

$$\sigma^2 X^2(\boldsymbol{x}, \mu) = (\boldsymbol{x} - \mu\sqrt{n}\,\hat{\boldsymbol{m}}) \cdot (\boldsymbol{x} - \mu\sqrt{n}\,\hat{\boldsymbol{m}}) \tag{8.3.7}$$

$$= \boldsymbol{x} \cdot \boldsymbol{x} - 2\mu\sqrt{n}(\boldsymbol{x} \cdot \hat{\boldsymbol{m}}) + \mu^2 n \tag{8.3.8}$$

$$= n\overline{x}^2 + (\boldsymbol{x} - \overline{x}) \cdot (\boldsymbol{x} - \overline{x}) - 2\mu\sqrt{n}(\boldsymbol{x} \cdot \hat{\boldsymbol{m}}) + \mu^2 n \tag{8.3.9}$$

$$= n(\overline{x} - \mu)^2 + (\boldsymbol{x} - \overline{x}) \cdot (\boldsymbol{x} - \overline{x}) \tag{8.3.10}$$

$$= n(\overline{x} - \mu)^2 + \sigma^2 X^2(\boldsymbol{x}, \overline{x}). \tag{8.3.11}$$

Since $X^2(\boldsymbol{x}, \mu)$ is χ^2_n distributed (the sum of the squares of n normally distributed variables) and the first term on the right is χ^2_1 distributed since $\overline{x}$ is a single normally distributed variable it follows that the last term is χ^2_{n-1} distributed. The

two parts of the data vector (8.3.4) are orthogonal so there will be no cross-terms or cross-correlation between the components. In other words, they are statistically independent!

We can apply a **chi-squared test** by calculating $X^2(\boldsymbol{x}, \overline{x})$ and seeing if its χ^2_{n-1} p-value is small. If it is we reject hypothesis that the data is constant. In this case, a one-sided test is advisable because if there is some variation in the data we would expect $X^2(\boldsymbol{x}, \overline{x})$ to be large rather than small. If $X^2(\boldsymbol{x}, \overline{x})$ is exceptionally small according to the χ^2_{n-1}-distribution, we have probably overestimated the errors.

This type of hypothesis is akin to doing Bayesian model selection in that it gives a criterion for rejecting a model irrespective of the specific values of the parameters, i.e. μ. If the data points were increasing with time, for example, you would expect $X^2(\boldsymbol{x}, \overline{x})$ to be large and you would reject hypothesis I.

Null Hypothesis III *Given that the signal is constant, it's value is equal to μ and the errors are Gaussian distributed with the known variance σ.*

This sounds a lot like hypothesis I, but it is not the same. We will not take into consideration the possibility that the signal is not constant when calculating distribution of our statistic. To do this we want a statistic that is not sensitive to the non-constancy of the signal. A statistic based only on the first part of the decomposition (8.3.4), the part not included in the previous test, would do exactly that. We can use the first term in (8.3.11)

$$X^2 = \sum_{i=1}^{n} \frac{(\overline{x} - \mu)^2}{\sigma^2} = \frac{n(\overline{x} - \mu)^2}{\sigma^2} \tag{8.3.12}$$

This is the magnitude of the projection of the data vector onto the mean vector $\hat{\boldsymbol{m}}$, a one dimensional space. It has only one degree of freedom and is thus χ^2_1 distributed. This can be used to put constraints on μ by excluding those values of μ whose p-value is below some threshold, say 5 or 1%.

8.4 The tail of three χ's

In general, for normally distributed data the likelihood can be written

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{C}|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\theta})\right)^T \boldsymbol{C}^{-1} \left(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\theta})\right)\right] \tag{8.4.1}$$

where $\boldsymbol{f}(\boldsymbol{\theta})$ is the, possibly nonlinear, relationship between the parameters and the prediction for the data. From this we know that if $\boldsymbol{f}(\boldsymbol{\theta})$ is the correct model

$$\chi^2(\boldsymbol{x}, \theta) = \left(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\theta})\right)^T \boldsymbol{C}^{-1} \left(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{\theta})\right) \tag{8.4.2}$$

is chi-squared distributed with n degrees of freedom where n is the number of data points.

For linear models we can go a bit further. In this case, $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{M}\boldsymbol{\theta}$, as discussed in chapter 6. The χ^2 can be written as

$$\begin{aligned}
\chi^2(\boldsymbol{x},\boldsymbol{\theta}) =& (\boldsymbol{x}-\boldsymbol{M}\boldsymbol{\theta})^T \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{M}\boldsymbol{\theta}) && (8.4.3)\\
=& \left(\boldsymbol{x}-\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})-\boldsymbol{M}\hat{\boldsymbol{\theta}}\right)^T \boldsymbol{C}^{-1} \left(\boldsymbol{x}-\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})-\boldsymbol{M}\hat{\boldsymbol{\theta}}\right) && (8.4.4)\\
=& (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + (\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}}) && (8.4.5)\\
& - (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}}) - (\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}})^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) && (8.4.6)
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood solution

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{x}. \tag{8.4.7}$$

The terms on line (8.4.6) are equal and both are zero. To see this consider

$$\begin{aligned}
\boldsymbol{M}^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}}) &= \boldsymbol{M}^T\boldsymbol{C}^{-1}\left[\boldsymbol{I}-\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\right]\boldsymbol{x} && (8.4.8)\\
&= \left[\boldsymbol{M}^T\boldsymbol{C}^{-1}-\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\right]\boldsymbol{x} && (8.4.9)\\
&= \left[\boldsymbol{M}^T\boldsymbol{C}^{-1}-\boldsymbol{M}^T\boldsymbol{C}^{-1}\right]\boldsymbol{x} && (8.4.10)\\
&= 0 && (8.4.11)
\end{aligned}$$

the first term in (8.4.6) contains this factor and the second term contains the transpose of this which must also be zero. This can be interpreted as the vector $\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$ being orthogonal to the vector $\boldsymbol{z} = (\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}})$ in the sense that $\boldsymbol{z}^T\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{z} = 0$. This also means that they are statistically independent since these vectors are normally distributed.

So the χ^2 can be decomposed into two parts that are statistically independent,

$$\begin{aligned}
\chi^2(\boldsymbol{x},\theta) =& (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + (\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}}) && (8.4.12)\\
=& \chi^2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) + \chi^2_{min}. && (8.4.13)
\end{aligned}$$

The second one contains no parameters and is clearly the minimum of $\chi^2(\boldsymbol{x},\theta)$ which is attained at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

This justifies the factorization of the likelihood that was used in section 10.1,

$$\mathcal{L} = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{C}|}} e^{-\frac{1}{2}\chi^2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})} e^{-\frac{1}{2}\chi^2_{min}}. \tag{8.4.14}$$

The question remains of how these two χ^2s are distributed and what they can be used for. You might think that since each one is constructed out of n normally distributed numbers they should both be χ^2 distributed with n degrees of freedom, but this is wrong.

The matrix $\boldsymbol{M}$ takes any point in parameter space and transforms it to a point in "data space", i.e. $\boldsymbol{x} = \boldsymbol{M\theta}$ is an n dimensional vector. But, in an over determined problem, the dimension of parameter space, I will call it m, is smaller than n so $\boldsymbol{M\theta}$ must be in a m dimensional subspace of data space. More accurately, the dimension is equal to the rank of $\boldsymbol{M}$ which is equal to m if it is full rank. Not all possible vectors are accessible by through $\boldsymbol{M}$.

The pseudo-inverse $\boldsymbol{M}^+ = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}$ gives us a way of going from data space to parameter space. Since it is a linear operator that goes from a higher dimensional space to a lower dimensional space it must have a null-space in data-space. In other words, there are vectors in data space that do not have a maximum likelihood solution in parameter space and if you add one of these vectors to the data the maximum likelihood solution, $\hat{\boldsymbol{\theta}}$, will not change.

Now consider the operator

$$\boldsymbol{P} = \boldsymbol{M}\boldsymbol{M}^+ = \boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}. \tag{8.4.15}$$

We can think of this as taking a data vector to parameter space and back again. In the process, the part of the data vector in the null-space of $\boldsymbol{M}^+$ is lost. It is easy to see that

$$\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P} \tag{8.4.16}$$

$$[\boldsymbol{P}\boldsymbol{x}]^T\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{P}\boldsymbol{y} = [\boldsymbol{P}\boldsymbol{x}]^T\boldsymbol{C}^{-1}\boldsymbol{P}\boldsymbol{y} \tag{8.4.17}$$

which are the requirements for a projection operator. So $\boldsymbol{P}$ projects the data vector onto the subspace of data-space that influences the parameters and removes the components that do not. $\boldsymbol{P}\boldsymbol{x} = \boldsymbol{M}\hat{\theta}(\boldsymbol{x})$ actually contains m normally distributed variable and not n. Likewise, the complement of $\boldsymbol{P}$, $\bar{\boldsymbol{P}} \equiv (\boldsymbol{I} - \boldsymbol{P})$ is a projection operator that projects into the complimentary subspace and $\bar{\boldsymbol{P}}\boldsymbol{x}$ contains $n-m$ normally distributed variable.

Projection operators take a vector and return only its components that are within its *range*, the subspace that they are projecting into. They remove any components that are within its null space. Any data vector can be decomposed into orthogonal parts

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} + \bar{\boldsymbol{P}}\boldsymbol{x} \tag{8.4.18}$$

$$= \boldsymbol{x}_k + \boldsymbol{x}_{n-k} \tag{8.4.19}$$

where $\boldsymbol{x}_k$ is in the k dimensional space that "influences" the model parameters and $\boldsymbol{x}_{n-k}$ is in the $n-k$ dimensional space that is independent of the parameters. This is a generalization of the projection onto the mean vector we saw in the previous section.

You can see that χ^2_{min} contains only vectors in the $\bar{\boldsymbol{P}}\boldsymbol{x}$ subspace

$$\chi^2_{min} = \left[(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\right]^T \boldsymbol{C}^{-1} \left[(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\right] = \left[\bar{\boldsymbol{P}}\boldsymbol{x}\right]^T \boldsymbol{C}^{-1} \left[\bar{\boldsymbol{P}}\boldsymbol{x}\right] \tag{8.4.20}$$

and $\chi^2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ contains only vectors in the $\boldsymbol{P}\boldsymbol{x}$ subspace – $\boldsymbol{M}\theta$ must also be in this subspace. The vanishing of the cross terms in (8.4.6) shows that these subspaces are orthogonal and statistically independent.

To summarize the geometric interpretation of what is going on here:

- $\boldsymbol{M}$ takes a parameter vector to data space, but since these spaces have different dimensions it cannot cover all of data space and will not be square.

- The matrix $\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}$ takes the data to the best fit model parameters - data space to a point in parameter space.

- The matrix $\boldsymbol{P} = \boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}$ projects the data onto the subspace that has a direct affect on the parameters of the model. This is a k dimensional subspace because there are k parameters. The extra $\boldsymbol{M}$ goes from parameter space back to data space.

- The matrix $\bar{\boldsymbol{P}} = \boldsymbol{I}-\boldsymbol{P} = \boldsymbol{I}-\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}$ projects the data into the subspace that does not affect the parameters. These are the degrees of freedom that are not absorbed by fitting the model.

This χ^2 can be used as a goodness-of-fit statistic for hypothesis testing even when the model is nonlinear. The null hypothesis is "The data is normally distributed, the parameters are related to the data points by $\boldsymbol{f}(\boldsymbol{\theta})$ **and** the parameters have values $\boldsymbol{\theta}$. We could further find the parameter set that minimizes χ^2 and compare its value to a χ^2_n distribution. If we can rule this model out at some confidence level we will know that all other parameter values are ruled out with higher confidence.

In the case of the a nonlinear model, it is not guaranteed that the number of degrees of freedom for χ^2_{min} is $n-m$. The most we can say is that it is less than or equal to $n-1$. This is because the equation for the maximum,

$$\frac{\partial\chi^2}{\partial\theta_\alpha} = 0, \tag{8.4.21}$$

does not necessarily consist of m independent equations.

An example that illustrates how different fitting a nonlinear model can be from fitting a linear model is the case of fitting the model

$$y = A\cos(\omega x + x_o) \tag{8.4.22}$$

to y and x values where A, ω and x_o are the parameters. This model can fit any data set perfectly, as long as there are no repeated values of x, so the minimum χ^2 will always be zero. The model is effectively undetermined even though there may be many more data points than parameters. There is no concept of the reduced number of degrees of freedom for χ^2_{min}. However, $\chi^2(A, \omega, x_o)$ is still χ^2_n distributed. The best fit parameters can be rejected on the bases of a left-sided hypothesis test! A perfect fit is improbably given that there is noise. There is a region of (A, ω, x_o)-space where $\chi^2(A, \omega, x_o)$ has reasonably probable values and a region where it has improbable values so χ^2 parameter estimation can be done.[1]

8.5 Hypothesis testing with linear models & Gaussian likelihoods

The two χ^2's, χ^2_{min} and $\chi^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ allow us to make more precise hypothesis tests in the case of linear models.

In the case of a linear model, χ^2_{min} will contain only $n - m$ normally distributed variables and thus the hypothesis that the model is the correct one should be rejected according to a χ^2_{n-m} distribution. This contains no specific parameters values. It is a global hypothesis test on whether the model can be ruled out irrespective of what the actual parameter values are. This is the kind of test that is not possible in Bayesian inference. Note that in the nonlinear case you can do the same test by finding χ^2_{min} numerically, but its significance should be assessed using a χ^2_n distribution because we don't know that χ^2_{min} contains only $n - m$ independent random numbers. This will be a weaker test because the distribution is broader.

$\chi^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ can be used for parameter estimation in the case of a linear model. The null hypothesis is that "Given that the model is correct and the errors are normally distributed, the parameters have values $\boldsymbol{\theta}$." The statistic will be χ^2_m distributed, which is potentially a much stronger statistical test than using the original χ^2 as you would do for a nonlinear model. The $\chi^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ essentially ignores all the components of the data that have no influence on the fitting of the parameters.

The **reduced** χ^2 or the χ^2 **per degree of freedom** is χ^2/n where n is the number of degrees of freedom. Since the mean of the χ^2 distribution is equal to its number of degrees of freedom, it is expected that the reduced χ^2 will be ~ 1.

[1] Note that this is not the case of fitting the discrete Fourier transform (DFT) because in that case ω and x_o are not parameters.

Problem 36. *Prove (8.4.16) and (8.4.16).*

Problem 37. *Prove that the range of matrix $\bar{\boldsymbol{P}}$ has dimension k in the following steps:*

1. *Show that $\bar{\boldsymbol{P}}\bar{\boldsymbol{P}} = \bar{\boldsymbol{P}}$ and that this implies that all the eigenvalues of $\bar{\boldsymbol{P}}$ are either 1 or 0.*
2. *Show that $tr\left[\bar{\boldsymbol{P}}\right] = n - k$ and why this implies that there are k eigenvalues that are zero and $n - k$ that are one.*

In the case of a nonlinear model we cannot be sure what the appropriate distribution is for χ^2_{max}, but we know that $\chi^2(\boldsymbol{x}, \boldsymbol{\theta})$ would be χ^2_n distributed if $\boldsymbol{\theta}$ is the correct model. A conservative hypothesis test would be to eliminate particular parameter sets where the p-values are small.

8.6 χ^2 model testing

Now let's revisit the question of how many parameters are too many parameters. If we fit a model to the data that is too simple we would expect χ^2_{min} to be large. If we add parameters to our model we would expect χ^2_{min} to fall if it is fitting the data better. If the model is over-fit we expect χ^2_{min} to be too small in the sense that it would be unlikely for it to be smaller than the observed value according to the χ^2_n distribution or χ^2_{n-m} for a linear model. This suggests the following procedure. Start with a simple model for which the χ^2_{min} is too large and then add parameters until the χ^2_{min} has a reasonable value in that its p-value is $\sim 1/2$.

For a linear model we can sharpen the discriminatory power somewhat. Let us say we have a linear model with k parameters and another model with $m = k - r$ parameters, $k > m$. We make the $X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ statistics for each model and relate them with

$$X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_M) = X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_K) + \left(X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_M) - X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_K)\right) \tag{8.6.1}$$

$$= X^2_K + \Delta X^2 \tag{8.6.2}$$

It can be shown that X^2_K and ΔX^2 are statistically independent and that $X^2_M \sim \chi^2_{n-m}$, $X^2_K \sim \chi^2_{n-k}$ and $\Delta X^2 \sim \chi^2_r$ in the same way as in section 8.4. Again this can be seen in terms of projections in data space. X^2_K contains only components that are not fixed by the k parameters in model K. X^2_M is in the larger space that is not constrained the m parameters. ΔX^2 is in the space that is a subspace of X^2_M's, but not in X^2_K's

- the space that is constrained by the extra parameters in X_K^2. So ΔX^2 and X_K^2 are in orthogonal spaces and ΔX^2 is in a $r = k - m$ dimensional subspace.

This suggests the following χ^2 test for model selection among nested linear models with Gaussian noise. The null hypothesis is that model K, the more complex one, is the correct model. In this case $\Delta X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_K) \sim \chi^2_{k-m}$. Usually one new parameter is introduced at a time so $r = k - m = 1$. ΔX^2 is the contribution to the total $X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_k)$ from the additional parameter. The measured ΔX^2 is compared to this distribution. If the p-value is small the additional parameter is not justified. You can use this to test which parameters are justified and which ones are not. Note that if a model is too complex we would expect $X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ to be too small, $< n - k$, so we should use a two sided test to evaluate the significance of a model.

Another common and related χ^2 based model selection test is called the **F-test**. Since ΔX^2 and X_K^2 are uncorrelated, χ^2 distributed variables, their ratio

$$f = \frac{\Delta X^2}{X_k^2}\left(\frac{n-k}{r}\right) \tag{8.6.3}$$

is a $F_{r,n-k}$ distributed variable.

In particular if we add one parameter to the model we expect that

$$f = \frac{\Delta X^2}{X_k^2}(n-k) \tag{8.6.4}$$

will be $F_{1,n-k}$. It terns out that if $f \sim F_{1,n-k}$ and $f = z^2$ then z is t-distributed with $n - k$ degrees of freedom. If the measured value of f (or z) has a small chance of occurring according to its distribution we conclude that it is justified to add this parameter.

These tests can be applied (with care) more generally than for just the linear Gaussian case, see section 12.7.

Problem 38. *Show that X_m^2 and ΔX^2 are statistically independent.*

8.7 frequentist confidence intervals

Once it has been determined that the best fit linear model has an acceptable χ^2_{min} it is time to find the error bars, or confidence intervals, for the parameters.

In section 6.1 we found that the posterior for a linear model is a Gaussian centered on the MLE (equation 6.1.11). It follows that

$$X^2(\boldsymbol{x}, \boldsymbol{\theta}) = X^2(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{M}^T \boldsymbol{C}^{-1} \boldsymbol{M} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{8.7.1}$$

where the likelihood is

$$\mathcal{L} = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{C}|}} e^{-\frac{1}{2}X^2(\boldsymbol{x},\boldsymbol{\theta})}. \tag{8.7.2}$$

You can see that $X^2(\boldsymbol{x},\hat{\boldsymbol{\theta}})$ was completely ignored in the Bayesian parameter estimate, while it is the only part that the global frequentist hypothesis test for the model was based on. The first term in (8.7.1) will be χ^2_{n-k} and the second one χ^2_k.

The boundaries of the confidence region (or interval in one dimension) in θ-space are drawn on contours of equal likelihood i.e.

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) = \text{constant} \tag{8.7.3}$$

The confidence level is taken from a χ^2_k distribution. Note that this is different from a Bayesian "**credibility region**" where the posterior is integrated within the boundaries of the region. We can look at this as marginalizing over all the modes of the data that do not effect the model fit and then using the modes that do effect the fit to constrain the model.

The simplest example is from our problem of finding the mean. Here

$$X^2 = \sum_i^n \frac{(x_i-\mu)^2}{\sigma^2} \tag{8.7.4}$$

$$= \sum_i^n \frac{(x_i-\overline{x})^2}{\sigma^2} + \frac{(\mu-\overline{x})^2}{\sigma^2/n} \tag{8.7.5}$$

The first part we saw before for testing if the signal is constant in section 8.3. The second part will be χ^2_1 distributed. $F_{\chi^2_1}(4) = 0.954$ so 95.4% confidence interval for μ is $\overline{x} - \frac{2\sigma}{\sqrt{n}}$ to $\overline{x} + \frac{2\sigma}{\sqrt{n}}$ or usually written $\mu = \overline{x} \pm \frac{2\sigma}{\sqrt{n}}$(95% cf).

Note that this does *not* mean that the mean has a 95% chance of being within this range. It means that if the mean where outside of this range the probability of getting a sample mean that is further away than was measured is less than 5%. (Kind of a convoluted statement really).

In general, if we gave a goodness-of-fit statistic $t(D;\boldsymbol{\theta})$ which is a function of the data D and the parameters $\boldsymbol{\theta}$ we find the confidence regions by plotting the contours of the cumulative distribution $F\left[t(D;\boldsymbol{\theta})\right] = 0.68$, 0.95, 0.99, etc. as a function of the parameters $\boldsymbol{\theta}$ (or in the case of a two-sided test $F\left[t(D;\boldsymbol{\theta})\right] - F\left[-t(D;\boldsymbol{\theta})\right]$). $t(D;\theta)$ could be any function that you would expect to be small (large) when the data fits the model well and large (small) when the data does not fit well (either in size or absolute value). Classical statistics may have known distributions, $F[t]$. The distribution of other statistics can be estimated by Monte Carlo, but this can be very time consuming since it needs to be done for every parameter set $\boldsymbol{\theta}$ to map out the confidence levels.

Problem 39. *Show that (8.7.3) is a function of only the $\boldsymbol{x}_k$ components of $\boldsymbol{x}$.*

8.7.1 Frequentest confidence and Bayesian credibility regions

As we have seen, the Bayesian credibility region represents a the fraction of the posterior probability. It says "there is a X% probability that the parameter is within this region." One can define a credibility region for one parameter or a subset of the parameters by integrating, or marginalizing, over the other parameters. This is because the rules of probability require this.

The frequentist assigns no probability to parameters. If the true parameter value is outside the confidence region then there would be less then a X% chance of getting data that fit the model worse than was measured. The region does not represent a probability directly. Integrals in parameter space have no meaning from a frequentist point of view. To find the confidence region for a subset of parameters we should *project* the boundary of the confidence region onto the parameters. When we say that if the value of parameter $\theta_1 = x$ is not likely to produce the observed data we mean no matter what the other parameter values are so we must take the values for the other parameters that produce the largest probability of producing the observed data given that $\theta_1 = x$.

8.8 Sufficient & ancillary statistics

We have investigating χ^2 hypothesis testing with linear models in some detail now. It illustrates many principles that can be applied more generally. One example is the difference between **sufficient statistics** and **ancillary statistics**.

A statistic, $t(\boldsymbol{d})$ is called a sufficient statistic for a parameter θ if it contains all the information in the data, $\boldsymbol{d}$, about that parameters. In this case the likelihood can be written

$$P(\boldsymbol{d}|\theta) = f(\boldsymbol{d})g\left(t(\boldsymbol{d})|\theta\right) \tag{8.8.1}$$

We have already seen in chapter 5 that the likelihood for independent Gaussian distributed data can be written in terms of only the sample mean, $\overline{x}$ and sample variance, Δ^2, so these are sufficient statistics in this case. You can see that from a Bayesian point a view the function $f(\boldsymbol{d})$ would drop out of the posterior and the data would not be there except through the sufficient statistics.

An ancillary statistic is in a sense the opposite. If $u(\boldsymbol{d})$ is ancillary its distribution is not dependent on any of the parameters,

$$P(\boldsymbol{d}|\theta) = f\left(u(\boldsymbol{d})\right) g\left(\boldsymbol{d}|\theta\right) \tag{8.8.2}$$

As we have seen, in the Gaussian / linear case the

$$-2\ln\mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) = X^2(\boldsymbol{d},\boldsymbol{\theta}) + const. \tag{8.8.3}$$

$$= X^2(\boldsymbol{d},\hat{\boldsymbol{\theta}}(\boldsymbol{d})) + X^2(\hat{\boldsymbol{\theta}}(\boldsymbol{d}),\boldsymbol{\theta}) + const. \tag{8.8.4}$$

So $X^2(\boldsymbol{d},\hat{\boldsymbol{\theta}}(\boldsymbol{d}))$ is an ancillary statistic and $X^2(\hat{\boldsymbol{\theta}}(\boldsymbol{d}),\boldsymbol{\theta})$ is sufficient for the linear parameters.

A sufficient statistic can be used to do parameter estimation and an ancillary statistic can be used to do a global goodness-of-fit or model selection test. However, it is not required that you have an ancillary statistic to do a goodness-of-fit test, although it does simplify things.

Chapter 9

Other hypothesis tests

In this chapter we look at some other hypothesis tests that are commonly used. These tests use something other than χ^2 as a goodness-of-fit statistic or use χ^2 as an approximation for some other statistic.

9.1 Pearson's correlation coefficient

Let us consider the problem of determining if two variables are correlated. For example, are body weight and life expectancy correlated or is the average age of stars correlated with the size of the galaxy they are in. **Pearson's correlation coefficient** is

$$r_{xy} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{9.1.1}$$

If the variable are uncorrelated we would expect, just through symmetry, this statistic to be zero on average. The quantity

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{9.1.2}$$

is t-distributed with $n-2$ degrees of freedom if the x_i's and y_i's are independent and normally distributed. This can be used to perform a hypothesis test for the absence of correlations. But if the data is not normally distributed you cannot evaluate the significance of the statistic in this way.

To test a nonzero correlation Fisher's transform of r

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{9.1.3}$$

is used to perform hypothesis tests. This statistic is more normally distributed for a range of distributions

$$F(r) \sim \mathcal{N}\left(F(\rho), \sigma^2 = \frac{1}{n-3}\right) \tag{9.1.4}$$

This can be used to test the hypothesis that the sample pairs are drawn from a multivariate normal distribution with correlation $\rho = C_{xy}/\sqrt{C_{xx}C_{yy}}$ where $\boldsymbol{C}$ is the covariance matrix.

9.2 Comparing data to a distribution

Let us return to the problem of determining how some measurements, such as star luminosities or photon energies or just any noisy data, are distributed. We tackled this problem with Bayesian parameters estimation already. In the frequentist context we can ask not only what the parameters of the distribution are, but also whether the data is consistent with the model distribution at all. We might ask if our data really is consistent with being normally distributed instead of just assuming that is the case, for example.

9.2.1 Q-Q plot

A quick and dirty, but often very effective, way to compare the distribution of some one dimensional data to a model distribution is to make a **Quantile-Quantile plot**. If the sorted data is x_i, and the cumulative the distribution is $F(x)$, then the Q-Q plot is i/n vs $F(x_i)$. i/n should be an approximation of the cumulative distribution so in the limit of large sample size this curve should converge to a line from $(0, 0)$ to $(1, 1)$. Figure 9.1 shows an example. If the sample size is small and/or the models being compared are not very different it might not be so clear which model is better. It helps to do some simulations to see what is expected for a particular distribution and sample size. As is, the Q-Q plot does not provide a quantitative reason to reject or accept a model distribution, but can be used as a qualitative aid to diagnose what is the biggest disagreement between the model and data.[1]

9.2.2 Binned data χ^2 test

One option for determining the distribution of data that is commonly used in astronomy is to bin the data, divide range of the data into intervals and count the number

[1] You could think of several ways of making this into a quantitative discriminator. They would probably be equivalent to or related to the standard tests that are discussed later in the section.

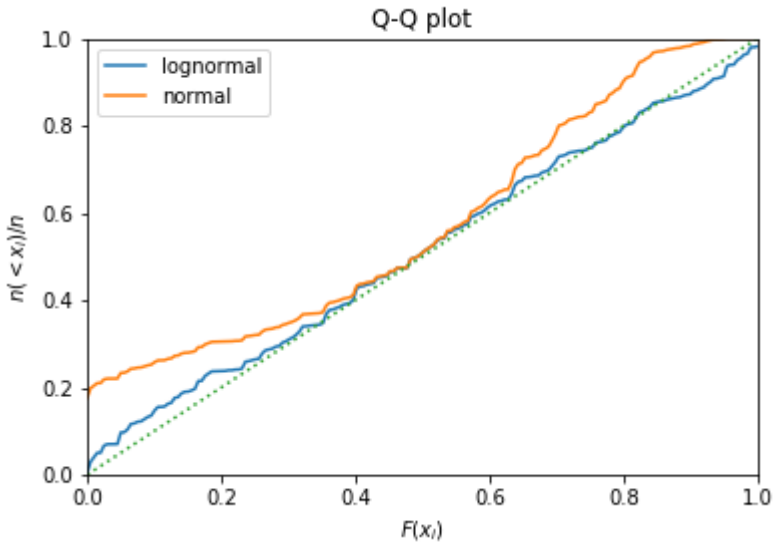

Figure 9.1: Quantile-Quantile plot. The data was drawn from a lognormal distribution with $n = 200$. A normal and lognormal distribution are compared. You can see that the correct distribution gives a curve that is much closer to the $y = x$ line. For smaller sample size the agreement might not be so clear. The flaring of the normal curve is a reflection of the normal distribution predicting more points below the minimum data point ($F(x_1) \simeq 0.2$ for the smallest data point) and less for the larger values that are in the sample ($F(x)$ goes to one well before the largest data point).

of data points in each bin. Let's say there are n_i observations in bin i and there are N measurements in total. Our model predicts that the probability of a given measurement being in bin i is p_i. The numbers in each bin will be distributed according to the multinomial distribution (section **??**):

$$P(\{n_i\}|N,\{p_i\}) = \frac{N!}{\prod_{i=1}^{k} n_i!}\prod_{i=1}^{k} p_i^{n_i} \tag{9.2.1}$$

Under the assumption that the number of counts in each bin is large we can use Sterling's approximation to simplify the probability

$$\ln P(\{n_i\}|N,\{p_i\}) = \ln(N!) + \sum_i^k \left[n_i \ln(p_i) - \ln(n_i!)\right] \tag{9.2.2}$$

$$\simeq N\ln(N) - N + \sum_i^k \left[n_i \ln(p_i) - n_i \ln(n_i) + n_i\right] \tag{9.2.3}$$

$$= N\ln(N) + \sum_i^k \left[n_i \ln(p_i) - n_i \ln(n_i)\right] \qquad \sum_i n_i = N \tag{9.2.4}$$

Sterling's approximation has been used so we have made the assumption that there are a large number of counts in each bin.

The mean and variance of the number counts are

$$E[n_i] = Np_i \qquad Var[n_i] = Np_i(1-p_i) \tag{9.2.5}$$

We can expand the log probability around the average number counts using

$$\ln P(n_i = Np_i) = N\ln(N) + \sum_i^k \left[Np_i \ln(p_i) - Np_i \ln(Np_i)\right] \tag{9.2.6}$$

$$= 0 \qquad \sum_i p_i = 1 \tag{9.2.7}$$

$$\left[\frac{\partial}{\partial n_i}\ln P(\{n_i\})\right]_{n_i=Np_i} = \left[\ln(p_i) - \ln(n_i)\right]_{n_i=Np_i} = -\ln N \tag{9.2.8}$$

$$\left[\frac{\partial^2}{\partial n_i^2}\ln P(\{n_i\})\right]_{n_i=Np_i} = \left[-\frac{1}{n_i}\right]_{n_i=Np_i} = -\frac{1}{Np_i} \tag{9.2.9}$$

So the expansion is

$$\ln P(\{n_i\}) \simeq -\ln N \sum_i (n_i - Np_i) - \sum_i \frac{1}{2Np_i}(n_i - Np_i)^2 + \mathcal{O}\left[(n_i - Np_i)^3\right] \tag{9.2.10}$$

$$= -\sum_i \frac{1}{2Np_i}(n_i - Np_i)^2 + \mathcal{O}\left[(n_i - Np_i)^3\right] \text{ using } \sum_i n_i = N \ , \ \sum_i p_i = 1 \tag{9.2.11}$$

So we can approximate the distribution of each n_i as being Gaussian so

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \tag{9.2.12}$$

will be approximately χ^2_{k-1} distributed because the one constraint that $N = \sum_i n_i$.

So a χ^2 test can be used to see if a particular distribution is consistent with the counts. This approximation is only valid if all the n_i are large however. If this approximation is not valid you could use this statistic, but find its distribution using Monte Carlo as will be discussed in a later chapter. Note also that binning the data always results in a lose of information and possibly a dependence on the usually arbitrary choice of bin boundaries. Setting the bins based on a criterion derived from the data (for example equal number of observations in each bin or no less than 10 in a bin) can invalidate the significance of the test.

9.2.3 Kolmogorov-Smirnov test

It is generally bad practice to bin data if it can be avoided. The Kolmogorov-Smirnov (KS) test is another frequentist test that is used to test if the data came from a particular distribution or whether two data sets came from the same distribution. It does not require binning the data which is an advantage over the test just discussed. This test and the ones in the next to sections are often used to **test normality**, i.e. consistency with the data being normally distributed. More generally, they can be used to test consistency with any distribution including luminosity functions, energy distribution functions, black-body spectrum, etc.

Consider the **empirical distribution function**

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \Theta(x_i \leq x) \tag{9.2.13}$$

$\hat{F}(x)$ is an unbiased estimator of the cumulative distribution, $F(x)$ so in the limit of $n \to \infty$ we would expect $\hat{F}_n(x)$ to converge to the true cumulative distribution,

$F(x)$. For a finite number of points it will be a series of steps. It is generally better to base statistical arguments on this rather than the commonly used alternative, a histogram, because of this property and it being independent of binning.

The KS statistic is

$$D_n = \max_x |\hat{F}_n(x) - F(x)| \tag{9.2.14}$$

i.e. the largest vertical distance between the sample and cumulative distributions. [2] Kolmogorov and Smirnov found that the distribution of this statistic is

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq t) = H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2} \tag{9.2.15}$$

for large n. This is independent of the distribution that is being tested $F(x)$.

The KS test is widely used, but as we will see there are modifications of this test that are more sensitive for most distributions. The KS statistic takes the largest distance between $\hat{F}_n(x)$ and $F(x)$ and because of this it is not sensitive to differences in the tails of the distribution.

9.2.4 two sample KS test

The hypothesis here is that both data sets come from the same data distribution. Let's say the empirical cumulative distributions for these two samples are $\hat{F}_n(x)$ and $\hat{G}_m(x)$. The statistic is the maximum vertical distance between the two sample cumulative distributions

$$D_{mn} = \max|\hat{F}_n(x) - \hat{G}_m(x)| \tag{9.2.16}$$

This statistic is distributed like

$$\lim_{n\to\infty} P\left(\sqrt{\frac{mn}{m+n}}D_{mn} \leq t\right) = H(t) \tag{9.2.17}$$

for large n.

[2] Note on notation: In statistics literature you will sometimes see the operator $\sup_x$ instead of $\max_x$. The difference is subtle. sup stands for "supremum" which is the smallest number that is larger than the argument within the allowed range.

9.2.5 Cremér-von Mises test

Like the KS test this test seeks to test the consistency of data with a given distribution, but uses the statistic

$$T_{CM} = n \int_{-\infty}^{\infty} dF(x) \left[\hat{F}(x) - F(x)\right]^2 \tag{9.2.18}$$

$$= \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(x_i)\right)^2 \tag{9.2.19}$$

where $x_1, \ldots, x_n$ are the sorted data points. The significance can be looked up using any decent computer statistical packages. This can also be used to compare two distributions, with a different significance level.

9.2.6 Anderson-Darling test

Another variation on this test that is usually more sensitive than either the KS or the Cremér-von Mises test is the Anderson-Darling test which uses the statistic

$$A^2_{AD} = n \int_{-\infty}^{\infty} dF(x) \frac{\left[\hat{F}_n(x) - F(x)\right]^2}{F(x)\left[1 - F(x)\right]} \tag{9.2.20}$$

$$= -n - \sum_{i}^{n} \frac{2i-1}{n} \left[\ln\left(F(x_i)\right) - \ln\left(1 - F(x_{n+1-i})\right)\right] \tag{9.2.21}$$

Again you need to look up the significance of the statistic using a software package,

Note that this test and the ones before compare one specific distribution to the data. They do not compare a family of distributions to the data. For example, as stated, they do not test if your data is consistent with any normal distribution, just the normal distribution you test with a fixed mean and variance. But they can be modified to test for consistency with a family of distributions. This can be done by fitting for the best parameters and then using Monte Carlo or bootstrap sampling to find the significance of the statistic (see section 9.3). Alternatively, a modification of the Anderson-Darling test exist for normality in general.

These statistics can also be interpreted as measures of the "distance" between two distributions. This is a concept we will return to in a later chapter.

9.3 Goodness-of-fit revisited

We have seen that for a linear model with normally distributed data we can find how well the model fits the data in a global sense using the χ^2 distribution (section 8.5 and

8.6). $X^2(D, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ are the best fit parameters is $\sim \chi^2_{N-k}$ distributed. The reduction in the number a degrees of freedom comes from the fact that we used the data to find the best fit parameters. For some other statistic or model we cannot in general make the split of the data space into subspaces that are dependent and independent of the parameters. We can generally calculate by Monte Carlo the distribution of a statistic for *specific parameter values*. For example, we can calculate the KS statistic for a Gaussian with mean and variance specified (μ and σ). This would not be a global test of Gaussianity however. In other words, we might ask if our statistical model is consistent with the model at all, irrespective of the best fit parameters when the model is not linear and/or the distribution is not normal. (People often quote the reduced χ^2 for the best fit model even when the model is not linear and/or the data is not normally distributed. Strictly speaking this is not correct although if the model can be safely expanded to linear order in the parameters around its best fit value it could be a good approximation.)

Fortunately there exists a simple Monte Carlo procedure for approximating the significance of these statistics for a whole family of models. The procedure goes like this:

1. Make a fake data set D^* with model $\hat{\boldsymbol{\theta}}$.
2. Find the best fit parameters for this data set $\hat{\boldsymbol{\theta}}^*$.
3. Calculate the goodness-of-fit statistic $\hat{A}^* = A(D^*; \hat{\boldsymbol{\theta}}^*)$
4. Repeat 1-3
5. Compare the cumulative distribution of $\hat{A}^*$ to the observed $\hat{A}$ to find its significance.

It can be shown that under very general conditions the limiting distribution ($n \to \infty$) of $\hat{A}^*$ is the same as $\hat{A}$ (Babu & Rao, 2004). There also exists bootstrap methods of doing this calculation.

9.4 Rank statistics

In many of the statistics we have talked about so far we have had to assume the data was Gaussian distributed or hope that this is a good approximation in some limit. Rank statistics avoid any requirement on how the data is distributed accept that data points must be independent. They do this by using the rank of the data rather than the values directly. If the data is sorted from least value to largest value the **rank** of a data point is where it appears in this list. In other words if a data point x_i has a rank

X_i there are $X_i - 1$ data points with smaller values. The advantage of use the rank is that we already know its distribution without knowing the underlying distribution of the data values x_i (assuming they are independent). X_i for a random data point has equal probability of being any number between 1 and n, the number of data points. Because statistics based on the rank do not depend on normality the are known as more **robust** than those that are dependent on normality (more on this later). They might not be as efficient (have smaller variance for the same amount of data) when the data is Gaussian distributed, but they wont go catastrophically wrong when the data is not Gaussian distributed.

9.4.1 Spearman's correlation coefficients

Here we revisit the problem of determining whether two sets of data points, x_i and y_i are correlated. We already met **Pearson's correlation coefficient**

$$r_{xy} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{9.4.1}$$

Spearman's correlation coefficient is the same thing, but using the ranks, X_i instead of the values x_i,

$$r_s = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i^n (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}} \tag{9.4.2}$$

This can be simplified by taking into account that the mean and variance of the ranks are always that same. Using these well know sums

$$\overline{X} = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} i = \frac{1}{2} n(n+1) \tag{9.4.3}$$

and

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} i^2 = \frac{1}{6} n(n+1)(2n+1) \tag{9.4.4}$$

The variance of both X_i and Y_i are

$$V_X = X_Y = \frac{1}{n} \sum_i (X_i - \overline{X})^2 = \frac{(n^2 - 1)}{12} \tag{9.4.5}$$

After some algebra

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_i (X_i - Y_i)^2 \tag{9.4.6}$$

Problem 40. *Show that (9.4.6) is true.*

Spearman's statistic is usually used with the null hypothesis that the variables are uncorrelated, i.e. to show that this hypothesis can be ruled out. To do this we will need the p-value or cumulative distribution of r_s when there are no correlations. The average $\langle r_s \rangle = 0$. This should be clear from the original definition (9.4.1). There is no exact analytic calculations for the distribution of r_s as far as I am aware, but you can find the exact significance of a value of r_s by a **permutation test**. If we sort the x_i's and there are no correlations then any order of the y_i's should be equally probably. We can calculate r_s for every possible permutation of the X_i's and see how many of those permutations have an r_s larger then the measured one. The number of permutations is of course $n!$ so this can get computationally expensive for large n. Figure 9.2 shows the results of this calculation for several values of n.

The permutation test, or a variation on it, is an option for calculating the significance of a statistic when all possible outcomes of the experiment given the null hypothesis are equally likely, discrete and finite in number (or more practically, small enough in number to be calculated). We encountered a similar situation in section 7.4.1 when we discussed bootstrap resampling. There is a finite number of bootstrap samples so if the number of data points is small these can all be calculated. This is not usually the case and sampling from them randomly is usually done to approximate the complete sum over boot strap samples. The same could be done here if the number is large. However in this case there exists some approximate solutions for large n.

It can be shown that

$$z = \sqrt{\frac{n-3}{1.06}} \operatorname{arctanh}(r_s) \tag{9.4.7}$$

is approximately $\mathcal{N}(0,1)$ distributed. This is know as the Fisher's z-transformation. It is also true that

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \tag{9.4.8}$$

is approximately t-distributed with $n-2$ degrees of freedom.

One disadvantage of using r_s is that it is a biased estimator for the true correlation ρ when it is not zero.

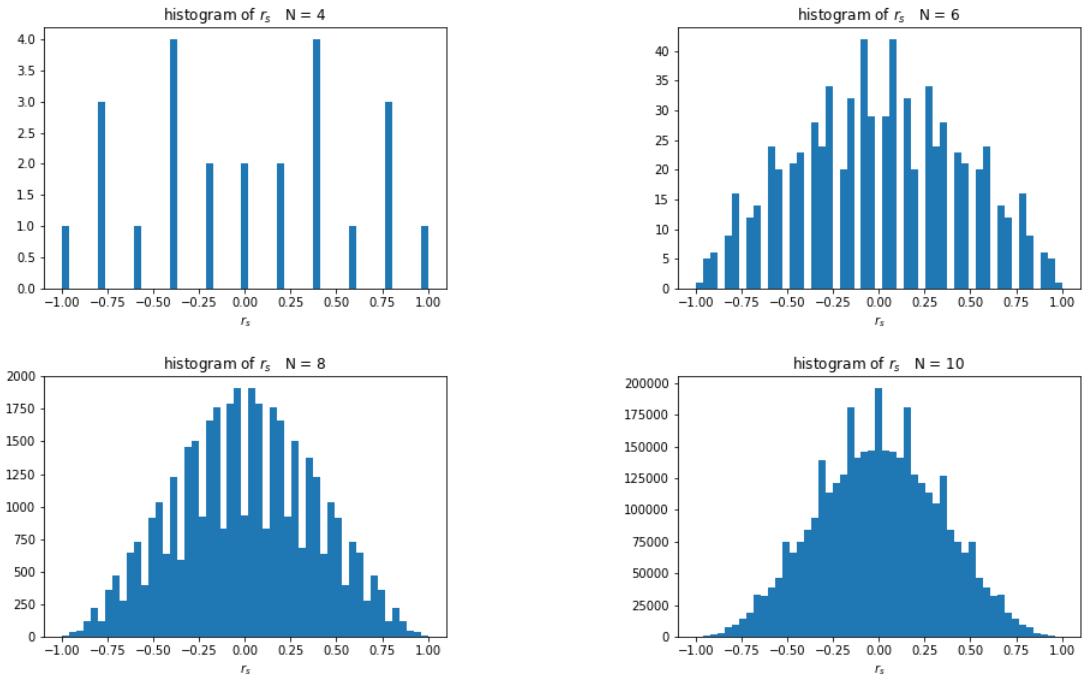

Figure 9.2: The distribution of Spearman's correlation coefficient, r_s, calculated under the hypothesis that there are no correlations by using all permutations of the ranks for one of the variables.

9.4.2 Kendall's correlation coefficient

Another rank statistic that is used for detecting correlations is Kendall's τ. Let's say we sort the data points x_i so that there ranks are $X_i = \{1, 2, \ldots, n\}$. If the variables are perfectly correlated then Y_i will be the same. If they are perfectly anti-correlated then $Y_i = \{n, n-1, \ldots, 1\}$. Let us define Q as the number of pairs of Y_i's that are out of order, the number of inversions. In other words if

$$h_{ij} = \begin{cases} 1 & Y_i > Y_j \\ 0 & \text{otherwise} \end{cases} \tag{9.4.9}$$

then

$$Q = \sum_{i<j} h_{ij} \tag{9.4.10}$$

So for $Y_i = \{1, 9, 6, 7, 5\}$ $Q = 5$. Kendall's correlation coefficient is

$$t = 1 - \frac{4Q}{n(n-1)} \tag{9.4.11}$$

For $Q = 0$, perfect correlation, $t = 1$ and for perfect anti-correlation $Q = n(n-1)/2$ and $t = -1$ and, as we expect for a correlation coefficient, the expectation value for uncorrelated data is $\langle t \rangle = 0$.

For the null hypothesis that the variables are not correlated we can calculate the distribution by calculating it for all permutations of Y_i as we did for Spearman's r_s. This calculation for some small values of n is displayed in figure 9.3. You can see from this plot that τ approaches normality more quickly with increasing n than r_s does. It also has a smaller variance. t is essentially normally distributed for $n > 10$ with a variance of

$$\sigma_t^2 = \frac{2(2n+5)}{9n(n-1)} \tag{9.4.12}$$

Another advantage of t is that it is an unbiased estimator of τ, the population statistic. r_s and t are closely related. In fact it is possible to show that

$$\langle r_s \rangle = \rho_s + \frac{3}{N+1}(\tau - \rho_s) \tag{9.4.13}$$

A disadvantage is that a particular value for τ might be hard to interpret. For a multivariant Gaussian it can be shown that

$$\tau = \frac{2}{\pi} \arcsin(\rho_{xy}) \tag{9.4.14}$$

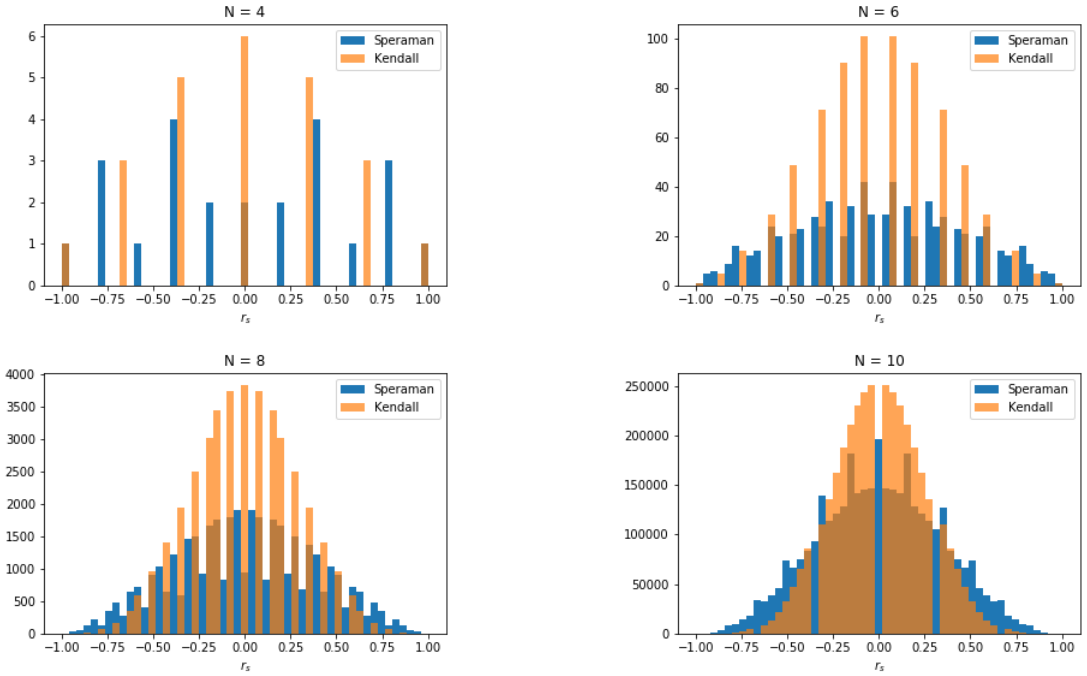

Figure 9.3: The distribution of Kendall and Spearman's correlation coefficients calculated under the hypothesis that there are no correlations by using all permutations of the ranks for one of the variables.

9.4.3 Wilcoxon's U test

Wilcoxon's U test (also called the **Mann-Whitney test**, **Wilcoxon-Mann-Whitney test** or **rank-sum test**) is a test for the equality of the means of two samples. In section 8.1 we saw a test for the difference of the means that relied on the underlying distributions being Gaussian. We can avoid this assumption without loosing much in efficiency by constructing a statistic out of ranks.

We have two samples x_i and y_i. Our hypothesis is that they come from the same distribution. We can put them together into one sample z_i and sort them. If they are taken from the same distribution we would expect the x_i's to appear randomly in the list, i.e. the ranks of one data set should be uniformly distributed. Several equivalent statics are used for this. There is simply the sum of the ranks

$$W = \sum X_i \tag{9.4.15}$$

where where X_i is the rank for the combined sample X and Y. It is also common to use

$$U = \sum_i^n X_i - \frac{1}{2} n_x(n_x + 1) \tag{9.4.16}$$

U is always between 0 and $n_x n_y$. The significance of this statistic can again be calculated by calculating it for all permutations of the ranks or, for larger n, by Monte Carlo, but it actually becomes quite close to normally distributed for only $n_x, n_y \gtrsim 8$ with a mean and variance

$$\langle U \rangle = \frac{n_x n_y}{2} \tag{9.4.17}$$

$$\sigma_U^2 = \frac{1}{12} n_x n_y (n_x + n_y + 1) \tag{9.4.18}$$

Problem 41. *Show that U is in the range $[0, n_x n_y]$.*

9.5 Bias and Statistics

In the Bayesian method we find the posterior for the parameters given the data. We can summarize this distribution by finding its mean, mode, variance, etc. These are statistics of the parameters although the probability distribution contains only the one data set that was observed. We are not concerned with repeated trials or the limit with an infinite amount of data.

In the frequentist approach a statistic is formed from the data. Sometimes this statistic is meant to be an estimate of a parameter in the model. In this case it is an **estimator**. We don't expect this estimator to equal the true value for every data set. If the average of this estimator, over all possible data sets of the same size, is not equal to the true value, the estimator is **biased**. If we increase the amount of data this bias will become smaller if our estimator is a good one. If the bias goes to zero for an infinitely large data set then we say it is **asymptotically unbiased**.

For our linear model the MLE is of courses linear in the data so, if the model is in fact the correct one, the MLE will be unbiased in this case. If the model is not linear or the true model contains more or less parameters then the model being fit, the parameter might be biased.

To illustrate these concept, let's say we have a model, $f(\boldsymbol{\theta}) = \boldsymbol{y}$, which relates some parameters $\boldsymbol{\theta}$ to some measurable quantities $\boldsymbol{y}$. Now through some theoretical ingenuity you are able to invert the model to get $f^{-1}(\boldsymbol{y}) = \boldsymbol{\theta}$. You might think that the best choice for an estimator would be $\tilde{\boldsymbol{\theta}} = f^{-1}(\boldsymbol{d})$ where $\boldsymbol{d}$ are the measured values of the $\boldsymbol{y}$ observables. But the data has noise in it so if f is not linear and the noise, $\boldsymbol{n}$ is additive

$$\left\langle \tilde{\boldsymbol{\theta}} \right\rangle = \left\langle f^{-1}(\boldsymbol{d}) \right\rangle = \left\langle f^{-1}(\boldsymbol{y} + \boldsymbol{n}) \right\rangle \neq \left\langle f^{-1}(\boldsymbol{y}) \right\rangle \tag{9.5.1}$$

even if $\langle \boldsymbol{n} \rangle = 0$. The estimator $\tilde{\boldsymbol{\theta}}$ is biased.

A simple example, let's say we want to measure the kth power of y. The estimator $\tilde{\theta}_k = d^k$ would have an average of

$$\left\langle \tilde{\theta}_k \right\rangle = \left\langle (y+n)^k \right\rangle = \sum_{i=0}^{n} \binom{n}{i} y^i \left\langle n^{k-i} \right\rangle \tag{9.5.2}$$

For $k > 2$ this would be a rather bad estimator.

Chapter 10

Bayesian model selection & model checking

In chapter 5 we considered Bayesian inference or parameter fitting. In section7.1 we encountered the problem of determining how many and which parameters are needed in a regression model. We addressed the problem with k-fold cross-validation and bootstrap resampling. Now we will look at how the Bayesian framework can be used to address this problem.

Let's say there are competing models that describe the data, but these models do not just differ from each other by having different values for their parameters. The models might have completely different parameters or one model might be the same as the other except that it has additional parameters. Which model is more strongly supported by the data? Is it justified to add the extra parameters? This is called model selection.

Let us consider a set of all possible models that explain the data $M_1, M_2, ...$ We can write down the posterior for model M_i using Bayes' theorem as in the parameter estimation case

$$P(M_i|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|M_i)P(M_i)}{P(\boldsymbol{D})} = \frac{P(\boldsymbol{D}|M_i)P(M_i)}{\sum_i P(\boldsymbol{D}|M_i)P(M_i)}. \tag{10.0.1}$$

It is difficult to imagine ever knowing *all* possible models so model selection is usually restricted to comparing the relative probability of two models, call them M_1 and M_2, by taking the ratio of their posteriors

$$O_{1,2} = \frac{P(M_1|\boldsymbol{D})}{P(M_2|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|M_1)}{P(\boldsymbol{D}|M_2)}\frac{P(M_1)}{P(M_2)} = B_{1,2}\frac{P(M_1)}{P(M_2)}. \tag{10.0.2}$$

$O_{1,2}$ is called the **odds** of model 1 relative to model 2 and $B_{1,2}$, the ratio of the model likelihoods, is know as **Bayes's factor**. If the prior probabilities are equal, as they

often are, then the odds is equal to Bayes' factor. Note that $P(\boldsymbol{D})$ cancels out so we avoid needing to know the probability of the data over all possible models. If the odds is large then model 1 is favored. If it is small then model 2 is favored. You can also take the log of the odds and then positive values would favor M_1 and negative, M_2.

How can we calculate $P(\boldsymbol{D}|M)$? In the parameter estimation problem we stayed within one model, or you could call it a family of models if each set of parameters counts as a different model. Because of this all the probabilities were conditional on this model being true although that was not explicitly shown. We can write Bayes' theorem again with the model conditionality explicitly shown

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)}{P(\boldsymbol{D}|M)}. \tag{10.0.3}$$

We can now see that $P(\boldsymbol{D}|M)$ in the odds (10.0.2) is actually the evidence for each model,

$$P(\boldsymbol{D}|M_i) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ P(\boldsymbol{D}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i) = \mathcal{E}(\boldsymbol{D}|M_i) \tag{10.0.4}$$

where the integral is over all of parameter space within model M_i. Bayes' factor is the ratio of evidences for two models.

The situation often arises where one has a standard model that explains the data and an extension to the model that includes some additional parameters. We will call these **nested models**. For example, the standard ΛCDM cosmological model and ΛCDM plus dark energy with an equation of state parameter that is not -1 ($w \equiv p/\rho \neq -1$) as it would be for a cosmological constant. Or the dark energy might be coupled to dark matter and there is a parameter describing the strength of this coupling. Or you have stellar evolution models that predicts the amount of lithium in a low mass star among other things. The standard model has no mixing in the atmosphere and the extended model has mixing regulated with an additional parameter.

For nested models, the extended model will always have a set of parameter values that fit the data as well as or better than the standard model since the standard model is the extended model with additional degrees of freedom to fit the data. Usually the standard model is identical to the extended model with the additional parameters fixed to some value (perhaps 0 or in the dark energy case $w = -1$). Let's label the likelihoods $\mathcal{L}_{st}(\boldsymbol{\theta}|\boldsymbol{D})$ for the standard model and $\mathcal{L}_{ex}(\boldsymbol{\theta}, \beta|\boldsymbol{D})$ for the extended model where β is the extra parameter. Let's denote the parameter values that maximize the standard model likelihood as $\hat{\boldsymbol{\theta}}_{st}$ and those that maximize the extended model likelihood as $(\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta})$. Then

$$\mathcal{L}_{ex}(\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta}) \geq \mathcal{L}_{st}(\hat{\boldsymbol{\theta}}_{st}). \tag{10.0.5}$$

Because of this one might be drawn to the conclusion that more complicated models are always as good as or better than less complicated ones. But, as discussed in section 7.1, the model might be over fit. This can also be said to violates Occam's principle, or razor, that the best model is the simplest one that is consistent with the observations (William of Ackham $\sim$ 1300).

For a more concrete example, you can always fit a line to two data points perfectly. If you add another data point the line generally wont go through all the points. You could add a parameter and fit a quadratic function and it would again go through all of the points. If you have n data points you can fit them perfectly with a nth order polynomial (as long as they all have different independent variable values). But if your model includes random noise in the data you would not expect the correct model to go through all the points. "Any theory that fits all the data is wrong, because some of the data is wrong." So when do you stop adding parameters? When does the model fit too well?

Although it is not immediately apparent, Bayesian model selection automatically incorporates a version of Occam's razor, but it is not in the form that one might expect. To demonstrate this let's consider an extended model with on extra parameters $\boldsymbol{\beta}$. The prior on this parameter will be $\pi(\boldsymbol{\beta})$. The standard model will be the extended one with $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. We will take the priors on the models to be equal ($P(M_1) = P(M_2)$). Bayes' factor between the models is

$$B_{2,1} = \frac{\int d\boldsymbol{\theta} \int d\beta \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\beta})\pi(\boldsymbol{\theta},\boldsymbol{\beta})}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\beta}_o)\pi(\boldsymbol{\theta})} = \frac{\langle \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\beta})\rangle^{\pi}_{\boldsymbol{\theta},\boldsymbol{\beta}}}{\langle \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\beta}_o)\rangle^{\pi}_{\boldsymbol{\theta}}} \tag{10.0.6}$$

where $\langle \ldots \rangle^{\pi}_{\boldsymbol{\theta}}$ denotes the average or expectation with respect to the prior on parameters $\boldsymbol{\theta}$. This shows us that for the extended model to be favored the average of the likelihood in the extended parameter space must be larger than its average in the standard parameter space. The extended space is larger and thus even if the maximum of the likelihood in this space is larger, it does not follow that the average will be larger.

To understand this a little better let's define the volume to which the likelihood alone constrains the parameters as

$$\mathcal{V}^{\mathcal{L}}_{\boldsymbol{\theta}} \equiv \frac{1}{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}})} \int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}). \tag{10.0.7}$$

Likewise we can define the volume to which the prior by itself constrains the parameters as $\mathcal{V}^{\pi}_{\boldsymbol{\theta}}$. In the case of a uniform prior $\pi(\boldsymbol{\theta}) = 1/\mathcal{V}^{\pi}_{\boldsymbol{\theta}}$. Now if the prior is approximately constant within the likelihood volume, as is the case when the prior is uniform and the likelihood is small at the boundaries of the prior, we can take the

prior out of the integrals in (10.0.6) and express Bayes' factor as

$$B_{2,1} = \frac{\left[\mathcal{V}^{\mathcal{L}}_{\boldsymbol{\theta},\boldsymbol{\beta}}/\mathcal{V}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\beta}}\right]}{\left[\mathcal{V}^{\mathcal{L}}_{\boldsymbol{\theta}}/\mathcal{V}^{\pi}_{\boldsymbol{\theta}}\right]} \frac{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ext}, \hat{\boldsymbol{\beta}}_{ext})}{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{st}, \boldsymbol{\beta}_o)} \tag{10.0.8}$$

The ratios of maximum likelihoods on the left will favor model 2 or be at least neutral, equal to 1. The factor in front shows that an extended model that expands the likelihood volume compared the prior volume will be favored. Counterintuitively, an extend model whose likelihood constrains the parameters to a small region, even if that region is far away from the original best fit region at $(\hat{\boldsymbol{\theta}}_{st}, \boldsymbol{\beta}_o)$, will be more favored than one that is less constraining. This is essentially because the prior defines our expectations for what range of values the new parameter should have and the extended model is penalized for contradicting that by restricting the range. The extended model must not only provide an island in parameters space where the fit is better. Its total statistical weight, integrated over parameters space must be larger.

In 10.0.8, the ratio of the volumes in square brackets, or something like it, is sometimes called **Occam's factor**. It can be interpreted as a measure of the width in parameter space of the posterior in the β dimension in the extended model compared to the width allowed by the prior. For the odds to favor the extended model the fit must not just be better, but so much better that it overpowers Occam's factor to make the odds greater than 1. In other words, if model M_1 unnecessarily restrictive on parameters β by setting it to β_o it is better to use model M_2 which allows it to be free.

Another illustrative extreme case is one where the prior on β is very narrow compared to its constraint from the likelihood. This could be the case if a previous experiment already constrained β much more strongly than the one we are considering here or it could be that the theory behind the model requires that this parameter be within a range within which it cannot significantly change the predictions for this data set. In this case

$$O_{2,1} = B_{2,1} = \frac{\int d\boldsymbol{\theta} \int d\beta \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta)\pi(\boldsymbol{\theta})\pi(\beta)}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o)\pi(\boldsymbol{\theta})} \tag{10.0.9}$$

$$\simeq \frac{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o)\pi(\boldsymbol{\theta}) \int d\beta \ \pi(\beta)}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o)\pi(\boldsymbol{\theta})} \tag{10.0.10}$$

$$\simeq \frac{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o)\pi(\boldsymbol{\theta})}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o)\pi(\boldsymbol{\theta})} \tag{10.0.11}$$

$$\simeq 1 \tag{10.0.12}$$

So if the extended model has a parameter that doesn't improve the fit to the data within its prior allowed range then this extended model will not be favored or disfavored over the simpler model. For this reason saying that Bayesian model selection

accounts for Occam's razor is misleading. Occam's principle is that a simpler model should be favored, but here we see that a model with extra superfluous, irrelevant parameters is not disfavored. This is what we want however. We can always add extra irrelevant parameters to a model that have no effect on its predictions. These models are identical in terms of their physical predictions so the data should not favor one over the other even if philosophers do.

It could be that a previous experiment justified the use of M_2 and thus by necessity constrained β. If this constraint is included in the prior for the current experiment which provide no further information about β the odds between these models will be one. The current experiment adds no more weight to our preference for model M_2 and detracts none. Note that it does not follow that the current experiment would not favor M_2 if a different prior were used.

One criticism of Bayesian model selection is that it depends on having a well justified prior distribution for the parameters. Normalization or boundaries of allowed parameter space are important whereas in the parameter estimation case normalization of the prior cancels out and the boundaries are only important if the likelihood is significant there. For this reason you can use a uniform or Jeffreys prior, for example, that extends to infinity . If you use an infinite uniform prior for a new parameter in the model selection problem you will always get an infinitely small odds! If you start from a state of ignorance what prior do you use? If you extent the prior to just some big number then the odds will depend on this sometimes arbitrary choice. For me this is a big ambiguity in applying Bayesian model selection to practical problems unless there is a well justified prior coming from theory (like in the case $0 < \Omega < 1$) or from a previous experiment.

Consider the following situation. The prior is uniform within a hypercube and the likelihood constrains the parameters completely so that it is effectively zero on all the boundaries of the hypercube and outside of it. The extended model adds another dimension to the hypercube for which the same is true. Now we calculate Bayes' factor between the models and decide if the extended model is justified. We then decide that we were too conservative and extend the range of the new parameter's prior to be twice as large. Nothing has changed about the data, the posteriors or the likelihoods, but Bayes' factor will be half of what it was before. The parameters have zero probability of being in this new volume, but yet its existence should change our judgment about the relative merits of these two models? Personally, I find this to be a problem that makes Bayesian model selection unsatisfactory.

We will find later that frequentist hypothesis testing offers an alternative to model selection that is more satisfying in many ways. There is also a method called posterior predictive p-values which combines some of the advantages of both methods. We will learn about this method in section 10.3.

10.1 Linear Guassian models

The case of a linear model with a Gaussian likelihood is particularly instructive and has a clear connection to χ^2 hypothesis testing discussed previously. As we saw in chapter 8.4, we can write the likelihood as

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{M}\boldsymbol{\theta}\right)^T \boldsymbol{C}^{-1}\left(\boldsymbol{x}-\boldsymbol{M}\boldsymbol{\theta}\right)\right] \tag{10.1.1}$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}}\right)^T \boldsymbol{C}^{-1}\left(\boldsymbol{x}-\boldsymbol{M}\hat{\boldsymbol{\theta}}\right)\right] \tag{10.1.2}$$

$$\times \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right] \tag{10.1.3}$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}}\ e^{-\frac{1}{2}\chi^2_{min}}\ e^{-\frac{1}{2}\chi^2(\boldsymbol{\theta})} \tag{10.1.4}$$

where these two χ^2s are defined here. To find the evidence we integrate over the parameters.

$$\mathcal{E}(\boldsymbol{x}) = \int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{10.1.5}$$

$$= \frac{1}{V_\theta}\int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) \tag{10.1.6}$$

$$= \frac{(2\pi)^{p/2}|\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}|^{-1/2}}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}V_\theta}\ e^{-\frac{1}{2}\chi^2_{min}} \tag{10.1.7}$$

where in line (10.1.6) it was assumed that the prior, $\pi(\boldsymbol{\theta})$, is uniform and the likelihood is small at the boundaries of allowed parameter space. The volume of parameter space is V_θ and p is the number of parameters.

The parameters of model 1 will be $\boldsymbol{\theta}$ and for model 2, $\boldsymbol{\beta}$. Bayes' factor is then

$$B_{21} \equiv \frac{\mathcal{E}_2(\boldsymbol{x})}{\mathcal{E}_1(\boldsymbol{x})} = \left[\frac{V_\theta}{V_\beta}\right]\left[(2\pi)^{(p_2-p_1)}\frac{|\boldsymbol{M}_1^T\boldsymbol{C}^{-1}\boldsymbol{M}_1|}{|\boldsymbol{M}_2^T\boldsymbol{C}^{-1}\boldsymbol{M}_2|}\right]^{1/2}\ e^{\frac{1}{2}\Delta\chi^2_{min}} \tag{10.1.8}$$

where $\Delta\chi^2_{min} = \chi^2_{min,1} - \chi^2_{min,2}$. This lends itself to the following interpretation. The first factor is the ratio of volumes in parameter space. If model 1 is nested in model 2 with one additional parameter, as considered before, than this is one over the range of parameter β. The second factor is the ratio of likelihood volumes for the two models in parameter space and the last factor is the ratio of the maximum likelihoods. Recall that $|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$.

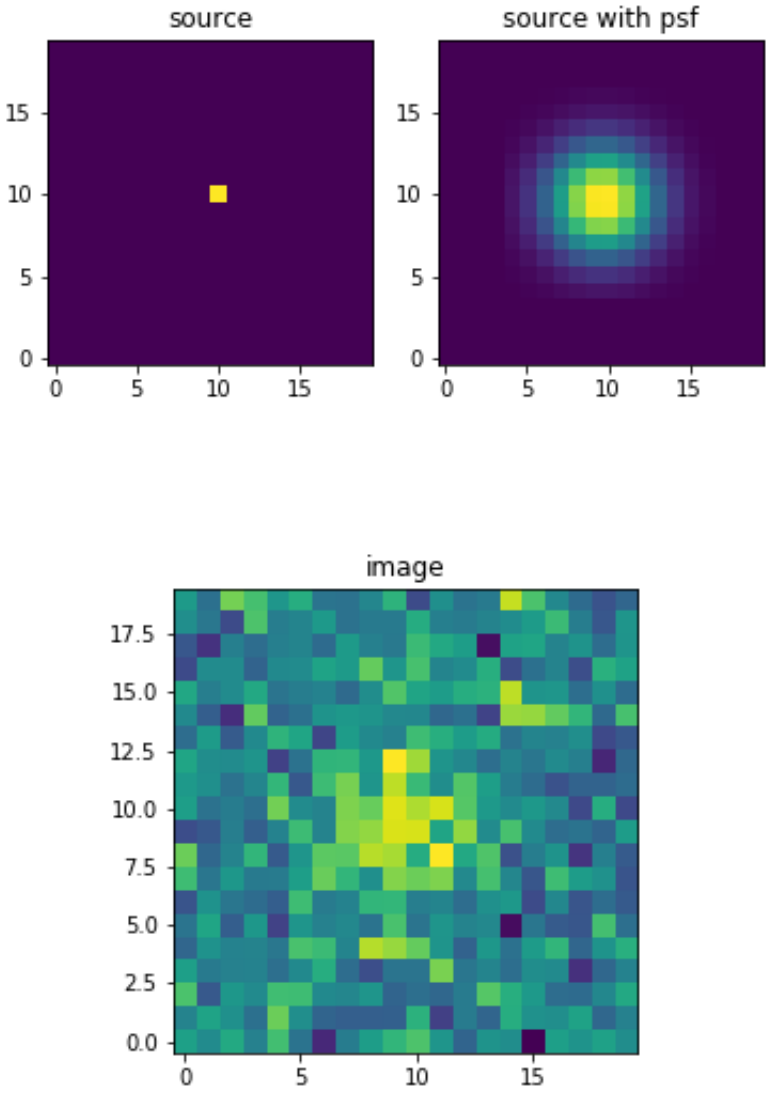

Figure 10.1: Simulated point source and point source convolved with psf. The source added to Gaussian uncorrelated noise.

10.1.1 Example: Object detection

Let us consider the case of an image made up of pixels. The noise in each pixel is independent with the same variance. We suspect there is an object at pixel j. The rest of the image is noise. There is a psf or blurring such that w_{ij} of the flux from position j goes into pixel i. There is also a uniform background.

Model 1 has no source at position j and model 2 has a source there. Model 1 has the parameter b for the background and model 2 has the additional parameter f_j for the flux from the source. The $\boldsymbol{M}$ matrices are

$$\boldsymbol{M}_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix} = \mathbf{1} \qquad \boldsymbol{M}_2 = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & w_{ij} \\ 1 & w_{(i+1)j} \\ \vdots & \vdots \end{pmatrix} = (\mathbf{1} \;\; \boldsymbol{w}) \tag{10.1.9}$$

Noise's covariance will be $C = \sigma^2 \boldsymbol{I}$ so,

$$\boldsymbol{M}_2^T \boldsymbol{C}^{-1} \boldsymbol{M}_2 = \frac{n}{\sigma^2} \qquad \boldsymbol{M}_2^T \boldsymbol{C}^{-1} \boldsymbol{M}_2 = \frac{1}{\sigma^2} \begin{pmatrix} n & \sum_i w_{ij} \\ \sum_i w_{ij} & \sum_i w_{ij}^2 \end{pmatrix}. \tag{10.1.10}$$

The psf will be normalized so that $\Sigma_i w_{ij} = 1$, otherwise it would leak flux. We will ignore leakage off the edge of the image. The maximum likelihood solution, equation (6.1.9), is

$$\begin{pmatrix} \hat{b} \\ \hat{f}_j \end{pmatrix} = \left(\boldsymbol{M}^T \boldsymbol{C}^{-1} \boldsymbol{M}\right)^{-1} \boldsymbol{M}^T \boldsymbol{C}^{-1} \boldsymbol{d} = \frac{1}{n \sum_i w_{ij}^2 - 1} \begin{pmatrix} \sum_j d_j \sum w_{ij}^2 - \sum_i d_i w_{ij} \\ n \sum_i d_i w_{ij} - \sum_i d_i \end{pmatrix}. \tag{10.1.11}$$

The posteriors are given by (6.1.11) with (10.1.10).

According to (10.1.8) Bayes' factor between these two models is

$$B_{21} \equiv \frac{\mathcal{E}_2(\boldsymbol{x})}{\mathcal{E}_1(\boldsymbol{x})} = \left[\frac{1}{\Delta f_j}\right] \left[\frac{2\pi n \sigma^2}{n \sum_i w_{ij}^2 - 1}\right]^{1/2} e^{\frac{1}{2}\Delta\chi^2_{min}}. \tag{10.1.12}$$

This does not take into account the requirement that f_j must be positive which could be taken care by changing the range of the f_j integral in () to 0 to ∞ instead of $-\infty$ to ∞. But considering (10.1.12) we can see the general behavior that Bayes' factor is dependent on the allowed range for the f_j, Δf_j. If we increase this range the evidence for there being a source goes down which doesn't seem desirable.

Another illustrative case this. Let us compare the models where there is no source to one where there is a source, but in this case the background, b, is known from some

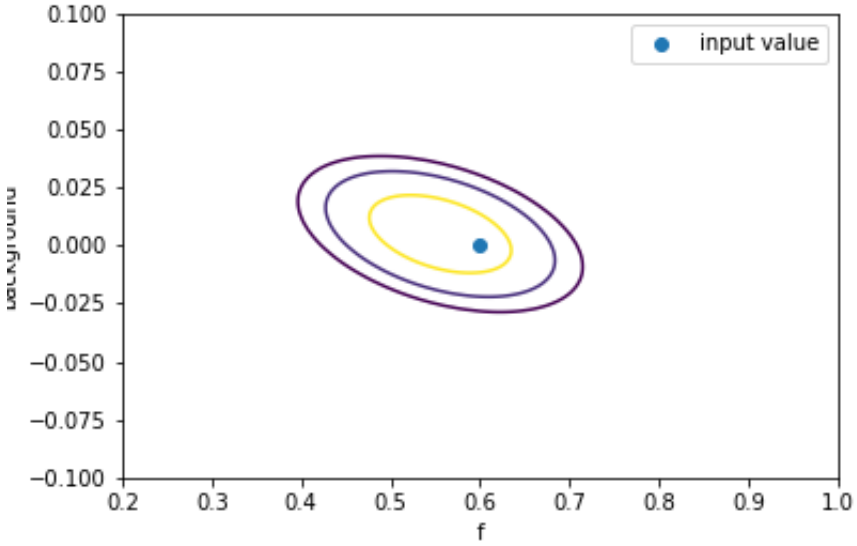

Figure 10.2: Posterior for the combination of parameters f and b. The contours contain 68%, 95% and 99% of the probability.

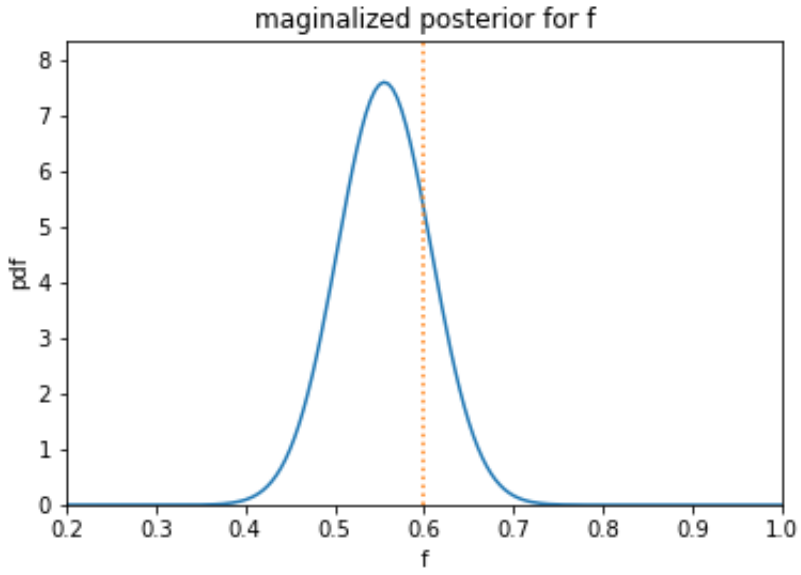

Figure 10.3: Marginalized posterior for parameter f. The dotted curve is the value used to generate the data.

previous calibration and has been subtracted. Model 1 has no free parameters and model 2 as one, f_j. The maximum likelihood solution for model 2 is

$$\hat{f}_j = \frac{\sum_i w_{ij} d_i}{\sum_i w_{ij}^2} \tag{10.1.13}$$

The evidences are

$$\mathcal{E}_1(f_j) = \frac{e^{-\frac{\sum_i d_i^2}{2\sigma^2}}}{(2\pi)^{n/2}\sigma^n} \quad \mathcal{E}_2(f_j) = \frac{e^{-\frac{\sum_i (d_i - w_{ij}\hat{f}_j)^2}{2\sigma^2}}}{(2\pi)^{n/2}\sigma^n} \int_0^\infty df_j \; e^{-\frac{\sum_i w_{ij}^2}{2\sigma^2}(f_j - \hat{f}_j)^2} \pi(f_j) \tag{10.1.14}$$

and Bayes' factor is

$$B_{21} = \frac{\mathcal{E}_2(f_j)}{\mathcal{E}_1(f_j)} = e^{\frac{1}{2}\Delta\chi^2_{min}} \int_0^\infty df_j \; e^{-\frac{\sum_i w_{ij}^2}{2\sigma^2}(f_j - \hat{f}_j)^2} \pi(f_j) \tag{10.1.15}$$

$$\Delta\chi^2_{min} = \frac{\sum_i d_i^2}{\sigma^2} - \frac{\sum_i (d_i - w_{ij}\hat{f}_j)^2}{\sigma^2}. \tag{10.1.16}$$

If we use a uniform prior up to $f_{\max}$ this becomes

$$B_{21} = \frac{\sigma}{f_{\max}} \sqrt{\frac{\pi}{2\sum_i w_{ij}^2}} \left[\mathrm{erf}\left(\sqrt{\frac{\sum_i w_{ij}^2}{2\sigma^2}} (f_{\max} - \hat{f}_j) \right) + \mathrm{erf}\left(\sqrt{\frac{\sum_i w_{ij}^2}{2\sigma^2}} \hat{f}_j \right) \right] e^{\frac{1}{2}\Delta\chi^2_{min}} \tag{10.1.17}$$

$$= \frac{\sigma}{f_{\max}} \sqrt{\frac{\pi}{2\sum_i w_{ij}^2}} \left[1 + \mathrm{erf}\left(\sqrt{\frac{\sum_i w_{ij}^2}{2\sigma^2}} \hat{f}_j \right) \right] e^{\frac{1}{2}\Delta\chi^2_{min}}. \qquad f_{\max} \gg \hat{f}_j \tag{10.1.18}$$

The $\Delta\chi^2_{min}$ factor will increase exponentially as the best fit flux, $\hat{f}_j$, gets larger, lending more support for model 2 and thus a detection. Not that the part in brackets would actually go down if $\hat{f}_j \gg f_{\max}$ which might seem strange. This is because the prior expectations are that a source should be no brighter than $f_{\max}$ so having a source brighter than this needs to be explained by a very unusually large contribution from the noise. However the $\Delta\chi^2_{min}$ factor will overpower it and increase B_{21} for very bright sources. This source might be unlikely in model 2, but it is even more unlikely in model 1.

Not liking the arbitrariness of $f_{\max}$, you might consider not using a uniform prior on f_i, but instead the expected luminosity function of sources as a prior. This might lead to some counter intuitive results. If a significant amount of the luminosity function is at $f_i \lesssim \sigma$ (as it almost always is for a power-law) than you would not

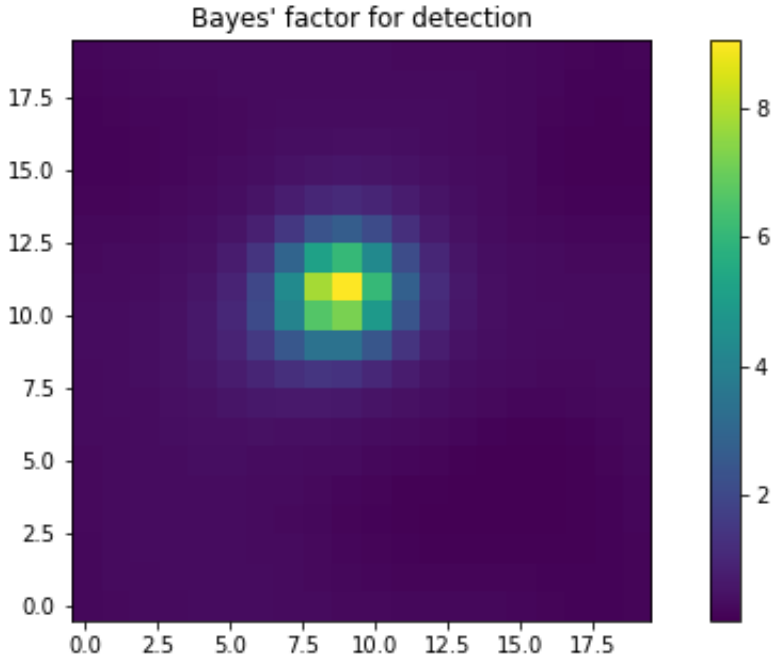

Figure 10.4: Bayes' factor for a source for each pixel of the image.

expect to detect a high proportion of the sources, but because, according to this prior, you expect most sources to produce no significant increase in the signal, Bayes' ratio would favor detection even when the data does not! This is because you are essentially testing whether a source that meets your prior expectations is present, but your expectation is that the source is unlikely to contribute to the data above noise levels. To see an easy illustration of this replace the prior with a Dirac delta function in (10.1.15), $\pi(f_i) = \delta(f_i)$. You will find that $B_{21} \geq 1$, but it will also be true that increasing $\hat{f}_j$ will increase B_{21}.

It is clear that if $\Delta\chi^2_{min}$ is large, model 2 should be favored. Bayesian model selection attempts to tell us precisely how big $\Delta\chi^2_{min}$ needs to be for the model to be accepted, but its answer is very strongly dependent on the prior distribution assigned to any new parameters. For me, the ambiguity as to a clear criterion for selecting a model and the dependence on the prior distribution even in regions of parameter space that are clearly ruled out by the data makes Bayesian model selection inherently suspect. We saw in chapter 8 that frequented hypothesis has a different answer to this question which might be more satisfactory, but is less widely applicable.

10.2 Ignore the prior & Bayesian Information Criterion (BIC)

A common alternative for model selection is to make the following "approximation". If we assume that the constraints on the parameters from the likelihood is much stronger than the constraint from the priors we can make the approximation

$$B_{2,1} = \frac{\int d\boldsymbol{\theta} \int d\beta \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta)\pi(\boldsymbol{\theta},\beta)}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)\pi(\boldsymbol{\theta})} \simeq \frac{\pi(\hat{\boldsymbol{\theta}}_{ext},\hat{\beta}_{ext})}{\pi(\hat{\boldsymbol{\theta}}_{st})} \frac{\int d\boldsymbol{\theta} d\boldsymbol{\beta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\beta})}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)} \tag{10.2.1}$$

and then the ratio of priors is simply dropped,

$$B_{2,1} \sim \frac{\int d\boldsymbol{\theta} d\boldsymbol{\beta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\boldsymbol{\beta})}{\int d\boldsymbol{\theta} \ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)}. \tag{10.2.2}$$

This does not make sense to me within the Bayesian framework. In the Bayesian context you can never integrate over the parameters without the prior because the prior actually defines the density of probability in the parameter space, it provides a metric on this space. However, outside of the Bayesian interpretation, the criterion might be justified in specific cases by simply showing that it works in simulations or analytically. This assumption is used to justify the BIC (Bayesian Information Criterion).

The BIC is an approximation that is sometimes used to make model selection much simpler. Consider the probability of the data given a model (eq. 10.0.4) again

$$P(\boldsymbol{D}|M_i) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M_i)\pi(\boldsymbol{\theta}|M_i) \tag{10.2.3}$$

$$= \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ e^{\ln \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M_i)}\pi(\boldsymbol{\theta}|M_i) \tag{10.2.4}$$

For clarity, let's suppress the dependents on the data and write $\mathcal{L}(\boldsymbol{\theta}, M_i)$. Now let's expand the the log likelihood around the maximum likelihood parameters for this model

$$\ln \mathcal{L}(\boldsymbol{\theta}, M_i) \simeq \ln \mathcal{L}(\hat{\boldsymbol{\theta}}, M_i) + \frac{1}{2}(\theta_i - \hat{\theta}_i)\left[\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\theta_j - \hat{\theta}_j) + \ldots \tag{10.2.5}$$

There is no linear term because we are expanding around the maximum likelihood. The matrix

$$\mathcal{I}_{ij} \equiv -\left[\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{10.2.6}$$

is sometimes called the information.[1] If there are n data points we can define $\bar{\mathcal{I}}_{ij} = \mathcal{I}_{ij}/n$ as the information per data point.

If we ignore higher order terms and we assume the prior $\pi(\boldsymbol{\theta}|M_i)$ is constant over the range of $\boldsymbol{\theta}$ where the likelihood is significant – i.e. the data constrains the model without any help from the prior – then we can approximate the integral

$$P(\boldsymbol{D}|M_i) \simeq \pi(\hat{\boldsymbol{\theta}}|M_i)\mathcal{L}(\hat{\boldsymbol{\theta}}, M_i)\frac{(2\pi)^{k/2}}{\sqrt{n^k|\bar{\mathcal{I}}|}} \tag{10.2.7}$$

where k is the number of parameters. In the limit of a large amount of data (compared to the number of parameters) we can ignore $|\bar{\mathcal{I}}|$ as it will be a factor of order 1 that will not change greatly between models. Likewise, we assume that the prior $\pi(\hat{\boldsymbol{\theta}}|M_i)$ does not favor any particular parameter set. The Bayesian Information Criterion (BIC) (Schwartz, 1978) for model M_i is defined as

$$\mathrm{BIC}_i \equiv k \ln n - 2 \ln\left[\mathcal{L}(\hat{\boldsymbol{\theta}}, M_i)\right] \tag{10.2.8}$$

so that

$$P(M_i|\boldsymbol{D}) \propto P(\boldsymbol{D}|M_i)P(M_i) \propto e^{-\mathrm{BIC}/2}P(M_i). \tag{10.2.9}$$

[1] This is not the Fisher information that will be discuss later because it is not averaged. It also should not be confused with Shannon's information.

The model with the smallest BIC is considered the best fit. You can see from its definition that more complex models are penalized by the $k \ln n$ term. This is the Occum's razer penalty function. If the BIC of two models differ by less than 2 there is considered to be no real reason to favor either one. If $|\Delta\text{BIC}| =$ 2-6 then there is some reason, 6 -10 is strong evidence and > 10 is considered very strong evidence that one is better than the other.

As a side note, there are other criterion used for the same purpose that differ from the BIC in that their penalty function is different and they are justified in different ways. For example the **Akaike information criterion** (AIC) is AIC $\equiv$ $2k - 2 \ln \mathcal{L}(\hat{\boldsymbol{\theta}}, M_i)$. Unlike the BIC, the AIC is not a consistent statistic, i.e. in the limit of infinite data there is a finite probability that the model with the lowest AIC will not be the correct one. A criterion based on **Minimum descriptive length** is also used. The interested reader might also look up the **Wald test** and the Lagrangian multiplier or **score test** for model selection problems.

Remember that we have made some approximations which might make the BIC invalid. In particular, if the number of independent data points compared to the number of parameters is not large the BIC will not be accurate. Also the BIC completely ignores the prior-volume for the parameters that we saw makes Bayesian model selection problematic. As a result the Bayesian Information criterion is not strictly speaking Bayesian. Also parameter degeneracies and unconstrained parameters will make $|\bar{\mathcal{I}}| = 0$ so k must be the number of non-degenerate, constrained parameters.

Problem 42. *Find the BIC for n independent identically normally distributed data points. The parameters are the mean and the variance.*

Problem 43. *Say we have records of the temperature in two cities in August that goes back centuries. Consider a model where the temperatures come in both cities are normally distributed with the same mean and variance. Now consider an alternative model where the mean and variance for the two cities are different. What criterion would the BIC give for favoring one of these models over the other?*

10.3 Bayesian model checking

To perform Bayesian model selection as discussed you need the evidence. You can hope to calculate the evidence analytically or numerically (see chapter 13) but, as we have seen, model selection by hypothesis testing has some real advantages, it is global and it is not sensitive to the prior. The parametric bootstrap method discussed in section 9.3 has some advantages, but its validity depends on asymptotic theorems which might not apply.

Hypothesis testing and model selection can be done using a very general hybrid method called **posterior predictive p-values** (PPP) (Protassov et al., 2002). The strategy is to use Bayesian prediction (section 6.6) based on the observed data to generate mock data sets that can be used to calculate the cumulative distribution of a goodness-of-fit statistic. The model is deemed inconsistent with the data if this predictive p-value is small. This is a kind of internal consistency check.

Let's look at the distribution of possible data for a given model. Using the product rule

$$p(\boldsymbol{x}) = \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}). \tag{10.3.1}$$

For the probability of the parameters, $p(\boldsymbol{\theta})$, we can use the posterior given the observed data, $\boldsymbol{d}$,

$$p(\boldsymbol{x}) = \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\boldsymbol{d})p(\boldsymbol{x}|\boldsymbol{\theta}). \tag{10.3.2}$$

For any statistic $T(\boldsymbol{x})$ we can use this to calculate its distribution

$$p(T) = \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\boldsymbol{d})p\left(T|\boldsymbol{\theta}\right) \tag{10.3.3}$$

or its cumulative distribution

$$F(T) = p(<T) = \int_{-\infty}^{T'} dT' \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\boldsymbol{d})p\left(T'|\boldsymbol{\theta}\right) \tag{10.3.4}$$

$$= \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\boldsymbol{d})p\left(<T|\boldsymbol{\theta}\right) \tag{10.3.5}$$

$$= \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\boldsymbol{d}) \int_{V(<T)} d\boldsymbol{x}\ p\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \tag{10.3.6}$$

where $V(<T)$ is the region in data-space where $T(\boldsymbol{x}) < T$. We can further write

$$p\left(<T|\boldsymbol{\theta}\right) = \int d\boldsymbol{x}\ \Theta\left(T > T(\boldsymbol{x})\right) p\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \tag{10.3.7}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} \Theta\left(T > T(\boldsymbol{x}_i)\right) \qquad \boldsymbol{x}_i \sim p\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \tag{10.3.8}$$

where the law of large numbers was used in the last line. The cumulative probability (10.3.6) can then be used to calculate the p-values. If the p-value is small, the observed data set is unlikely and the model can be rejected.

Unlike Bayesian inference, this can be used for global model checking or goodness-of-fit. The prior distribution does enter into it through the posterior, but as the amount of data increases it becomes independent of the prior which is a great advantage over using evidence ratios to do model selection. How powerful this test is depends on what statistic, $T(\boldsymbol{x})$, is used, as it is for traditional hypothesis testing. You are free to choose any statistic that you think will be sensitive to inconsistencies between the model and the data.

In simple cases equation (10.3.6) can be calculated analytically, but in many cases it cannot. It can be approximated however with the following steps

1. Generate parameters set $\boldsymbol{\theta}_i$, $i = 1 \dots N$ taken from the posterior $p(\boldsymbol{\theta}|\boldsymbol{d})$ using converged MCMC (chapter 13) or another technique.

2. For each parameter set $\boldsymbol{\theta}_i$ generate a data set x_i taken from the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$. This is often a Gaussian, Poisson or another classical distribution that can be easily sampled from.

3. Calculate the statistic $T_i = T(\boldsymbol{x}_i)$ for all the x_i's

4. Create the empirical cumulative distribution for the T_i's to evaluate the p-value. In other words, the estimated for the right-sided, single tail p-value is

$$p = \frac{1}{N} \sum_{i}^{N} \Theta \left(T_i > T(\boldsymbol{d}) \right) \tag{10.3.9}$$

Problem 44. *Show that for the case of normally distributed data, a linear model and $T = \chi^2_{min}(\boldsymbol{x})$, as defined in section 8.4, this is the same global goodness-of-fit test as discussed there.*

Chapter 11

categorical variables

So far we have dealt primarily with continuous variables. There are many important cases where the variable of interest is categorical, it takes on discrete values. These problems come up with sorting objects into classes and in trying to determine if being in a particular class is related to having a particular characteristic or not.

11.1 Contingency tables

Contingency tables are a way of examining the distributions of **categorical variables**, variables that can take one of a finite number of values. For example, dead or alive, class I or class II radio source, color of a car, etc. A typical null hypothesis might be, "Taking a particular medication does not change the risk of cancer?", "Smoking does not increase the chance of suicide" or "Elliptical galaxies are just as likely to have AGN as spiral galaxies?". A contingency table might look something like table 11.1. Usually, one is not interested in the distribution of one of the variables. We are not investigated in the distribution of star types for example only if the star type is related to whether it has a planet. In medical testing we would not be investigating how many people took the medication in so far as this is chosen by the researchers and not an outcome of the trials. A contingency table might also be used to test how well a test works. For example, a detection test or a medical test for a disease in which case the columns might be has the disease and doesn't have disease and the rows might be tested positive and tested negative.

Precisely how the significance of the contingency table is evaluated depends on how the experiment was designed and what null hypothesis is tested. Let us consider a generic 2x2 contingency table given in table 11.2.

In analyzing a contingency table the assumption is that the observations fall into the different combinations of characteristics randomly, but that the probabil-

star type	planets	no planets	total
F	10	134	144
G	15	97	112
K	2	30	32
	27	261	288

Table 11.1: Contingency table of completely made up data.

	type I	type II	row totals
category A	a	b	$r_1 = a + b$
category B	c	d	$r_2 = c + d$
column totals	$c_1 = a + c$	$c_2 = b + d$	$n = a + b + c + b$

Table 11.2: 2x2 Contingency table.

ities might not be equal. The total number of observations is fixed. In general, we can assign probabilities to each combinations of characteristics, p_a, p_b, p_c and p_d. Since all these must add up to one we can eliminate one of them. Let it be p_d. The correct distribution is the multinomial. For the 2x2 case this is

$$P(a, b, c|p_a, p_b, p_c, n) = \frac{n!}{a!b!c!(n-a-b-c)!}(p_a)^a(p_b)^b(p_c)^c(1-p_a-p_b-p_c)^{n-a-b-c} \tag{11.1.1}$$

where the p's are the probability for each case. There are three independent measurements since the sum of the observations is fixed at n.

One might encounter a problem where you want to determine all probabilities, but there is usually some relationship between the probabilities one wishes to test. The most common question is "Is the probability of being of type I dependent on the category or not?" This situation is the case for **Fisher's exact test** [1] which provides a clear hypothesis test for the equality of the distributions between columns. It is sometimes said that in this case the trials are continued until a predetermined number of outcomes are found. We keep searching for planets until we have found 10, for example. But in practice I think it is better stated that in this case the probability is contingent on both the number of instances in each row ($r_1 \dots$) and the number in each column ($c_1 \dots$). With these constraints there is only one independent measurement for the 2x2 case which can be taken to be a. The probability for the case were the columns are equally likely, the usual null hypothesis, is the hypergeometric

[1] Supposedly it was first used to test biologist Muriel Bristol's claim to be able to taste whether milk had been added to her cup before or after the tea. I don't know what the conclusion was.

distribution .

$$P(a|c_1, r_1, r_2, n) = \frac{\binom{r_1}{a}\binom{r_2}{c}}{\binom{n}{c_1}} = \frac{\binom{r_1}{a}\binom{r_2}{c_1-a}}{\binom{n}{c_1}} \tag{11.1.2}$$

a itself can be used as the goodness-of-fit statistic and this distribution used to evaluate its significance. You might remember that this is the probability of drawing a "balls" of type I in r_1 draws from a bag of n balls that has c_1 balls of type I and c_2 balls of type II, without replacement. You can see how this applies in this case.

Barnard's test and **Boschloos's test** are exact tests used to analyze contingency tables that are more powerful than Fisher's but require more computational work. These tests are available in statistical software packages. They work by explicitly finding all the possible tables, subject to the constraints, that fit the null hypothesis less well than the observed one.

When the number of rows and columns becomes larger it becomes harder to formulate a hypothesis test. When the numbers are large ($\gtrsim$ 10 in each bin) there is an alternative. The problem is very similar to the binned χ^2 test discussed in section 9.2.2 only in this case each of the rows are independent samples from a multinomial distribution and the null hypothesis is that each row has the same distribution. Let's generalize to an arbitrary number of rows and columns. The probability of getting n_{jk} events in the j-th row k-th column is

$$P = \prod_{j=1}^{r} P_{multinom}(n_{j1}, \ldots, n_{jc}|r_j, \{p_1, \ldots, p_c\}) \tag{11.1.3}$$

The null hypothesis requires that the probabilities p_k depend only on the columns. r_j is the total number of events in row j.

Following the same logic as in section 9.2.2 we can arrive at χ^2-like statistic

$$X^2 = \sum_{j=1}^{r}\sum_{k=1}^{c} \frac{(n_{jk} - r_j p_k)^2}{r_j p_k} \tag{11.1.4}$$

$r_j p_k$ is the expected number of events in row j column k given the null hypothesis. We don't know the best fit probabilities p_k a priori, but if there are a large number of events we can estimate them with

$$p_k \simeq \frac{\sum_{j=1}^{r} n_{jk}}{n} = \frac{c_k}{n} \tag{11.1.5}$$

where n is the total number of events and c_k is the total number of events in the k-th column. Putting these estimates into the X^2 statistic gives

$$X^2 = \sum_{j=1}^{r}\sum_{k=1}^{c} \frac{(n_{jk} - r_j c_k/n)^2}{r_j c_k/n} \tag{11.1.6}$$

This is the traditional statistic for contingency tables when the sample is sufficiently large.

There are r constraints for the row sums and c constraints for the column sums, but both of these sets must add up to the same number so in total there are $r+s-1$ constrains. There are $c\,r$ entries in a table so the number of degrees of freedom is $cr-(r+s-1)=(r-1)(c-1)$. So the above statistic is $\chi^2_{(r-1)(c-1)}$ distributed in the limit of many observations.

If it has been determined that the rows are consistent with being distributed in the same way one might want to know what the probability of being type I is. Going back the multinomial distribution, each row is an independent sample and so $p_a+p_b=1$ and $p_c+p_d=1$. The probability can the written

$$P\left(a,c|r_1,r_2,p_a,p_c,n\right)=\binom{r_1}{a}\binom{r_2}{c}(p_a)^a(1-p_a)^{r_1-a}(p_c)^c(1-p_c)^{r_2-c} \tag{11.1.7}$$

If the probabilities are the same for the rows, $p_a=p_c=p$ in which case

$$P\left(a,c|r_1,r_2,p,n\right)=\binom{r_1}{a}\binom{r_2}{c}p^{c_1}(1-p)^{n-c_1} \tag{11.1.8}$$

$$=\binom{n}{c_1}p^{c_1}(1-p)^{n-c_1} \tag{11.1.9}$$

We now have one unknown parameter, p, the probability of being of type I, and one independent measurements, c_1. The hypothesis that type I has the probability p can be done using c_1 as a statistic and this distribution to evaluate its significance. You could also use this likelihood to calculate the posterior for p.

If the hypothesis that each row is distributed identically is rejected one might want to constrain the parameters p_a and p_c. In this case the rows are completely separate so separate hypothesis tests or posteriors should be calculated for each row.

Problem 45. *Calculate X^2 for the contingency table 11.2. Is there evidence for the existence of planets being dependent on the type of star?*

11.2 logistic regression or classification

Regression with a categorical dependent or target variable is called logistic regression. The goal is to build a model that takes some continuous or categorical independent (or feature) variables and predicts which class the object or event it is in. The model is trained or fit using data where the independent (or target) variable or variables are known. In machine learning language this would be a supervised learning problem.

An astronomical example is trying to sort objects into stars, galaxies and quasars based on their photometry and/or images.

Let's say we are interested in predicting some outcome Y that can be true or false (win or lose, pass or doesn't pass, live or die). The probability of Y being true might be dependent on some independent variable or variables X. We wish to be able to predict the probability of Y given an X.

Each trial has a different probability of being correct, p_i, depending on the variable $\boldsymbol{x}_i$. For the two possibilities $y_i = 0$ or 1 the likelihood is binomial

$$\mathcal{L}(\{y_i\}) = \prod_{i=0}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{11.2.1}$$

The idea behind logistic regression is to make a model for the p_i's that depends on the $\boldsymbol{x}_i$'s and some parameters $\boldsymbol{\theta}$. The best fit parameters can be found by maximizing the likelihood. In this way instead of making a hard boundary in "feature space", $\boldsymbol{x}$, on one side of which we predict $y = 1$ and on the other $y = 0$ we instead predict the odds of y at some point $\boldsymbol{x}$. This allows for the possibility that different outcomes y could occur at the same $\boldsymbol{x}$. If y is dependent on $\boldsymbol{x}$ we would expect that in some region of $\boldsymbol{x}$-space the outcome will be close to certain, $y \sim 1$ or $y \sim 0$, and in other regions no definitive determination can be made, $y \sim 1/2$.

The most popular model is a linear one for the log of the odds $p_i/(1-p_i)$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \theta_o + \sum_j \theta_j x_j \tag{11.2.2}$$

The free parameters are the θ_j's. The sum is over all the independent variables. The odds implies that the probability is in the form of a **sigmoid function**, $S(u)$,

$$S(u) = \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u} \tag{11.2.3}$$

$S(u)$ is 1 for large u and 0 for large $-u$. The probability that the ith data point will be true is given by

$$p_i = S\left(\theta_o + \sum_j \theta_j x_j\right) \tag{11.2.4}$$

This is one example of an **activation function**.

This is plugged into the likelihood (11.2.1) for the training set. The maximum likelihood solution for $\boldsymbol{\theta}$ must be found by numerical means (usually Newton-Raphson or some variant) because no analytic solution exists.

11.2.1 multinomial logistic regression

What if there are more than 2 classes? Let's say there are K classes. The probability of being in class i is modeled as

$$p_i(\boldsymbol{x}) = \frac{1}{1+\sum_{k\neq p} e^{\boldsymbol{\theta}^k\cdot\boldsymbol{x}}} \times \begin{cases} e^{\boldsymbol{\theta}^i\cdot\boldsymbol{x}} & , \quad i \neq p \\ 1 & , \quad i = p \end{cases} \tag{11.2.5}$$

where $k = p$ is the "pivot class". You can see that this model for the probabilities is normalized by construction. There are now $(K-1)\times N_f$ coefficients $\boldsymbol{\theta}^k$ where N_f is the number of features. This function, called a **softmax** function, will ideally be close to one for one of the classes and close to zero for the others if the model is working well.

There are several strategies for finding the coefficients. One is simply to maximize the multinomial likelihood with respect to these coefficients by numerical means. Another is to find the coefficients $\boldsymbol{\theta}^k$ by doing a binomial logistic fit between class k and the pivot class p. This is repeated for all $K-1$ classes besides p.

The probability of an observed case being in class i is p_i so the likelihood for a single case is just p_i. The p_i's will be different in each case depending on the independent variable, or features, in each case, $p_i(\boldsymbol{x}_j)$. The total likelihood will be the product of the probabilities for each case

$$\mathcal{L} = \prod_{j=0}^{N} p_{i_j}(\boldsymbol{x}_j|\boldsymbol{\theta}^i) \tag{11.2.6}$$

where i_j is the class that case j is found to be in. This can also be written

$$\mathcal{L} = \prod_{j=0}^{N}\prod_{i=0}^{K} p_i(\boldsymbol{x}_j|\boldsymbol{\theta}^i)^{t_i^j} \tag{11.2.7}$$

where t_i^j is one when i is the class of the jth case and zero otherwise. This way of representing the classification with an array of length K where all entries are zero except the true one is called **one hot encoding**.

cross-entropy loss

In machine learning and in software made for it, maximizing the likelihood (11.2.7) or (11.2.1) often goes by a different name and is given a different interpretation. If

we take the log of the likelihood (11.2.7) and multiply by -1 we get

$$L = -\log \mathcal{L} = \sum_{j=0}^{N} L_j \tag{11.2.8}$$

$$= -\sum_{j=0}^{N} \left[\sum_{i=0}^{K} t_i^j \log p_i(\boldsymbol{x}_j|\boldsymbol{\theta}) \right] \tag{11.2.9}$$

The part in the brackets with the negative sign is often called the **cross-entropy** of the jth case. Here t_i^j is interpreted as a probability distribution of the data jth data point, which it isn't, but you could consider it the bootstrap approximation of the distribution. Minimizing this entropy is equivalent to maximizing the likelihood.

The interpretation is that the cross-entropy expresses the loss of certainty that comes from using the classifier rather than the label t_i which are taken to be certain. Remember, t_i is 1 for the class that the case falls into and 0 for all others classes. If the classifier is perfect $p_i = 1$ when $t_i = 1$ and $p_i = 0$ for all other classes. In this case $L_j = -\log(1) = 0$ which is the smallest it can be: perfect certainty, no information loss when using the classifier instead of the true values. In the binary case

$$L_j = -\left[t_j \log(p_j) + (1 - t_j) \log(1 - p_j)\right] = \begin{cases} \log(p_j) & t_j = 1 \\ \log(1 - p_j) & t_j = 0 \end{cases} \tag{11.2.10}$$

We will return to the concept of entropy in chapter 14, but for now we can accept that minimizing the cross-entropy is equivalent to maximizing the likelihood.

Chapter 12

Maximum Likelihood, Fisher Information, Error Forecasting and Experimental Design

12.1 The Maximum Likelihood Estimator

One usually would like a specific value for a parameter to represent the outcome of an experiment. One particular choice, and often the only reasonable one, is the maximum likelihood estimator or **MLE** which we will signify by $\hat{\boldsymbol{\theta}}$. It is the parameter values where the likelihood is maximized

$$\frac{\partial \mathcal{L}}{\partial \theta_i}(\hat{\boldsymbol{\theta}}) = 0 \tag{12.1.1}$$

We have already seen that an explicit form for the MLE can be found in the case of a linear model and Gaussian distributed data with known covariances (section 6.1) and that it is unbiased ($< \hat{\boldsymbol{\theta}} >= \boldsymbol{\theta}$). When the model is nonlinear and/or the noise is to be measure simultaneously the MLE is often found numerically.

For example one of the simplest ways of finding the MLE numerically is as follows. We start at some point in parameter space $\boldsymbol{\theta}_o$. The Taylor expansion of the log-likelihood around this point is

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{\theta}_o) + \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}_o)}{\partial \theta_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)_i + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)_i \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta}_o)}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)_j + \dots \tag{12.1.2}$$

$$= \ln \mathcal{L}(\boldsymbol{\theta}_o) + (\boldsymbol{\theta} - \boldsymbol{\theta}_o) \cdot \boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}_o) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)^T \boldsymbol{F}(\boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o) + \dots \tag{12.1.3}$$

where the curvature or Hessian matrix is

$$F_{ij}(\boldsymbol{\theta}) = \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \tag{12.1.4}$$

If we take the gradient of the Taylor expansion with respect to $\boldsymbol{\theta}$ we get

$$\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}_o) + \boldsymbol{F}(\boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o). \tag{12.1.5}$$

We want to find the maximum where $\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}) = 0$. This expansion will not be perfect in general, but we can, in most cases, use it to find the maximum iteratively. The process is to calculate $\boldsymbol{F}$ and $\boldsymbol{\nabla} \ln \mathcal{L}$ at the current point and then step to the next point with

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \boldsymbol{F}^{-1}(\boldsymbol{\theta}_n)\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}_n) \tag{12.1.6}$$

This is repeated until $\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}) = 0$ to within some tolerance. This finds the maximum quickly in many cases. There are many more sophisticated algorithms for finding the maximum of a scalar function in n-dimensions and ones that don't require calculating the Hessian which can sometimes be difficult. It is also possible that there are multiple maxima and this will converge on a local maximum and not the global one. Boundaries to parameter space can also complicate things.

In section 12.6 we will see that the MLE has special properties when the amount of independent data is large and certain regularity conditions on the likelihood are met. But first let us look at some general properties of the likelihood function.

12.2 Fisher information and the minimum variance limit

Let us derive an important limitation on all estimators. The normalization of the likelihood is of course

$$\int d^n x \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = 1 \tag{12.2.1}$$

Taking the derivative of this with respect to a parameter θ_i gives

$$\int d^n x \frac{\partial}{\partial\theta_i} \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \int d^n x \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) \frac{\partial \ln \mathcal{L}}{\partial\theta_i} \tag{12.2.2}$$

$$= \left\langle \frac{\partial \ln \mathcal{L}}{\partial\theta_i} \right\rangle = 0 \tag{12.2.3}$$

since $\langle \dots \rangle = \int d^n x \mathcal{L}(\boldsymbol{x})(\dots)$. Differentiating this again gives

$$\int d^n x \left(\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \mathcal{L}}{\partial \theta_j} + \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \mathcal{L} \right) = \int d^n x \left(\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \mathcal{L} + \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \mathcal{L} \right) = 0 \quad (12.2.4)$$

In other words

$$\mathcal{F}_{ij} \equiv \left\langle \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \right\rangle = - \left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right\rangle \quad (12.2.5)$$

where $\mathcal{F}_{ij}$ is known as the **Fisher information matrix** (not to be confused with the Hessian $\boldsymbol{F}$ matrix which is not averaged).

Say we have an estimator for the parameter θ_i which we will call $\tilde{\theta}_i(\boldsymbol{x})$. Its mean will be

$$\left\langle \tilde{\theta}_i \right\rangle = \int d^n x \, \tilde{\theta}_i(\boldsymbol{x}) \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \theta_i + b(\boldsymbol{\theta}) \quad (12.2.6)$$

where $b(\boldsymbol{\theta})$ is the bias which could be zero or not. Taking the differential of this with respect to θ_i gives

$$\int d^n x \, \tilde{\theta}_i(\boldsymbol{x}) \frac{\partial \mathcal{L}}{\partial \theta_i} = 1 + \frac{\partial b}{\partial \theta_i} \quad (12.2.7)$$

$$\int d^n x \, \tilde{\theta}_i(\boldsymbol{x}) \mathcal{L} \frac{\partial \ln \mathcal{L}}{\partial \theta_i} = 1 + b' \quad (12.2.8)$$

$$\left\langle \tilde{\theta}_i(\boldsymbol{x}) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle = 1 + b' \quad (12.2.9)$$

It follows from (12.2.3) that

$$\left\langle (\tilde{\theta}_i(\boldsymbol{x}) - \theta_i) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle = 1 + b' \quad (12.2.10)$$

since the extra term will be zero. This is the covariance between the estimator and the derivative of the log likelihood. The Cauchy-Schwarz inequality applies to any covariance so

$$\left[\left\langle (\tilde{\theta}_i(\boldsymbol{x}) - \theta_i) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle \right]^2 \leq Var[\tilde{\theta}_i] Var \left[\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right] = Var[\tilde{\theta}_i] \mathcal{F}_{ii} \quad (12.2.11)$$

or

$$Var[\tilde{\theta}_i] \geq \frac{(1 + b')^2}{\mathcal{F}_{ii}} \quad (12.2.12)$$

If the estimator is unbiased $b' = 0$. This is called the **Cramér-Rao limit** or inequality. It puts an absolute bound on the variance of any estimator of a parameter. An estimator that reaches this bound is called an **efficient estimator** or EE. It is the best you can do (at least in terms of the variance) so if you can prove that your estimator reaches this limit and is unbiased there is no need to look any further for a better one. Not all problems have an EE. For example there is no EE for σ of Gaussian distributed data with zero mean. The **efficiency** of an estimator is the ratio of its variance relative to the minimum variance limit.

The Fisher matrix is sometimes called simply the information. It can be interpreted as a measure of how much information the data contains about a parameter. The Cramér-Rao limit is one reason for this interpretation. Note also that $\mathcal{F}_{ii}$ is average of the curvature or Hessian matrix of the log-likelihood. $\boldsymbol{\mathcal{F}}$ measures the rate at which the posterior drops off from its maximum in parameter space on average, i.e. how pointy the peak is. Note that the Fisher matrix is not a function of any data set. It is a property of the statistical model.

Directly from its definition it is easily shown that the Fisher matrix is symmetric and transforms like a tensor under changes of the parameters from a set $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$,

$$\mathcal{F}'_{ab} = \frac{\partial \theta_i}{\partial \theta'_a} \mathcal{F}_{ij} \frac{\partial \theta_j}{\partial \theta'_b} \tag{12.2.13}$$

Problem 46. *Prove that the sample mean is an efficient estimator of the mean of N uncorrelated Gaussian variables.*

Problem 47. *What is the efficiency of the median as an estimate of the mean in the uncorrelated Gaussian case?*

Problem 48. *Consider the estimator $A = a\overline{x} = \frac{a}{N}\sum_i^N x_i$ for the mean. What is the value of a that minimizes the variance $\langle (A-\mu)^2 \rangle$? What is the efficiency of this estimator? Does this violate the minimum variance limit?*

Problem 49. *For n identically normally distributed variables with known mean show that the variance estimator*

$$S_n^2 = \frac{1}{n}\sum_i^n (x - \overline{x})^2 \tag{12.2.14}$$

has an efficiency greater than one.

12.3 Forecasting and the Fisher matrix

In planning experiments and astronomical surveys it is often necessary to predict how well particular parameters will be measured. No one would fund a satellite or particle accelerator without some idea of how well it will measure things of interest. One way of forecasting these errors that is in wide use in cosmology is to use the Fisher matrix and the Cramér-Rao limit on the variance. One finds an expression for the log-likelihood and takes its derivatives. Then one picks fiducial parameter values, usually the values expected, and then averages using the same likelihood to get the Fisher matrix. Then the Cramér-Rao limit is used

$$Var[\theta] = \sigma_\theta^2 \simeq \frac{1}{\mathcal{F}_{\theta\theta}} \tag{12.3.1}$$

There are several criticisms of this method of forecasting errors. One is that for different fiducial parameter values the Fisher matrix can be quite different. Another is that the Cramér-Rao limit is not likely to be reached in practice because there is no EE and/or there are unaccounted for systematic errors which dominate when the statistical errors are small. Still another is that, as we will see, it does not account for degenercies between parameters, although in section 12.4 we will see that there are approximations that try to take this into account.

12.3.1 Example: Simple Cosmological Supernovae

As a simple example let's consider a simplified version of the famous type Ia supernova (SN) surveys that established that the Universe is accelerating in its expansion and won Perlmutter, Schmidt and Riess the Nobel prize in 2011. There exists a relationship between the width of a type Ia supernova's light curve, i.e. the length of time it is bright, and it's peak luminosity. For this exercise let's assume that the SNe brightnesses have already been corrected using this relationship and that the error in the corrected magnitudes are Gaussian distributed. (The uncertainty in this relation is not actually small enough that this can be assumed, but we will simplify the problem for now.) We will call the corrected brightnesses b_i. The corrected intrinsic peak luminosity, L_o, is unknown but the same for all SNe. The observed brightness is related to the intrinsic luminosity through the luminosity distance

$$D_L(z, H_o, \Omega_m, \Omega_\Lambda) = \frac{(1+z)c}{H_o} \int_0^z dz' \, \frac{1}{\sqrt{\Omega_m(1+z')^3 + 1 - \Omega_m}} = \frac{c}{H_o} d_L(z, \Omega_m) \tag{12.3.2}$$

where z is the SN's redshift, H_o is the Hubble constant, Ω_m is the average density of the Universe in units of the critical density and Ω_Λ is the cosmological constant

in the same units. Here it has been assumed that the Universe is geometrically flat although this is not necessary. In this case the density in the cosmological constant is $\Omega_\Lambda = 1 - \Omega_m$. $d_L(z, H_o, \Omega_m)$ is the luminosity distance in "Hubble lengths".

Our goal might be to figure out how many supernovae will be required to measure Ω_Λ to say 10%. This being astronomy the measurements and errors are usually given in magnitudes. The magnitude of the SN will be

$$m = M_o + 5\log_{10}(D_L(z, H_o, \Omega_m)) \tag{12.3.3}$$
$$= M_o + 5\log 10(H_o/c) + 5\log_{10}(d_L(z, \Omega_m)) \tag{12.3.4}$$

where M_o is an undetermined constant which includes the intrinsic peak luminosity. With these assumptions the likelihood is

$$\ln \mathcal{L}(\boldsymbol{m}, \boldsymbol{z}|\Omega_m, \Omega_\Lambda) = -\frac{1}{2}\sum_i \frac{1}{\sigma_i^2}\left(m_i - M_o - 5\log 10(H_o/c) - 5\log_{10} d_L(z_i, \Omega_m)\right)^2 - \frac{1}{2}\sum_i \ln(2\pi\sigma_i^2 \tag{12.3.5}$$

$$= -\frac{1}{2}\sum_i \frac{1}{\sigma_i^2}\left(m_i - \tilde{M}_o - \mu(z_i, \Omega_m)\right)^2 - \frac{1}{2}\sum_i \ln(2\pi\sigma_i^2) \tag{12.3.6}$$

Note that because M_o and H_o come into the likelihood only as a product there is no way data could determine them separately without additional information. They are **degenerate parameters** in that they cannot be disentangled from one anther. Sometimes these degeneracies are obvious, as in this case, and sometimes they are not.

Now let's find the Fisher matrix

$$\frac{\partial}{\partial\Omega_m}\ln \mathcal{L}(\boldsymbol{m}, \boldsymbol{z}|\Omega_m) = -\sum_i \frac{1}{\sigma_i^2}\left(m_i - \tilde{M}_o + \mu(z_i, \Omega_m)\right)\frac{\partial\mu(z_i)}{\partial\Omega_m} \tag{12.3.7}$$

To find the Fisher matrix you have the choice of taking another derivative or squaring this. I'll choose to take another derivative

$$\frac{\partial^2}{\partial\Omega_m^2}\ln \mathcal{L}(\boldsymbol{m}, \boldsymbol{z}|\Omega_m) = -\sum_i \frac{1}{\sigma_i^2}\left[\left(\frac{\partial\mu(z_i)}{\partial\Omega_m}\right)^2 + \left(m_i - \tilde{M}_o - \mu(z_i, \Omega_m)\right)\frac{\partial^2\mu(z_i)}{\partial\Omega_m^2}\right] \tag{12.3.8}$$

If we take the average of this the second term will be zero because according to the likelihood $\langle m_i\rangle = \tilde{M}_o + \mu(z_i, \Omega_m)$ so

$$\mathcal{F}_{\Omega_m\Omega_m} = -\left\langle \frac{\partial^2}{\partial\Omega_m^2}\ln \mathcal{L}(\boldsymbol{m}, \boldsymbol{z}|\Omega_m)\right\rangle \tag{12.3.9}$$

$$= \sum_i \frac{1}{\sigma_i^2}\left(\frac{\partial\mu(z_i)}{\partial\Omega_m}\right)^2 \tag{12.3.10}$$

The other components of the Fisher matrix are

$$\mathcal{F}_{M_o M_o} = \sum_i \frac{1}{\sigma_i^2} \tag{12.3.11}$$

$$\mathcal{F}_{M_o \Omega_m} = \sum_i \frac{1}{\sigma_i^2} \frac{\partial \mu(z_i)}{\partial \Omega_m} \tag{12.3.12}$$

where

$$\frac{\partial \mu}{\partial \Omega_m} = 5 \log_{10}(e) \frac{\partial}{\partial \Omega_m} \ln d_L(z) = -2.17147 \frac{(1+z)}{2 d_L(z)} \int_0^z dz' \, \frac{(1+z')^3 - 1}{(\Omega_m (1+z')^3 + (1-\Omega_m))^{3/2}} \tag{12.3.13}$$

We don't yet know the redshifts of the supernovae that will be observed. However we can guess from past observations and/or the survey strategy what the redshift distribution is likely to be. Let us say that it is something like $f(z) \propto x^\alpha e^{-z/z_o}$. Using this we can convert the sums into integrals

$$\sum_i \to n \int dz \; f(z) \tag{12.3.14}$$

So that for example

$$\mathcal{F}_{\Omega_m \Omega_m} = \frac{n}{\sigma^2} \int dz \; f(z) \left(\frac{\partial \mu(z)}{\partial \Omega_m} \right)^2 \tag{12.3.15}$$

were $f(z)$ is normalized to one and σ^2 has been approximated as constant for all supernovae.

For 1 supernovae, $\sigma_m = 0.3$ mag, redshift distribution parameters $\alpha = 2$ and $z_0 = 0.15$ the Fisher matrix is $\mathcal{F}_{\Omega_m \Omega_m} = 1.67$, $\mathcal{F}_{M_o Mo} = 11.1$, $\mathcal{F}_{\Omega_m M_o} = 3.18$ for the fiducial model $\Omega_m = 0.3$. At a different point in parameters space, $\Omega_m \neq 0.3$, this will change. And for a different redshift distribution this would change.

12.4 The Asymptotic Normal Approximations

Let us expand the likelihood around the MLE (or MPE for a uniform prior)

$$\ln \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) \simeq \ln \mathcal{L}(\boldsymbol{d}|\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \cdot \left. \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left. \frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \mathcal{O}\left(|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^3 \right) \tag{12.4.1}$$

$$= \ln \mathcal{L}(\boldsymbol{d}|\hat{\boldsymbol{\theta}}) + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left. \frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \mathcal{O}\left(|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^3 \right) \tag{12.4.2}$$

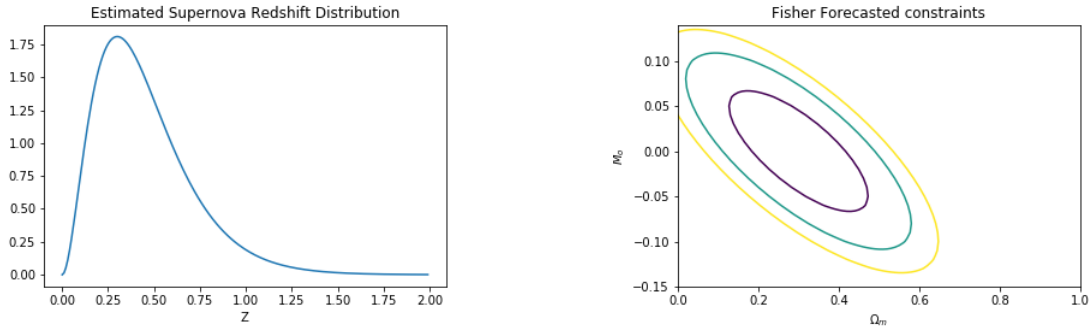

Figure 12.1: The estimated supernova redshift distribution on the left and the forecasted constraints on Ω_m and the peak luminosity normalization. Here 100 supernovae are assumed and $\sigma_m = 0.3$ mag. The redshift distribution parameters are $\alpha = 2$ and $z_0 = 0.15$.

where the second line comes from the requirement that $\hat{\boldsymbol{\theta}}$ be the maximum. As we have seen, there are no higher order terms for a linear model. When the model is nonlinear we would expect this approximations to get better as the amount of data gets larger and the constraints on the parameters get stronger. In fact, if some conditions are met, $\hat{\theta}$ is guaranteed to approach normality as the amount of data gets larger, see section12.6.

Ignoring the higher order terms, the average log-likelihood will be

$$\langle \ln \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) \rangle \simeq \left\langle \ln \mathcal{L}\left(\boldsymbol{d}|\hat{\boldsymbol{\theta}}\right) \right\rangle - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{12.4.3}$$

This leads us to approximate the posterior of a future experiment as

$$p(\boldsymbol{\theta}) \simeq \frac{|\boldsymbol{\mathcal{F}}|}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \tag{12.4.4}$$

at least near its peak. With this you see that the Cramér-Rao limit (12.3.1) is the *conditional* variance for one parameter given this posterior, the with all other parameters held fixed.

Following the rules for manipulating multivariant Gaussian distributions discussed in section **??** we can find some useful properties of this approximation. The *parameter* covariance matrix will be $\boldsymbol{\mathcal{F}}^{-1}$. The variance of a single parameter *after marginalizing* over the all the other parameters is

$$\sigma_\theta^2 \simeq \left[\boldsymbol{\mathcal{F}}^{-1}\right]_{\theta\theta} \tag{12.4.5}$$

You can also find the marginalized posterior for a subsample of parameters by inverting $\boldsymbol{\mathcal{F}}$, removing the rows and columns that correspond to the marginalized parameters and then inverting back to get $\boldsymbol{\mathcal{F}}$ in that smaller space and then use (12.4.4).

Another very handy property of this approximation is that you can easily add priors on the parameters from other experiments (at least up to second order in the log of the likelihood). Since the log of the posterior is the sum of the log of the likelihood and the prior it follows that

$$\boldsymbol{\mathcal{F}}^{tot} = \boldsymbol{\mathcal{F}} + \boldsymbol{C}^{-1}_{\text{prior}} \tag{12.4.6}$$

$\boldsymbol{C}^{-1}_{\text{prior}}$ could be the precision matrix of the parameters from some previous experiment or the Fisher matrix from some other possible experiment. For example we might ask, "What are the constraints on the cosmological parameters from the supernova experiment discussed above combined with the constraints we already have from CMB observations?" Because measurements of the cosmological parameters in particular tend to have large degeneracies, the answer to this question is not obvious. It could be that one experiment has parameters that the other does not. In this case the rows and columns of $\boldsymbol{\mathcal{F}}$ corresponding to the parameters that the experiment does not have should be set to zero which corresponds to no constraint.

Problem 50. *Show that if the likelihood depends only on a single combination of two parameters through a function $ff(\theta_1, \theta_2)$, that is*

$$\ln \mathcal{L}(\boldsymbol{x}|\theta_1, \theta_2) = \ln \mathcal{L}\left(\boldsymbol{x}|f(\theta_1, \theta_2)\right) \tag{12.4.7}$$

then the Fisher matrix will have a determinant of zero. Assume the derivatives of $f(\theta_1, \theta_2)$ do not vanish. What is the eigenvector that has a zero eigenvalue in terms of the derivatives of $f(\theta_1, \theta_2)$? This is a direction of degeneracy.

You will see by solving the above problem that degenerate combinations of the parameters correspond to eigenvectors of $\boldsymbol{\mathcal{F}}$ with eigenvalues of 0. Their existence will make $\boldsymbol{\mathcal{F}}$ non inevitable. If this is the case, the degenerate combinations of parameters should be found and replaces with a smaller set of non degenerate parameters before taking the inverse.

Any constraint plot derived from the approximate posterior (12.4.4) will be a series of ellipses. Traditionally one plots the contours that contain 0.68, 0.95 and 0.99. The correct contour levels for $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ can be found using a χ^2 distribution function from a statistical software library. Figure 12.1 shows such a plot for our simplified cosmological supernova example.

Problem 51. *Show that the area or volume of an ellipsoid in n dimensional parameter space with*

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) < X^2 \tag{12.4.8}$$

is

$$V = \frac{1}{\sqrt{|\boldsymbol{\mathcal{F}}|}} \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} \frac{X^n}{(n-1)} \tag{12.4.9}$$

and specifically in 2 dimensions $V = \pi |\boldsymbol{\mathcal{F}}|^{-1/2} X^2$*. For this reason* $\sqrt{|\boldsymbol{\mathcal{F}}|}$ *is sometimes used as a* **figure of merit** *because it is a single number that signifies how well an experiment will constrain a combination of parameters.*

One last note on Fisher matrix forecasting. It is approximate and depends only on the average of an expansion around the peak of the posterior. This approximation can break down when the constraints are not very strong compared to nonlinearities in the model and when there are significant nonlinear degeneracies in the parameters which is often the case in the cosmological setting. It also estimates the variances in the parameters with their minimum possible value, which is optimistic. For these reasons and others it might not give accurate estimates of the errors that will eventually be achieved. However, this method can be of great use in designing experiments or planning a survey strategy. If you want to measure one or a few parameters in particular and there is freedom in the experimental design (amount of data, range of an independent variable, whether to survey a large area of sky shallowly or a smaller area more deeply) you can calculate the Fisher estimate of the errors for different experimental designs and find the optimal values.

Problem 52. *The parameters magnitude,* m*, and redshift,* z*, have a Fisher matrix*

$$\begin{pmatrix} F_{m,m} & 0 \\ 0 & F_{zz} \end{pmatrix} \tag{12.4.10}$$

What is the Fisher matrix for the parameters D *and* z *if the absolute luminosity, or magnitude, is perfectly known* ($m = 2.5 \log_{10}(L/D^2) + m_o$)*?*

Now let's say the absolute magnitude ($M = 2.5 \log_{10}(L)$) *is not known, but, from prior information, we know that it has a distribution* $M \sim \mathcal{N}(M_o, \sigma_M)$*. What is the Fisher matrix for* D*,* z *and* M*? If you marginalize over* M *what is the Fisher matrix for* D *and* z*?*

12.5 Fisher Matrix with Gaussian Distributed Data

If the data is Gaussian distributed and the mean, $\boldsymbol{\mu}$ and/or the covariance, $\boldsymbol{C}$, of the distribution depends on some parameters α β then the Fisher matrix takes the form

$$\boldsymbol{\mathcal{F}}_{\alpha\beta} = \boldsymbol{\mu}_{,\alpha}^{T} \boldsymbol{C}^{-1} \boldsymbol{\mu}_{,\beta} + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{C}^{-1} \boldsymbol{C}_{,\alpha} \boldsymbol{C}^{-1} \boldsymbol{C}_{,\beta}\right] \tag{12.5.1}$$

Problem 53. *Show that equation (12.5.1) is correct. Use the identities*

$$\boldsymbol{C}^{-1}{}_{,\beta} = -\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1} \quad \textit{and} \quad \frac{d}{d\beta}\ln|\boldsymbol{C}| = tr\left[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\right] \tag{12.5.2}$$

This form of the Fisher matrix comes up a lot in Cosmology. In the standard cosmological model, the Fourier modes of the primordial density field are Gaussian distributed which results in the same being true for the spherical harmonic modes of the Cosmic Microwave Background (CMB) and for the Fourier modes of the distribution of galaxies (at least on large scales). The power spectrum of these modes is dependent on Cosmological parameters and departures from General Relativity if they exist. The Fisher matrix is used in forecasting constraints and in numerical algorithms for finding best fit parameters using large data sets. I is also often used as a substitute for $\boldsymbol{F}$ in (12.1.6) when finding the maximum likelihood because $\boldsymbol{F}$ can be computationally expensive and $\boldsymbol{\mathcal{F}}$ often works just as well.

12.5.1 independent samples

Consider the special case of n independent normally distributed measurements with the same means and variances. In this case the covariance matrix is

$$\boldsymbol{C} = \sigma^2\boldsymbol{I} \qquad \boldsymbol{C}^{-1} = \frac{1}{\sigma^2}\boldsymbol{I} \tag{12.5.3}$$

where $\boldsymbol{I}$ is the identity matrix. From(12.5.1) we can find that

$$\mathcal{F}_{\mu\mu} = \frac{n}{\sigma^2} \tag{12.5.4}$$

$$\mathcal{F}_{\mu\sigma^2} = 0 \tag{12.5.5}$$

$$\mathcal{F}_{\sigma^2\sigma^2} = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}\boldsymbol{I}\boldsymbol{C}^{-1}\boldsymbol{I}\right] \tag{12.5.6}$$

$$= \frac{1}{2\sigma^4}\mathrm{tr}\left[\boldsymbol{I}\boldsymbol{I}\boldsymbol{I}\boldsymbol{I}\right] \tag{12.5.7}$$

$$= \frac{n}{2\sigma^4} \tag{12.5.8}$$

So there is no unbiased estimator for the mean that have a variance smaller than σ^2/n. We found back in section (4.1) that the sample mean is unbiased and has a variance of σ^2/n so it is an *efficient* estimator. On the other hand, in section 4.2, we found that the variance of the standard estimator for the variance, S_n^2, is $2\sigma^4/(n-1)$ where as from the Fisher matrix above we can see that the Cramér-Rao limit is $2\sigma^4/n$ so S_n^2 has an efficiency of $(n-1)/n$.

12.5.2 Fisher matrix for a galaxy survey

The distribution of matter in the Universe on large scales is predicted to be a Gaussian random field. In a Gaussian random field the Fourier modes are statistically independent Gaussian variables with variances $P(\boldsymbol{k}, z)$ which is the power spectrum. A model is often used where the average number of galaxies in a cell, or volume in space, is proportional to the density in that cell. The actual number of galaxies in the cell is Poisson distributed around that average. The power spectrum of a Poisson field with out any density fluctuation would be equal to the inverse of the galaxy density, n. So the total covariance matrix for the modes of the galaxy distribution is the sum of the Gaussian and Poisson contributions,

$$C_{\boldsymbol{k}\boldsymbol{k}'} = \langle \delta_{\boldsymbol{k}} \delta'^{*}_{\boldsymbol{k}} \rangle = \langle \delta_{\boldsymbol{k}} \delta_{-\boldsymbol{k}} \rangle \delta^{K}_{\boldsymbol{k}\boldsymbol{k}'} = \left(P(\boldsymbol{k}, z) + \frac{1}{n} \right) \delta^{K}_{\boldsymbol{k}\boldsymbol{k}'} \tag{12.5.9}$$

where $\delta^{K}_{\boldsymbol{k}\boldsymbol{k}'}$ is the Kronecker delta. We can use the result from section 12.5 for the Fisher matrix of Gaussian data to get

$$\mathcal{F}_{\theta\beta} = \frac{1}{2} \sum_{\boldsymbol{k}} \left[\frac{1}{(P_{\boldsymbol{k}} + 1/n)^2} \frac{\partial P_{\boldsymbol{k}}}{\partial \theta} \frac{\partial P_{\boldsymbol{k}}}{\partial \beta} \right] \tag{12.5.10}$$

$$= \frac{1}{2} \sum_{\boldsymbol{k}} \left[\left(\frac{n P_{\boldsymbol{k}}}{n P_{\boldsymbol{k}} + 1} \right)^2 \frac{\partial \ln P_{\boldsymbol{k}}}{\partial \theta} \frac{\partial \ln P_{\boldsymbol{k}}}{\partial \beta} \right] \tag{12.5.11}$$

The volume of a cell in discrete Fourier space is $V_{cell} = \frac{(2\pi)^3}{V_{\rm survey}}$ where $V_{\rm survey}$ is the volume of the survey. So $2\pi k^2 dk d\mu / V_{cell} = V_{\rm survey} k^2 dk d\mu / (2\pi)^2$ where $\mu = \cos(\theta)$, the cosine of the angle between the radial direction and $\boldsymbol{k}$. So the sum over Fourier modes can be substituted with an integral

$$\sum_{\boldsymbol{k}} \to \frac{V_{\rm survey}}{(2\pi)^2} \int_{-1}^{1} d\mu \int_{k_{min}}^{k_{max}} k^2 dk. \tag{12.5.12}$$

and the Fisher matrix is

$$\mathcal{F}_{\theta\beta} = \frac{1}{8\pi^2} \int_{-1}^{1} d\mu \int_{k_{min}}^{k_{max}} k^2 dk \, \frac{\partial \ln P(k,\mu)}{\partial \theta} \frac{\partial \ln P(k,\mu)}{\partial \beta} V_{\rm eff}(k,\mu) \tag{12.5.13}$$

where the effective survey volume is

$$V_{\rm eff}(k,\mu) = \left[\frac{n P_k}{n P_k + 1} \right]^2 V_{\rm survey} \tag{12.5.14}$$

This Fisher matrix ignores many things that exist in a real galaxy redshift surveys. Nonlinear structure formation which causes the density field to be non-Gaussian for large $\boldsymbol{k}$ (small scales) destroys information about the cosmological parameters. The scale dependent and possibly nonlinear bias that relates galaxy density to mass density has not been considered and there are many sources of noise and incompleteness. As a result, this is really a limit to how well a survey could possibly do in constraining parameters.

12.6 Asymptotic behavior of the maximum likelihood estimator

The maximum likelihood estimator (MLE) has some special properties in the limit of a large amount of data that makes it special and a popular option for an estimator if it can be calculated. After our discussions in this chapter we are in a position to understand why the MLE has these properties.

They are the following:

The log likelihood of n independent data points, or data sets, x_n can be written

$$\ln\left(\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\right) = \ln\left(\prod_i^n L(x_i|\boldsymbol{\theta})\right) = \sum_i^n \ln\left(L(x_i|\boldsymbol{\theta})\right) \tag{12.6.1}$$

If $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ is the MLE and $\boldsymbol{\theta}_o$ is the true value of the parameter, under some often satisfied requirements on the regularity of the likelihood function, in the limit of large amounts of independent data:

1. $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_o$ in probability.[1] This means that as the amount of data gets very large the MLE will eventually get arbitrarily close to the true value. In other words the MLE is a *consistent* statistic.

2. The MLE is asymptotically normally distributed

$$(\hat{\theta} - \theta_o) \sim \mathcal{N}\left(0, \frac{1}{nF_{\theta\theta}}\right) \tag{12.6.2}$$

   where $F_{\theta\theta}$ is the Fisher matrix for $L(\boldsymbol{x}|\theta_o)$.

What follows is a "physicist's proof" of these properties, meaning that it is not mathematically rigorous. I think it illustrates some more general methods for doing

[1] x_n is said to convergence to x_o "in probability" if for all $\epsilon > 0$ $\lim_{n\to\infty} P(|x_o - x_n| > \epsilon) = 0$. This is often denoted $x_n \xrightarrow{p} x$.

such proof and the limitations of this theorem without getting too lost in the details. You can find more rigorous proofs in many statistics textbooks (for example Surfling (1980)).

Let us define a scaled version of the log likelihood

$$l_n(\theta) = \frac{1}{n} \ln\left(\mathcal{L}(\boldsymbol{x}|\theta)\right) = \frac{i}{n} \sum_I^n \ln\left(L(x_i|\theta)\right) \tag{12.6.3}$$

Let us also define the function

$$l(\theta) = E\left[\ln\left(L(x|\theta)\right)\right] = \int dx\ L(x|\theta_o) \ln\left(L(x|\theta)\right) \tag{12.6.4}$$

By the law of large numbers (the sample mean is equal the the distributions mean for an infinitely large sample) $l_n(\theta) \to l(\theta)$. This also applies to derivatives of $l_n(\theta)$:

$$l_n(\theta) = \frac{1}{n} \ln\left(\mathcal{L}(\boldsymbol{x}|\theta)\right) = \frac{1}{n} \sum_I^n \ln\left(L(x_i|\theta)\right) \to l(\theta) \tag{12.6.5}$$

$$l_n'(\theta) = \frac{1}{n} \frac{\partial}{\partial\theta} \ln\left(\mathcal{L}(\boldsymbol{x}|\theta)\right) = \frac{1}{n} \sum_I^n \frac{\partial}{\partial\theta} \ln\left(L(x_i|\theta)\right) \to E\left[\frac{\partial}{\partial\theta} \ln\left(L(x|\theta)\right)\right] \tag{12.6.6}$$

$$l_n''(\theta) = \frac{1}{n} \sum_I^n \frac{\partial^2}{\partial\theta^2} \ln\left(L(x_i|\theta)\right) \to E\left[\frac{\partial^2}{\partial\theta^2} \ln\left(L(x|\theta)\right)\right] = -F_{\theta\theta}(\boldsymbol{\theta}) \tag{12.6.7}$$

where the true Fisher matrix is $F_{\theta\theta} = F_{\theta\theta}(\boldsymbol{\theta}_o)$. By the central limit theorem $l_n(\theta)$, being the sum of random numbers, is asymptotically normally distributed.

For all θ

$$l(\theta) \leq l(\theta_o) \tag{12.6.8}$$

This is true because

$$\begin{aligned}
l(\theta) - l(\theta_o) &= E\left[\ln\left(L(x|\theta)\right) - \ln\left(L(x|\theta_o)\right)\right] && (12.6.9)\\
&= E\left[\ln\left(\frac{L(x|\theta)}{L(x|\theta_o)}\right)\right] && (12.6.10)\\
&\leq E\left[\left(\frac{L(x|\theta)}{L(x|\theta_o)}\right) - 1\right] \qquad \ln(x) \leq x - 1 && (12.6.11)\\
&= \int dx\ L(x|x_o)\left(\frac{L(x|\theta)}{L(x|\theta_o)} - 1\right) && (12.6.12)\\
&= \int dx\ L(x|\theta) - \int dx\ L(x|\theta_o) && (12.6.13)\\
&= 1 - 1 && (12.6.14)\\
&= 0 && (12.6.15)
\end{aligned}$$

So the property 1) follows from

1. $\hat{\theta}$ maximizes $l_n(\theta)$ by definition.
2. θ_o maximizes $l(\theta)$ by equation (12.6.8).
3. Since $l_n(\theta) \to l(\theta)$ for large n and for all θ, there maxima must be the same in this limit.

Now for the distribution of $\hat{\theta}$. The mean value theorem of calculus holds that

$$f(a) = f(b) + f'(c)(a - b) \tag{12.6.16}$$

for some c between a and b, $c \in [a, b]$. We can apply this theorem to $l_n'(\hat{\theta})$ which is zero because of the definition of $\hat{\theta}$

$$l_n'(\hat{\theta}) = l_n'(\theta_o) + l_n''(\tilde{\theta})(\hat{\theta} - \theta_o) \tag{12.6.17}$$

so

$$(\hat{\theta} - \theta_o) = -\frac{l_n'(\theta_o)}{l_n''(\tilde{\theta})} \tag{12.6.18}$$

As n gets larger $\hat{\theta} \xrightarrow{p} \theta_o$ and since $\tilde{\theta}$ is between them it must be that $\tilde{\theta} \xrightarrow{p} \theta_o$. The variance numerator is

$$Var\left[l_n'(\theta_o)\right] = E\left[\left(l_n'(\theta_o)\right)^2\right] - E\left[l_n'(\theta_o)\right]^2 \tag{12.6.19}$$

$$= E\left[\left(\frac{1}{n}\sum_i^n L'(x_i|\theta_o)\right)^2\right] - E\left[l_n'(\theta_o)\right]^2 \tag{12.6.20}$$

$$= \frac{1}{n^2}\sum_i^n E\left[\left(L'(x_i|\theta_o)\right)^2\right] - E\left[l_n'(\theta_o)\right]^2 \tag{12.6.21}$$

$$= \frac{1}{n}F_{\theta\theta} - E\left[l_n'(\theta_o)\right]^2 \tag{12.6.22}$$

$$\to \frac{F_{\theta\theta}}{n} - E\left[l'(\theta_o)\right]^2 \tag{12.6.23}$$

$$= \frac{F_{\theta\theta}}{n} - 0 \tag{12.6.24}$$

The denominator goes to $F_{\theta\theta}$ by 12.6.7. So

$$Var\left[(\hat{\theta} - \theta_o)\right] \to \frac{1}{nF_{\theta\theta}} \tag{12.6.25}$$

And by the central limit theorem $l_n'(\theta_o)$ is normally distributed.

12.7 Likelihood ratio test

The fact that the maximum likelihood estimator is asymptotically unbiased and normally distributed under some conditions motivates applying the statistical tests that where introduced in sections 8.1 through 8.6 for linear Gaussian models. Specifically we look at the likelihood ratio test for model selection here which is a generalization of the $\Delta\chi^2$ model selection discussed in section 8.6.

Again consider nested model M and K with numbers of parameters m and k with $m > k$. The goodness-of-fit statistic is

$$T = -2\ln\left[\frac{\mathcal{L}(D|\hat{\boldsymbol{\theta}}_M)}{\mathcal{L}(D|\hat{\boldsymbol{\theta}}_K)}\right] = -2\left[\ln\left[\mathcal{L}(D|\hat{\boldsymbol{\theta}}_M)\right] - \ln\left[\mathcal{L}(D|\hat{\boldsymbol{\theta}}_K)\right]\right] \tag{12.7.1}$$

The value of T_H is > 0. In the limit that the MLE is normally distributed T is χ^2_{m-k} distributed like the ΔX^2 statistic discussed in section 8.6 . The model M can be rejected using the p-value from this distribution. You can also do an F-test for model selection under the same conditions.

The statistic

$$T(\boldsymbol{\theta}) = -2\ln\left[\frac{\mathcal{L}(D|\boldsymbol{\theta})}{\mathcal{L}(D|\hat{\boldsymbol{\theta}})}\right] = -2\left[\ln\left[\mathcal{L}(D|\boldsymbol{\theta})\right] - \ln\left[\mathcal{L}(D|\hat{\boldsymbol{\theta}})\right]\right] \tag{12.7.2}$$

can be used to do parameter estimation under the same conditions. It will be χ^2_k distributed where k is the number of parameters in the limit of large amounts of data.

Before applying these tests it is important to check that the conditions for the convergence to normality apply. Among the conditions are that the amount of data be sufficiently large, the models M and K are nested and that the true parameter value $\boldsymbol{\theta}_o$ is not on the edge of the allowed range. This last requirement comes about because the regularity conditions on the likelihood eluded to in the last section are not satisfied in this case. One requirement is that the likelihood must be thrice differentiable which is clearly not satisfied at the edge of allowed parameter space.

As discussed by Protassov et al. (2002), this test is not valid for the common problem of detection, but is often incorrectly used for it. For example, detected a spectral line or source on an image. In this case the model for the data with the line would be something like

$$F(\lambda_i) = \theta_o F_o(\lambda_i) + \theta_1 f(\lambda_i) \tag{12.7.3}$$

where $F_o(\lambda_i)$ is the background or continuum and $f(\lambda_i)$ is the line profile. The model without the line is corresponds to $\theta_1 = 0$. An emission line or source cannot have a negative contribution so $\theta_1 \geq 0$ so the parameter θ_1 is on the edge of the parameters space. The limit discussed in section 12.6 does not apply in this case.

Chapter 13

Numerical Sampling methods

Ideally one is able to write out an analytic expression for the likelihood or posterior and perform integrals over it analytically or by standard numerical integration methods to find expectation values of statistics or the integrated probability for a variable being in a certain region. However, sometimes these integrals are very difficult to perform because the dimension of parameter(data)-space is high and sometimes there isn't an analytic expression for the probability (for example when a simulation is used to go between parameters and predictions).

The next best thing to integrating over an analytic function is having a large sample of deviates drawn from the distribution. With a sufficiently large sample drawn from a distribution one can use the **law of large numbers** to estimate any expectation value

$$E[g(x)] = \int_{-\infty}^{\infty} d^n x \; p(\boldsymbol{x}) g(\boldsymbol{x}) \simeq \frac{1}{n} \sum_i g(\boldsymbol{x}_i) \tag{13.0.1}$$

where the $\boldsymbol{x}_i$'s are drawn from the distribution $p(\boldsymbol{x})$.

In one dimension it is often possible to efficiently sample from a standard distribution function and any good statistical software package will have functions to do this. There are several methods used to find these deviates such as rejection and transformation methods.

13.1 probability integral transform

We already know how to transform variables. If we have a pdf $p(x)$ than in a new variable f the pdf is $p(x)\frac{dx}{df}$. If we require the distribution in f to be uniform then $p(x)\frac{dx}{df} = const.$ or $cdf = p(x)dx$ and when properly normalized $\int_0^{f(x)} df = F(x)$ is the cumulative distribution function. So if you can invert the cumulative distribution

function to get the quantile function $x = F^{-1}(f)$ then you can draw f from a uniform distribution between 0 and 1 and the corresponding x will be distributed according to the pdf $p(x) = \frac{d}{dx}F(x)$. More formally:

Theorem: $y = F_X(x)$ is uniformly distributed if x is distributed such that it's cumulative distribution is $F_X(x)$.

Proof:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) && (13.1.1)\\ &= P\left(F_X(x) \leq y\right) && (13.1.2)\\ &= P\left(x \leq F_X^{-1}(y)\right) && (13.1.3)\\ &= F_X\left(F_X^{-1}(y)\right) && (13.1.4)\\ &= y && (13.1.5) \end{aligned}$$

So y is uniformly distributed.

For example, say you want a random point within a sphere of radius R. The pdf is

$$p(r, \theta, \phi) dr d\theta d\phi \propto r^2 d\cos(\theta) d\phi \tag{13.1.6}$$

The cumulative distribution for the radius is

$$F(r) = \left(\frac{r}{R}\right)^3 \tag{13.1.7}$$

So inverting this gives

$$r = RF^{1/3} \tag{13.1.8}$$

So you can draw a uniform number from 0 to 1 and in this way find a random radius. This same method could be used to find a position for a random particle within a particular density profile for example.

In D dimensional space a random point within a D-ball (the interior of a D-1 sphere) can be found in the following steps

- Draw D normally distributed numbers, $\boldsymbol{x}$. Since this is an isotropic distribution the vector $\boldsymbol{x}/|\boldsymbol{x}|$ is uniformly distributed on the unit (D-1)-sphere.
- Calculate $|\boldsymbol{x}|^2 = \sum_i^D x_i^2$.
- Draw a uniform number, F, between 0 and 1.

- Calculate the new radius with $r = R_{\max} F^{\frac{1}{D}}$.
- Renomalize the vector

$$\boldsymbol{y} = \frac{r}{|\boldsymbol{x}|}\boldsymbol{x} \tag{13.1.9}$$

See Press et al. (2007) for more information on generating random deviates in one dimension.

13.2 numerical confidence levels

Frequentist hypothesis testing and parameter confidence intervals are based on comparing some statistic of the data to the distribution of that statistic given a certain hypothesis or parameter set. One can think of many statistics whose distribution is hard or impossible to calculated analytically. For example, say you have a model that requires a lengthy calculation to predict the number of solar neutrinos you will detect. The inputs to this calculation – temperature and density profile of the sun, scattering cross-section of various nucleons, etc – are not perfectly known so the predictions are not perfectly known even in terms of the average rate. What is the distribution of the rate or the number of neutrinos that will be detected over a certain period of time? Can your model be ruled out?

The situation is like this

$$\textbf{Initial conditions / input variables} \rightarrow \textbf{simulation / theory} \rightarrow \textbf{observables} \tag{13.2.1}$$

Another example, say you have a numerical simulation that starts with some primordial gas cloud and predicts the number of globular clusters in the galaxy that forms. Each time you run the simulation with random initial conditions taken from a reasonable distribution – mass of cloud, random density fluctuations in the cloud, etc. – you get out a different numbers of globular clusters. You observe the number of globular clusters in the galaxies around us. You derive a statistic from them – for example, the average number of globular clusters or the maximum number of globular clusters or the minimum number of globular clusters. You find that this statistic isn't the same as in your simulations. Can you rule out your model and conclude that there is something incorrect in your simulation?

From a sample one can easily estimate the probability of a statistic being larger then X with k/n where k is the number of samples above and n is the total number

of trials. This is only an approximation however. How do we know the accuracy of probability or confidence level given a sample?

For frequentist this topic falls into the subject of quantile estimation that was touched on in section 4.6. Here the boundary of the interval that contains some fraction of the probability is estimated from the sample. Unfortunately the distribution of this estimate depends on the distribution itself so it is not possible to determine how good the estimate is without assuming something about the distribution. But if you knew the distribution you wouldn't be trying to estimate its quantile from a sample.

Here is where being a bit flexible with ideology comes in handy because we can find a Bayesian constraint on the frequentist confidence level that assumes nothing about the underlying distribution. If the probability of a statistic being larger than X is p then we know that the probability of k samples out of n being larger than X is given by the binomial distribution

$$P(k|p,n) = \binom{n}{k} p^k (1-p)^{n-k} \tag{13.2.2}$$

Assuming a uniform prior and using the integral form of the beta function (appendix A.4 equation A.4.6) , we can renormalize this to get the posterior for p

$$P(p|n,k) = \frac{\Gamma(n+2)}{\Gamma(n-k+1)\Gamma(k+1)} p^k (1-p)^{n-k} \tag{13.2.3}$$

$$= \frac{(n+1)!}{(n-k)!k!}\, p^k (1-p)^{n-k} \tag{13.2.4}$$

for $0 \leq p \leq 1$ and zero otherwise. This is true no matter what the underlying distribution is.

The mode of this distribution can be found in the usual way (take the derivative of $\ln P(p)$ and set it equal to zero)

$$p_{ML} = \frac{k}{n} \tag{13.2.5}$$

which is just what you might have guessed. If 5% of the distribution is larger than X then $\sim 5\%$ of the sample should be larger than X.

The average of the posterior is

$$\langle p \rangle = \frac{(n+1)!}{(n-k)!k!} \int_0^1 dp\; p^{k+1} (1-p)^{n-k} \tag{13.2.6}$$

$$= \frac{(n+1)!(k+1)!}{k!(n+2)!} \tag{13.2.7}$$

$$= \frac{(k+1)}{(n+2)} \tag{13.2.8}$$

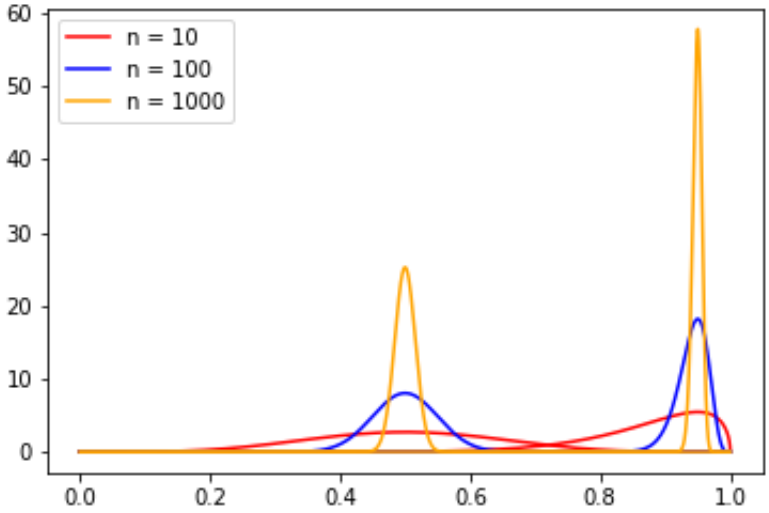

Figure 13.1: The posterior for the cumulative probability up to a boundary given that the fraction of samples above the boundary is 50% (center) and 95% (right). The the total number of samples is as in the legend.

, not equal to the mode, and we can find its variance

$$\langle p^2 \rangle = \frac{(n+1)!}{(n-k)!k!} \int_0^1 dp\; p^{k+2}(1-p)^{n-k} \tag{13.2.9}$$

$$= \frac{(k+2)(k+1)}{(n+3)(n+2)} \tag{13.2.10}$$

so

$$\sigma_p^2 = \langle p^2 \rangle - \langle p \rangle^2 \tag{13.2.11}$$

$$= \frac{(k+1)}{(n+2)} \left[\frac{(k+2)}{(n+3)} - \frac{(k+1)}{(n+2)} \right] \tag{13.2.12}$$

$$\simeq \frac{3\langle p \rangle \left(1 - \langle p \rangle\right)}{n} + \mathcal{O}(1/n^2) \tag{13.2.13}$$

The distribution is plotted in figure 13.1. The distribution is narrower for k/n near the extremes, $k \sim n$ and $k \sim 0$, for the same number of samples, but in these cases one is usually more concerned with accuracy since the difference between 95% confidence and 99% confidence is large while the difference between 25% and 50% doesn't make

much difference since neither one would exclude the hypothesis significantly, i.e. it is only in cases of high significance that you need a lot of samples.

If you want to know the p-value to an accuracy of 0.001 you will need

$$n \gtrsim \frac{3\langle p\rangle\left(1-\langle p\rangle\right)}{(0.001)^2} = 3\times 10^6\langle p\rangle\left(1-\langle p\rangle\right) \tag{13.2.14}$$

simulations. For $\langle p\rangle = 0.99$ this is $n \gtrsim 3\times 10^4$.

13.3 Monte Carlo Integration & Importance Sampling

A related numerical technique to the subjects that will be discussed here is Monte Carlo Integration. This is a way of using the law of large numbers (13.0.1) to estimate a multidimensional integral that cannot be done analytically or by using a standard one dimensional method such as the trapezoids or Romberg. It is typically used when the number of dimensions is high and/or the boundaries to the region of integration are complicated, and there is no other choice.

We want to estimate the integral of some function $g(\boldsymbol{x})$,

$$\int_{\partial V} d^n x\; g(\boldsymbol{x}) = \int_{\partial V} d^n x\; p(\boldsymbol{x})\left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \tag{13.3.1}$$

$$= \int_{-\infty}^{\infty} d^n x\; p(\boldsymbol{x})\left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x})}\right)\Theta(\boldsymbol{x}\in V)) \tag{13.3.2}$$

$$= \left\langle\left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x})}\right)\Theta(\boldsymbol{x}\in V)\right\rangle_p \tag{13.3.3}$$

$$\simeq \frac{1}{n}\sum_{x_i\in V}\left(\frac{g(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right)\Theta(\boldsymbol{x}_i\in V) \qquad\qquad \boldsymbol{x}_i\sim p(x) \tag{13.3.4}$$

where the x_i's are drawn from the distribution $p(x)$. This is guaranteed to converge to the correct answer as $n\to\infty$ as long as $p(\boldsymbol{x})\neq 0$ everywhere that $g(\boldsymbol{x})$ is not within the volume of integration. The estimated error on this would be

$$\pm\frac{1}{n}\sqrt{\sum_{x_i\in V}\left(\frac{g(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right)^2 - \frac{1}{n}\left(\sum_{x_i\in V}\frac{g(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right)^2} \tag{13.3.5}$$

The closer the sampling distribution $p(\boldsymbol{x}_i)$ is to $g(\boldsymbol{x}_i)$ the better will be the estimate. If some standard probability distribution that can be sampled from easily

resembles the function to be integrated this can be useful. Quite complicated algorithms can be derived from this where the volume is partitioned and the sampling function $p(\boldsymbol{x})$ refined adaptively to improve convergence. We will not go into those here, but the connection to what proceeded this section and what follows should be clear.

Problem 54. *Write a program that calculates the value of π by finding the area within a unit circle by Monte Carlo. Draw uniform deviates from within a square that circumscribes the circle. The area of the circle is on average the area of the circle times the fraction of deviates that are within the circle. What is the error in this estimate of π?*

When MC integration is applied to probability distributions themselves it is called **importance sampling**. If you want to know the expectation value of $f(\boldsymbol{x})$ given a pdf $p(\boldsymbol{x})$, but you cannot sample directly from $p(\boldsymbol{x})$ you can estimate it by sampling from a distribution $q(\boldsymbol{x})$. The expectation value is then

$$E[f(\boldsymbol{x})] = \int_{-\infty}^{\infty} d\boldsymbol{x}\ f(\boldsymbol{x})p(\boldsymbol{x}) \tag{13.3.6}$$

$$= \frac{1}{n}\sum_{i}^{n} f(\boldsymbol{x}_i)\left(\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}\right) \qquad \boldsymbol{x}_i \sim q(\boldsymbol{x}) \tag{13.3.7}$$

$$= \frac{1}{n}\sum_{i}^{n} f(\boldsymbol{x}_i)w_i \qquad w_i = \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \tag{13.3.8}$$

It is required that $q(\boldsymbol{x}) > 0$ everywhere $p(\boldsymbol{x}) > 0$ for this to converge to the correct expectation value. The closer $q(\boldsymbol{x})$ is to $p(\boldsymbol{x})$ the quicker this will converge. This inspires an estimate for $p(x)$ reminiscent of the bootstrap pdf

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i} w_i \delta^D(\boldsymbol{x} - \boldsymbol{x}_i) \tag{13.3.9}$$

13.3.1 importance sampling in Bayesian inference

While addressing a Bayesian inference problem it is often the case that you can write down the likelihood and the prior, but you cannot integrate their product over parameter space. As a result you cannot find the evidence or the means, variances, covariances etc. of the posterior. Remember also that the evidence is needed to do Bayesian model comparison. The model might be nonlinear and the number of parameters large and/or the likelihood might not be a simple function. Again we sample from some distribution $q(\boldsymbol{\theta})$ and for the evidence we want to integrate

$f(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ so

$$\mathcal{E}(\boldsymbol{d}) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{13.3.10}$$

$$\simeq \frac{1}{n}\sum_{i}^{n} \frac{\mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)} \qquad \boldsymbol{\theta} \sim q(\boldsymbol{\theta}) \tag{13.3.11}$$

Now if we want to average over the posterior $p(\boldsymbol{\theta}|\boldsymbol{d}) = \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\mathcal{E}(\boldsymbol{d})$ we can use (13.3.8) with this estimate for the normalization to get

$$E[f(\boldsymbol{\theta})] = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\boldsymbol{d}) f(\boldsymbol{\theta}) \tag{13.3.12}$$

$$\simeq \frac{\sum_i^n w_i\, f(\boldsymbol{\theta}_i)}{\sum_i^n w_i} \quad \text{where } w_i = \frac{\mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)} \tag{13.3.13}$$

$$\equiv \bar{f} \tag{13.3.14}$$

An estimate of the variance of this is

$$\sigma_{\bar{f}}^2 = \frac{\sum_i^n w_i^2 \left(f(x_i) - \bar{f}\right)^2}{\left[\sum_i^n w_i\right]^2} \tag{13.3.15}$$

which can be monitored until the desired accuracy is obtained.

This is nice, but often not very useful when there are more than a couple of parameters and/or the likelihood constrains the parameters to a much smaller volume than the prior does. The reason is that in a high dimensional space the volume where the likelihood is large is often a tiny, a tiny fraction of the volume allowed by the prior and you generally don't know where it is in parameters space before you start. This is one manifestation of the **curse of high dimensionality**.

There are **adaptive importance sampling** algorithms that update the sampling distribution to better fit the function to be integrated as the calculation proceeds. For example, you could sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimates of the mean and variance of the posterior based on the previous samples. The sampling becomes more efficient and the answer more stable as time goes on. Many other strategies are possible and widely used. The final weights w_i and samples $\boldsymbol{\theta}_i$ can then be used to find the mean and covariance, or other moments, of the posterior. There are also algorithms based on **Sequential Importance Sampling** or SIS where new points are chosen based on the existing points. For a review of this subject see Tokdar & Kass (2010).

13.4 Curse of high dimensionality

Consider a problem with D parameters so the parameter space is D-dimensional. One approach might be to map out the posterior by evaluating it on a regular grid. If we divide each dimension into 10 ,which would be very coarse total number, the total number of grid points will be 10^D. If D is a modest number of 10 then this is 10 billion! Just to make a very coarse map that will probably not be good enough to calculate anything quantitative.

Another way of seeing this is to look at the volume in spheres. Let us say that the likelihood constrains the parameters to a spherical region in parameters space of radius R and that the prior to a modestly larger region of radius $2R$. The ratio of the volumes is $(2R/R)^D = 2^D$ which is 1,024 for $D = 10$ so if you drew points from the prior only one in a thousand would land in the region where the likelihood is large. If the radius where 10 times smaller it would be one in 10 billion. And the dimensionality could easily be larger.

It is clear from these considerations that when the number of parameters gets large some stratagem must be used to pick points that are not simply drawn at random from the prior, but are more efficient at exploring the likelihood. Adaptive importance sampling is one approach. Another stratagem that is more commonly encountered in astrophysics is Markov Chain Monte Carlo or MCMC.

13.5 Markov Chains

In statistics a **chain** is an ordered series of random numbers, $\boldsymbol{x}_1 \ldots \boldsymbol{x}_n \ldots$ where the conditional probability of each element given the other elements is specified – $p(\boldsymbol{x}_n|\boldsymbol{x}_1 \ldots)$. You can think of the whole chain as being a single random object. The theory on chains is extensive. They can be used to model everything from gambling to the stock market to chemical reactions and many other things. There are many different types of chains with different properties. Here we will concentrate only on the type of chain that is commonly used in scientific inference problems and the properties that are important to this application and we will do so with informal definitions and no proofs.

A **Markov chain** is a chain where the conditional probability of any element $\boldsymbol{x}_n$ can be expressed as a function of only the previous element $\boldsymbol{x}_{n-1}$, (The future depends only on the present and not on the past, although the present does depend on the past.) The probability $p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)$ is known as the Markov chain's **transition kernel**. If the transition kernel is independent of n it is said to be *time-homogenious*. The chains we are interested in are **ergodic chains**. To be ergodic the chain must be

1. irreducible - A chain starting at any state $\boldsymbol{x}_o$ can reach any other state after a finite number of steps, not necessarily 1 step.

2. aperiodic - The chain will not return to the same state after some fixed number of steps and all multiples of this number of steps.

3. positive recurrent - The expectation value for the number of steps between any two states is finite.

It is also true that a Markov chain is ergodic if there is a number N such that any state can be reached from any other state in N steps and any number of steps larger than N.

The most important consequence of ergodicity is that the chain has a unique **stationary distribution** $f(\boldsymbol{x})$ such that

$$\int_{-\infty}^{\infty} d^D x_n \; f(\boldsymbol{x}_n) p(\boldsymbol{x}_{1+n}|\boldsymbol{x}_n) = f(\boldsymbol{x}_{1+n}) \tag{13.5.1}$$

which means that we can produce chains whose states are distributed according to $f(\boldsymbol{x})$ if we can find a transition kernel that satisfies this requirement. And the law of large numbers will apply

$$E[g(\boldsymbol{x})] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} g(\boldsymbol{x}_n) \tag{13.5.2}$$

Note that the transition kernel is not unique for a particular $f(\boldsymbol{x})$. We can also select one or two parameters and make a histogram which should be a representation of the marginal stationary distribution.

Also note that having a stationary distribution does *not* mean that each element in the chain is independent, i.e. $p(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n) \neq f(\boldsymbol{x}_{n+1}) f(\boldsymbol{x}_n)$. In fact, as we will see, states that are very far separated in the chain are not always independent, but as the separation increases they will eventually become independent.

Problem 55. *If you have a Markov Chain with transition probability $p(\boldsymbol{x}_{1+n}|\boldsymbol{x}_n)$ for all n what is the probability of $p(\boldsymbol{x}_{1+n}|\boldsymbol{x}_{n-1})$?*

13.5.1 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is based on finding a transition kernel that will have any desired stationary distribution. The kernel satisfies **detailed balance**:

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n) f(\boldsymbol{x}_n) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}) f(\boldsymbol{x}_{n+1}) \tag{13.5.3}$$

for all n. Detailed balance is often used in statistical physics for example in Einstein's famous derivation of stimulated emission. You can easily see that if $p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)$ satisfied detailed balance it will also satisfy (13.5.1).

You can think of this as if there were a flow of points out of state $\boldsymbol{x}_n$ into $\boldsymbol{x}_{n+1}$ and a counter flow out of $\boldsymbol{x}_{n+1}$ into $\boldsymbol{x}_n$. The flow is proportional to the probability of being in the first state times the probability of transitioning. Detailed balance requires that the flow and counter flow between every pair of states are equal. The stationary state will then be the steady state and the time the chain spends in a given state will be proportional to $f(\boldsymbol{x})$.

The HM algorithm is as follows. Starting at state $\boldsymbol{x}_n$

1. Choose a new trial point $\boldsymbol{x}_t$ from a **proposal distribution** $q(\boldsymbol{x}_t|\boldsymbol{x}_n)$.

2. Calculate

$$\alpha(\boldsymbol{x}_t, \boldsymbol{x}_n) = \min\left\{1, \frac{q(\boldsymbol{x}_n|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_n)}\frac{f(\boldsymbol{x}_t)}{f(\boldsymbol{x}_n)}\right\} \tag{13.5.4}$$

3. If $\alpha < 1$ draw a uniform deviate between 0 and 1. If α is large than this number accept the trial state and set $\boldsymbol{x}_{n+1} = \boldsymbol{x}_t$. Otherwise set $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n$. In other words, accept the trial state with probability $\alpha(\boldsymbol{x}_t, \boldsymbol{x}_n)$.

4. repeat

In this case we can easily see how detail balance is satisfied by this algorithm

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) = q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)\alpha(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n)f(\boldsymbol{x}_n) \tag{13.5.5}$$

$$= \begin{cases} q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) & , \quad q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) < q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) \\ q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) & , \quad q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) > q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) \end{cases} \tag{13.5.6}$$

and

$$p(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) = q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})\alpha(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) \tag{13.5.7}$$

which will be the same as above in the same cases so (13.5.3) is satisfied.

13.5.2 choosing a proposal distribution

Although the MCMC is guaranteed to converge under the conditions mentioned above, it might take a *very long time.* Like age of the Universe long time if your not careful. The chain moves around parameter space in a random walk and if it

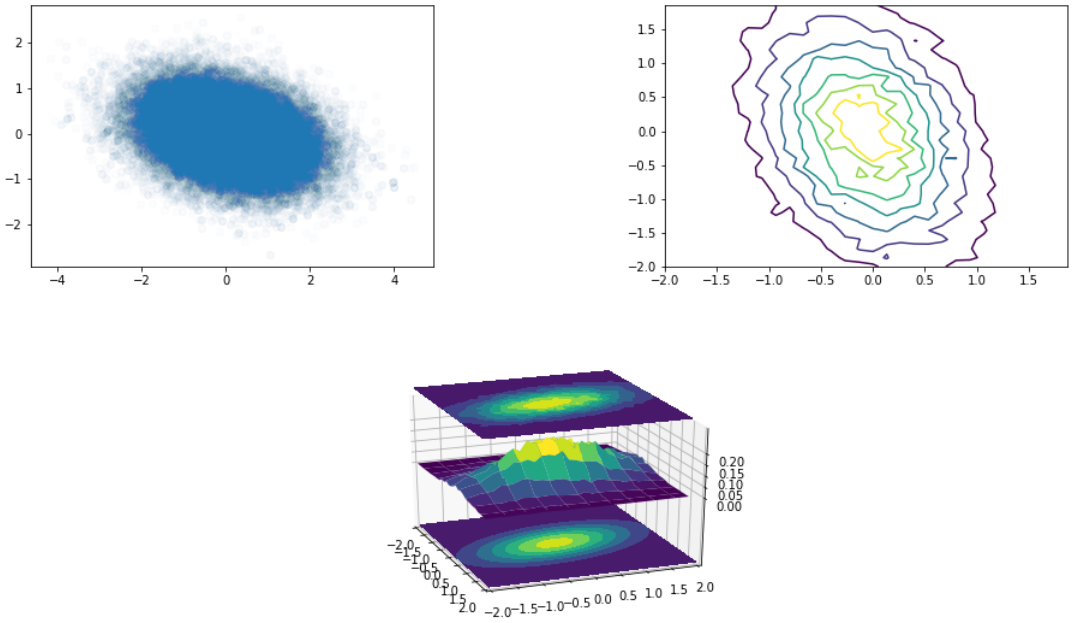

Figure 13.2: On the top left are the points from a Metropolis-Hastings Markov Chain for a simple 2 dimensional Gaussian. On the top right is a contour plot of the 2d histogram of those points. Below is the target distribution and two representations of the histogram.

does not reach every region of significant probability many times it will not be a good approximation of an independent sampling from the stationary distribution. To achieve good *mixing* the **rejection rate** of proposed moves must not be too high or too low. If it is too high the chain will have many duplicated points that will not fill parameter space in an even way. If the rejection rate is too low the chain will move, but not fast enough to get around the space. A rule of thumb is that you want a rejection rate of about 80%, i.e. an acceptance rate of 20%. This rate can be changed by adjusting the proposal function $q(\boldsymbol{x}_t|\boldsymbol{x}_n)$.

There is a great deal of freedom in choosing a proposal function and finding the right one for a particular problem is a bit of an art. Often (as in the original algorithm) the proposal distribution is symmetric, $q(\boldsymbol{x}_t|\boldsymbol{x}_n) = q(\boldsymbol{x}_n|\boldsymbol{x}_t)$, so that it doesn't come into α at all. Since we need to sample from $q(\boldsymbol{x}_t|\boldsymbol{x}_n)$ it makes sense to use a standard distribution with a well implemented random deviate generator. A popular choice is the mulitvariant Gaussian centered on the current point so $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{y}$ where $\boldsymbol{y}$ is samples from a multivariate Gaussian. But the σ's (or the covariance matrix $\boldsymbol{C}$) is not specified. These variances need to be adjusted until an acceptable rejection rate is found. Reducing σ's tends to decrease the rejection rate. When the σ's are large steps tend to put the proposed new point into regions that are far away from a peak in the probability and thus are rejected.

There is nothing stopping one from making steps in one dimension at a time as long as all dimensions are eventually explored. One can for example cycle through the parameters or pick a parameter at random each step. Sometimes this can improve the convergence.

To initialize the chain one must guess a point in parameter space. This will usually not be a place of high probability unless you are a good guesser. The chain will be attracted by the high probability regions (assuming there is some gradient in the posterior which might not be the case because of numerical underflow problems), but might take a while to get there. During this **burn in period** the chain is not near its stationary distribution. For this reason one usually discards the first part of the chain. There is no perfect method for determining how long the burn in period should be. You can look at a plot of the parameters vs steps and usually, but not always, it moves rapidly across parameter space and then settles into some location like in figure 13.6). Other times the maximum of the distribution can be found by some minimization technique such as was discussed in section **??** and then MCMC is used to map the posterior to find variances and covariances. If the chain starts near a maximum it might not be necessary to discard a burn in period. If this is not the case it is often useful to use a proposal distribution that jumps further during this burning stage while the chain is searching for the peak(s) and then reduce the jumps later for a new chain to get an acceptable rejection rate.

The biggest difficulties with MCMC arise when:

- The *initial guess* is so far from any peaks and the probability is so flat out there that the chain never finds a peak. It is sometimes the case that in low probability regions the calculation of $f(\boldsymbol{x})$ has a numerical underflow error or is dominated by numerical noise in which case the chain may wander around without getting anywhere.

- The *parameters are degenerate*. Imagine a distribution $f(\boldsymbol{x})$ that has a narrow ridge. If the proposal distribution is isotropic it will be either too wide in one direction so that the rejection rate is too high or it will be too small and creep along the ridge very slowly. In the case of a linear degeneracy you might be able learn something about the distribution and then make your proposal distribution anisotropic in a way that improves convergence. A nonlinear degeneracy is much more of a problem. In this case a proposal distribution that works well at one point in space will not work well at another. Imagine a $f(\boldsymbol{x})$ that is a function of $x_1^2 + 2x_2^2$ (subscripts are parameters not stages in the chain here). The "peak" or ridge will be an ellipse. If the ridge is very narrow a good proposal distribution will be narrow in the direction perpendicular to the ridge, but this is not the same direction everywhere. When possible, one should try to eliminate known nonlinear degeneracies by changing the variables. (For example $p = x_1^2 + 2x_2^2$ would be a better variable to use in this toy example.)

- The distribution has *multiple modes*. This is probably the hardest problem to deal with. If there are multiple peaks in the distribution that are separated by regions of low probability then the chain can easily get caught in one peak where its probability of transitioning to the other is very small. You might adjust the proposal distribution to get a good rejection rate for one peak, but that might make the probability of jumping between peaks effectively zero (see figure 13.4). This problem is exacerbated in large dimensional space because peaks that might not seem to be far away by their Euclidean distance, d are in a volume that goes up like d^D where D is the dimension of space. A defense against this is to run multiple chains with different random initial states and see if they find different modes.

Problem 56. *Gibbs Sampling: Say there are k parameters. At each step only one parameter is updated. The current parameter to be updated will be $x^{(i)}$. Show that the rejection rate will be zero ($\alpha(\boldsymbol{x}_t, \boldsymbol{x}_n) = 1$) for the the proposal function*

$$q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n) = f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)}) \tag{13.5.8}$$

where $\boldsymbol{x}^{(i-)}$ are all the other parameters that are not being updated this step and $f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})$ is the conditional target distribution.

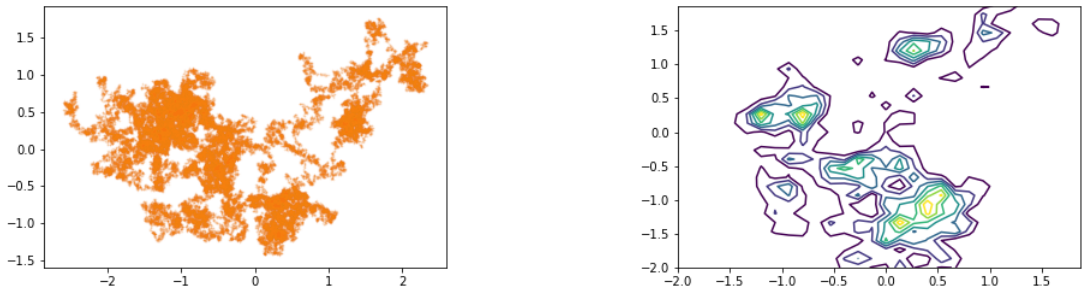

Figure 13.3: This is the same simple 2D Gaussian as shown in figure 13.2, but here the width of the proposal distribution was chosen to be too small ($\sigma = 0.01$ here and 0.5 there).

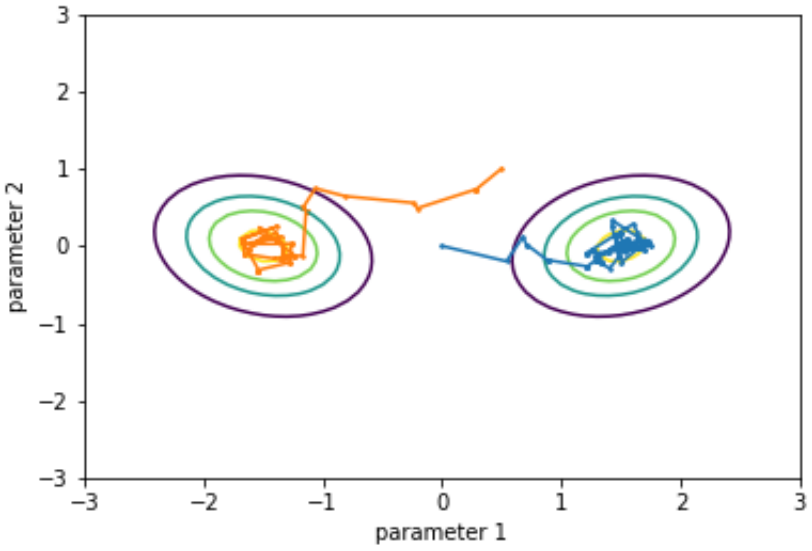

Figure 13.4: Two chains started at different initial points in a multimodal distribution. In this case the chance of jumping between modes within a single chain is small.

The above problem shows that a special choice for the proposal function can result in zero rejections. **Gibbs sampling** can be much faster other forms of MH because of this property, but the catch is that you need to be able to sample efficiently from $f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})$. In certain circumstances one might know this conditional probability, but not be able to calculate properties of the joint probability analytically. To my knowledge this situations doesn't come up often in statistical inference, but does in modeling physical or social processes.

13.5.3 example

Figures 13.5 through 13.8 show an example. The (fake) data is shown in 13.5. The likelihood function is Gaussian with varying but known σ's. The model here is $y = p_1(xp_2)^2 + 3(p_1 + p_2)^2$. Figure 13.6 shows the burn in period. You can see that the chain seems to have been attracted to some steady state solution. Figure 13.7 shows a chain 100,000 points long. The acceptance rate for this chain was 21.6%. An isotropic Gaussian proposal distribution was used with $\sigma = 0.3$.

13.5.4 convergence

It is critical that one knows when the chain has converged. Unfortunately there is no fool proof ways to determine this. One thing you can do is calculated the autocorrelation for each of the parameters as a function of the *lag*, m, the separation in the chain. It can be defines as

$$C_{\alpha,\beta}(m) = \frac{\sum_{i=1}^{N-m}(\alpha_i - \overline{\alpha})(\beta_{i+m} - \overline{\beta})}{\sqrt{\left(\sum_{i=1}^{N-m}(\alpha_i - \overline{\alpha})^2\right)\left(\sum_{i=m}^{N}(\beta_i - \overline{\beta})^2\right)}} \tag{13.5.9}$$

where α and β are parameter values. In the case of the autocorrelation $\alpha = \beta$. $C_{\alpha,\beta}(0) = 1$ by construction. Distant points along the chain should not be correlated so this function should oscillate about zero for large lag, m. The first time this function drops to zero or near zero is an estimate of the **correlation length**. Let us call this N_{corr}. Points separated by less than the correlation length will not be independent. You can define an effective number of independent samples in the chain as

$$N_{eff} = \frac{N_{chain}}{N_{corr}} \tag{13.5.10}$$

We want this number to be large. We also want any statistic we are interested in to depend on a number of points that is much larger than N_{corr}. For example the

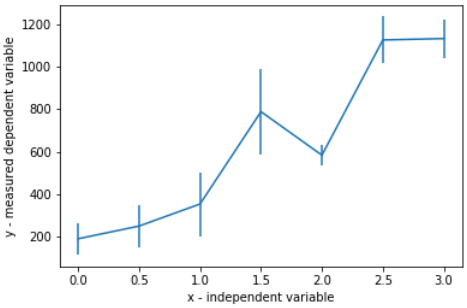

Figure 13.5: Simulated data.

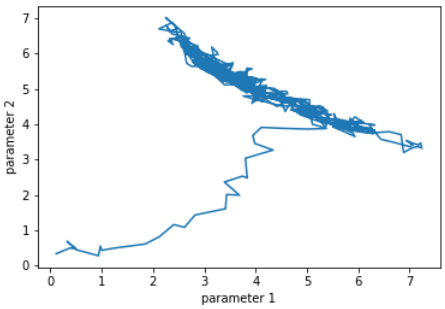

Figure 13.6: The first 1,000 steps of the MCMC starting at an initial guess $(0, 0)$.

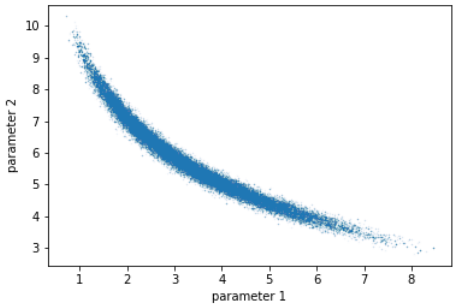

Figure 13.7: The 100,000 steps after discarding the first 1,000.

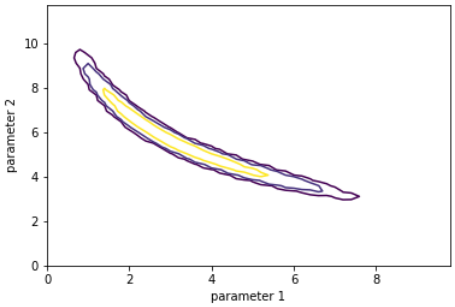

Figure 13.8: Contours surrounding 68%, 95% and 99% of the points.

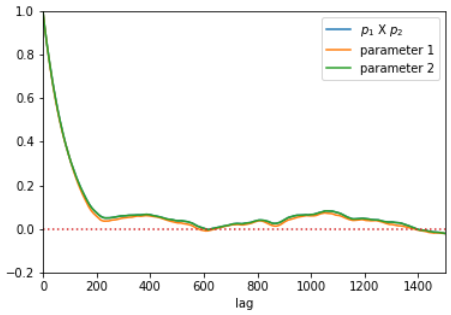

Figure 13.9: The correlation coefficient as a function of lag in the MCMC chain. Shown are the autocorrelation for the two parameters and the cross-correlation between them.

difference between the 95% and 99% contour levels depend on only 4% of the particles. This might be smaller than N_{corr}.

Figure 13.9 shows the correlation function for the example given above. You can see that the value of N_{corr} is not precisely defined and this curve is a bit different every time the calculation to run over. We can say $N_{corr} \sim 200-600$ to be conservative. This means that our length 100,000 chain actually only has about 500 to 200 effectively independent samples.

In practice this criterion can be fooled. For example a chain that is caught in one mode of a multimodal distribution might appear to be converging nicely.

A somewhat more sophisticated method that takes into account multiple chains is the **Gelman-Rubin diagnostic** $\hat{R}$ (Gelman & Rubin (1992)). If we have m independent chains each of length n and θ_i^α is the ith parameter value of the αth chain we can define the following quantities:

$$\overline{\theta}^\alpha = \frac{1}{n}\sum_i^n \theta_i^\alpha \qquad \overline{\overline{\theta}} = \frac{1}{m}\sum_\alpha^m \overline{\theta}^\alpha \tag{13.5.11}$$

$$s_\alpha^2 = \frac{1}{n-1}\sum_i^n (\theta_i^\alpha - \overline{\theta}^\alpha)^2 \qquad B = \frac{n}{m-1}\sum_\alpha^m \left(\overline{\theta}^\alpha - \overline{\overline{\theta}}\right)^2 \tag{13.5.12}$$

$$W = \frac{1}{m}\sum_\alpha^m s_\alpha^2 \qquad V = \frac{n-1}{n}W + \frac{M+1}{nm}B \tag{13.5.13}$$

$$\hat{R} = \sqrt{\frac{V}{W}} \tag{13.5.14}$$

$\hat{R}$ is an estimate of the factor by which the variance in θ can be reduced by continuing the chains. A $\hat{R} \sim 1$ is a good sign. This should be done for all the parameters of interest.

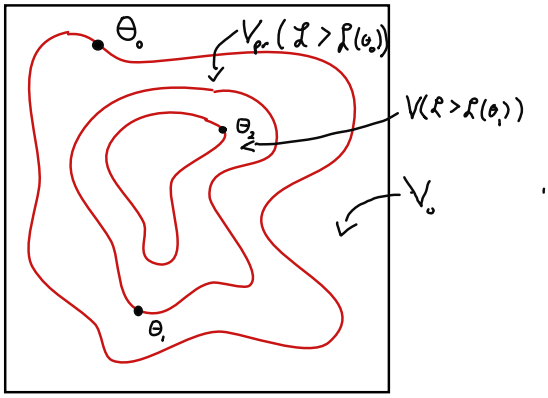

Figure 13.10: Nested sampling starts with a random set of points in parameter space. The average volumes between the contours are equal.

13.5.5 variations

There are many variations to the basic HM MC algorithm such as *Differential Evolution MCMC* or DEMCMC (ter Braak 2006, ter Braak & Vgurt 2008, Nelson et al. 2014), *Affine-Invariant Ensemble MCMC* (Goodman & Weare 2010, Foreman-Mackey et al. 2013) , *Hamiltonian sampling MCMC*, *Gibbs sampling* (see problem 56) and Parallel Tempering MCMC (Gregory, 2006). These try to ameliorate the basic problems with MCMC – adjusting the proposal function to fit the problem, dealing with degeneracy and multimodality. Some of them involve multiple chains that run in parallel and communicate with each other and/or they have adaptive ways of finding better proposal distributions. Many implementations of these algorithms can be found on the internet. For a review of MCMC methods see Neal (1998) (https://www.cs.toronto.edu/ radford/res-mcmc.html).

Problem 57. *You do an MCMC calculation for parameters θ_1 and θ_2 using a Gaussian proposal function. You now decide that it would be better if the first parameter were* $\ln(\theta_1)$. *You change the proposal function to be a Gaussian in* $\ln(\theta_1)$. *You change nothing else. Do you expect to get a different answer for the summery statistics?*

13.6 nested sampling & calculation of evidence

Another numerical technique that has become widely used for solving the Bayesian inference problem and in astrophysics in particular is **nested sampling**. The application of this to the inference problem is due to Skilling (2004a) (see also the book Silvia & Skilling (2006)).

Nested sampling is primarily a Monte Carlo integration technique applied to calculating the evidence. You will recall that the evidence is

$$\mathcal{E} = \int d^n\theta \; \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{13.6.1}$$

For the moment let's take the prior $\pi(\boldsymbol{\theta})$ to be uniform, but restricted to a finite volume in parameter space. Let's find N_a random points in this volume using a standard random number generator $\{\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_{N_a}\}$. Now we evaluate the likelihood at each of the points and sort them so that $\mathcal{L}(\boldsymbol{\theta}_1) < \mathcal{L}(\boldsymbol{\theta}_2) < \cdots < \mathcal{L}(\boldsymbol{\theta}_n)$. We know from our study of the extreme values (section 4.5) and Monte Carlo confidence levels (section 13.2) that the probability (in this case proportional to the volume) a new point having $\mathcal{L}(\boldsymbol{\theta}) < \mathcal{L}(\boldsymbol{\theta}_1)$ can be estimated as $P(\mathcal{L}(\boldsymbol{\theta}) < \mathcal{L}(\boldsymbol{\theta}_1)) \simeq 1/N_a$. Or the volume (probability) with a larger likelihood is

$$V_{pr}(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1)) \simeq \left(1 - \frac{1}{N_a}\right) V_o \tag{13.6.2}$$

where V_o is the initial volume of the parameters space. Now let's pick another random point from the volume but accept it only if its likelihood is larger then minimum previously found, $\mathcal{L}(\boldsymbol{\theta}_1)$. Once we have found a good point we discard $\boldsymbol{\theta}_1$, add the new point to the list and resort them. The new lowest point might now be the old $\boldsymbol{\theta}_2$ or it might be the new point. Now we can apply the same argument to find an estimate of the volume with $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}(\boldsymbol{\theta}_1)$, but now with the condition that all the points are required to be within the volume V_{pr} so the new volume is $V_{pr}^{(2)} = V_{pr}(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1)) \simeq \left(1 - \frac{1}{N_a}\right) V_{pr}$. If we continue to do this we will in the nth cycle get an estimate for the volume of

$$V_{pr}(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1^n)) = V_{pr}^n \simeq \left(1 - \frac{1}{N_a}\right)^n V_o \tag{13.6.3}$$

Where $\boldsymbol{\theta}_1^n$ is the point in the set of N_a points with the smallest likelihood after n steps. We store all the $\boldsymbol{\theta}_1^n$'s and $\mathcal{L}_n \equiv \mathcal{L}(\boldsymbol{\theta}_1^n)$.

The volume in parameter space (or probability according to the prior) associated with the likelihood $\mathcal{L}_n$ can be found by interpolation

$$v_n = \frac{1}{2}\left(V_{pr}^{n-1} - V_{pr}^{n+1}\right) \tag{13.6.4}$$

Using this we can estimate the evidence as

$$\mathcal{E} \simeq \sum_{n=1}^{M} \mathcal{L}_n v_n \tag{13.6.5}$$

where M is the total number of cycles used. Typically the calculation is continued until new cycles changed $\mathcal{E}$ by less than the desired accuracy.

Any expectation value for any function of the parameters can then be estimated with

$$E\left[f(\boldsymbol{\theta})\right] \simeq \frac{\sum_{n=1}^{M} f(\boldsymbol{\theta}_1^n)\mathcal{L}_n v_n}{\mathcal{E}} \tag{13.6.6}$$

The approximation is often made that

$$V_{pr}^n = \exp\left[\ln(V_{pr}^n)\right] \tag{13.6.7}$$

$$= \exp\left[n \ln\left(1 - \frac{1}{N_a}\right)\right] V_o \tag{13.6.8}$$

$$\simeq \exp\left[-\frac{n}{N_a}\right] V_o \tag{13.6.9}$$

so

$$v_n \simeq \frac{1}{2} e^{-\frac{n}{N_a}} \left(e^{+\frac{1}{N_a}} - e^{-\frac{1}{N_a}}\right) V_o \tag{13.6.10}$$

$$\simeq \frac{e^{-\frac{n}{N_a}}}{N_a} V_o \tag{13.6.11}$$

You can see that the volume goes exponentially down with n.

What I have called volume here (V_{pr}^n and v_n) could just as well be called the probability according to the prior. If the prior is not uniform then the algorithm works just the same as long as the random points are drawn from the prior. This might be possible using a standard numerical library or in some cases a MCMC is used for this.

Feroz & Hobson (2008) show that an estimate of the variance in $\ln \mathcal{E}$ is

$$\sigma^2_{\ln \mathcal{E}} \simeq \frac{`H}{N_a} = \frac{1}{N_a} \sum_{n=1}^{M} \frac{\mathcal{L}_n v_n}{\mathcal{E}} \ln\left(\frac{\mathcal{L}_i}{\mathcal{E}}\right) \tag{13.6.12}$$

where H is an estimate of the *relative entropy* (more on this later).

13.6.1 optimization

So far the nested sampling algorithm automatically zooms in exponentially on regions of high posterior in parameter space and can estimate the integrals in high dimensional space without gridding or making assumptions about the form of the posterior. But as it zooms in, it becomes exponentially less efficient since most of the points that are drawn randomly from the prior will not have likelihoods that are above the current minimum.

The fix for this is to draw the points from a smaller and smaller region that always contains the entire region with $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_n$. One popular software package that does this is called *multinest* (Feroz & Hobson, 2008). In this case points are drawn uniformly from inside an ellipsoid that shrinks around the active points. The difficulty is keeping the ellipsoid from shrinking too quickly and cutting off some of the high $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_n$ region while at the same time shrinking it quickly enough to make the algorithm efficient. There is typically a few parameters involved with this that require adjustments along with the number of active points. Drawing a point from within an ellipsoid in D dimensional space can be done efficiently by drawing a point from inside a D-sphere by the method given in section 13.1 and then stretching it with the axis ratios of the ellipsoid.

A more recent publicly available implementation of nested sampling is called PolyCord (Handley et al., 2015) which uses a slice sampling strategy to improve efficiency.

13.7 Simulated Annealing

Consider the following family of distributions

$$p_\lambda(\boldsymbol{\theta}|\boldsymbol{D}) \propto \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{D})^\lambda \pi(\boldsymbol{\theta}) \tag{13.7.1}$$

where λ is a real value ≥ 0. You can see that $\lambda = 0$ corresponds to the prior and $\lambda = 1$ corresponds to the posterior. A strategy for improving the efficiency of an MC chain might be to start chains with $\lambda = 0$ where it is usually easy to run an efficient chain and then slowly increase λ through a series of values λ_n while letting the chain equilibrate for each value. In this way the points will be attracted to the regions of high posterior values while at the same time hopefully hopping out of false maxima before the chain gets stuck in them. If we continue beyond $\lambda = 1$ toward ∞ we can even use this to find the maximum.

This procedure helps in finding maximum of the posterior, but the resulting chain is only guaranteed to converge to $p(\boldsymbol{\theta}|\boldsymbol{D})$ in the same way that the final chain with $\lambda = 1$ is guaranteed to do so. Alone it might help reduce the burn in period, but in some cases it might require more evaluations than the burn in period would have required.

However, there are some interesting variations on this theme which are guaranteed to converge to the posterior and generally make the sampling more efficient. They generally calculate weights for each sampled point as they progress through the λ_n's. The BayesSys algorithm (Skilling, 2004b) and Annealed Importance Sampling (Neal, 2001) combine aspects of simulated annealing with importance sampling and Markov Chains. These methods can also be used to calculate the evidence as well as the posterior.

13.7.1 statistical physics analog

As the name suggests, there is an interesting correspondence between simulated annealing and statistical physics. If we define energy of state $\boldsymbol{\theta}$ as

$$E(\boldsymbol{\theta}) = -\ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{D})\pi(\boldsymbol{\theta}) \tag{13.7.2}$$

then by (13.7.1)

$$p_\lambda(\boldsymbol{\theta}) \propto e^{-\lambda E(\boldsymbol{\theta})} = e^{-\frac{E(\boldsymbol{\theta})}{k_B T}}. \tag{13.7.3}$$

We can recognize this as the Gibbs distribution if we make the identification $\lambda = 1/(k_B T)$ with k_B being Boltzmann's constant, as is done in the last equality. So we can think of $p_o(\boldsymbol{\theta})$ ($\lambda = 0$) as the distribution of states at infinite temperature. As $\lambda \to 1$ the temperature is reduced and the thermal energy plays a smaller part in the distribution. $\lambda = \infty$ corresponds to $T = 0$, the classical ground state.

The analogy can be stretched further by associating the evidence

$$\mathcal{E}[\lambda] = \int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{D})^\lambda \pi(\boldsymbol{\theta}) \propto \int d\boldsymbol{\theta}\ p_\lambda(\boldsymbol{\theta}|\boldsymbol{D}) \tag{13.7.4}$$

with the canonical partition function.

13.8 Approximate Bayesian Computation (ABC)

Another method that has gotten some attention recently is ABC also known as "Likelihoodless Bayesian Inference" (Akeret et al. (2015); Kacprzak et al. (2018) and references there in for example). This is for the case where the likelihood is not known in closed form, but one can simulate data sets given a set of parameters. This approach of simulating data from parameters is often called **forward modeling** as apposed to going from data to parameters through a likelihood function.

In this case we have some method of simulating data given a set of parameters, $\boldsymbol{\theta}$. This could be a cosmological n-body simulation or a simulation of an image or

spectra that includes additive and multiplicative noise, non detections, biases etc. A single set of parameters will not correspond to a uniquely data set so every time the simulation is run you get a different data set, $\boldsymbol{D}^*$ even for the same $\boldsymbol{\theta}$.

We would expect that if the parameters are "near" their true values the simulated data, $\boldsymbol{D}^*$ will be "near" the observed data $\boldsymbol{D}$. Let us invent some measure of distance between data set which we will denote $\rho(\boldsymbol{D}, \boldsymbol{D}^*)$. The number of simulations that land within $\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon$ would be proportional to the probability. It follows that

$$\frac{p\left(\boldsymbol{\theta}_1|\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon\right)}{p\left(\boldsymbol{\theta}_2|\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon\right)} = \frac{p\left(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon|\boldsymbol{\theta}_1\right)\pi(\boldsymbol{\theta}_2)}{p\left(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon|\boldsymbol{\theta}_2\right)\pi(\boldsymbol{\theta}_1)} \simeq \frac{N_{\boldsymbol{\theta}_1}(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon)}{N_{\boldsymbol{\theta}_2}(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon)}\frac{N_{\boldsymbol{\theta}_2}}{N_{\boldsymbol{\theta}_1}} \quad (13.8.1)$$

where Bayes' theorem is used for the first equality, $N_{\boldsymbol{\theta}}$ is the number of simulations run with parameters $\boldsymbol{\theta}$ and $N_{\boldsymbol{\theta}}(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon)$ is the number of those simulations whose data set satisfies the constraint. The second equality requires that the parameters are sampled according to the prior, i.e. $N_{\boldsymbol{\theta}_2}/N_{\boldsymbol{\theta}_1} \sim \pi(\boldsymbol{\theta}_2)/\pi(\boldsymbol{\theta}_1)$.

If the parameters are sampled according to the prior then we can make the approximation

$$p\left(\boldsymbol{\theta}|\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon\right) \sim \frac{N_{\boldsymbol{\theta}}(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon)}{N(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon)} \sim \frac{1}{N(\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon)}\sum_i \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i^*) \quad (13.8.2)$$

where the $\boldsymbol{\theta}_i^*$ are the parameter sets that resulted in data sets within ϵ of the observed data. If ϵ is very small one expects this probability to convergence to the true posterior. Symbolically

$$p\left(\boldsymbol{\theta}|\rho(\boldsymbol{D}, \boldsymbol{D}^*) < \epsilon\right) \xrightarrow{\epsilon \to 0} p(\boldsymbol{\theta}|\boldsymbol{D}) \quad (13.8.3)$$

You can imagine doing a huge number of simulations and then throwing away all the ones with $\rho(\boldsymbol{D}, \boldsymbol{D}^*) > \epsilon$. If ϵ is small enough the $\boldsymbol{\theta}$ values of these remaining simulations should be distributed according to $p(\boldsymbol{\theta}|\boldsymbol{D})$. In practice this is very inefficient. As we have seen, the target volume in parameter space can be very small relative the total parameter space volume. In addition, in this case the data space has an even higher dimensionality and the region within ϵ of the observed data is usually even smaller. For this reason ABC is usually avoided whenever it is possible to write down an accurate likelihood.

To mitigate these problems the strategy is usually to start with a large ϵ and then reduce it while simultaneously reducing the volume in $\boldsymbol{\theta}$-space from which the prior is sampled. You can look at some of the methods for doing this in the references cited above.

What is $\rho(\boldsymbol{D}, \boldsymbol{D}^*)$? The most obvious choice would be a least-squares or Euclidean distance type cost function,

$$\rho(\boldsymbol{D}, \boldsymbol{D}^*) = \sum_i (d_i - d_i^*)^2 \quad (13.8.4)$$

In practice one often does not required a point-by-point matching of simulated data and real data because this reduces the acceptable volume in data space. Often some statistics of the data are calculated and $\rho(\boldsymbol{D}, \boldsymbol{D}^*)$ is constructed out of them.

For example, you might be simulating the distribution of matter in the universe and be interested in models that match the observed power spectrum, but it is not required that there be a galaxy cluster at the same place as in the real galaxy survey. For such a case one might use

$$\rho(\boldsymbol{D}, \boldsymbol{D}^*) = \sum_k \frac{(P_k - P_k^*)^2}{P_k^2} \tag{13.8.5}$$

where P_k is the power spectrum of the data and the sum is over the measured Fourier modes.

Chapter 14

Information and entropy

> [I]n studying probability theory, it was vaguely troubling to see reference to "gaussian random variables", or "stochastic processes", or "stationary time series", or "disorder", as if the property of being gaussian, random, stochastic, stationary, or disorderly is a real property, like the property of possessing mass or length, existing in Nature. Indeed, some seek to develop statistical tests to determine the presence of these properties in their data...
>
> Once one has grasped the idea, one sees the Mind Projection Fallacy everywhere; what we have been taught as deep wisdom, is stripped of its pretensions and seen to be instead a foolish non sequitur. The error occurs in two complementary forms, which we might indicate thus: (A) (My own imagination) $\rightarrow$ (Real property of Nature), (B) (My own ignorance) $\rightarrow$ (Nature is indeterminate)

Jaynes [E. T. JAYNES (1989). PROBABILITY THEORY AS LOGIC Ninth Annual Workshop on Maximum Entropy and Bayesian Methods. pp. 1–16.

14.1 information content of data

Shannon, studying signal processing, wanted to find an expression for the *information* in a digital massage. The idea is that certainty about the state of a system/message implies known information. If we have only a distribution for the state of a system we have less information about it than if we knew the exact state, i.e. the case where the probability of state i is $p_i = 1$. Shannon in 1948 tried to quantify the amount of information that is required to go from a particular distribution to certainty. He argued that a measure of information should satisfied the following axioms.

Shannon's Axioms:

- Axiom I : S is a real continuous function of the probabilities p_i, $S[p_1, \ldots, p_m]$.
- Axiom II : If all p_i's are equal, $p_i = 1/m$, then $S[1/m, 1/m, \ldots, 1/m]$ is an increasing function of m. If all states are equally probable increasing the number of states increases the uncertainty.
- Axiom II: The grouping property. For all possible inclusive groupings $g = 1 \ldots N$ of the states $i = 1 \ldots n$ we must have

$$S = S\left[\{P\}\right] + \sum_g P_g S_g \tag{14.1.1}$$

  where

$$P_g = \sum_{i \in g} p_i \tag{14.1.2}$$

Surprisingly there is a unique functional that satisfies these requirements called the **Shannon entropy** given by

$$S[p] = -\sum_i^m p_i \ln p_i \tag{14.1.3}$$

If the sign is changed it is called the Shannon **information**. This is considered the beginning of information theory, but as we will see, and you probably know, this expression had already been found in a different context.

The third axiom is controversial and to many people not satisfactory. Other entropies are possible and have been proposed. In a minute we will see that a variation of the Shannon entropy that is often equivalent in practice has a much more satisfying justification to my mind.

Some properties of $S[p]$ are:

- $S[p] \geq 0$
- For the uniform probability case $p_i = 1/m$, the maximum ignorance case, $S[1/m, \ldots, 1/m] = \ln(m)$.
- In the case of complete certainty one of the probabilities will be 1 and all the others 0. In this case $S[1, 0, \ldots, 0] = 0$.
- If there are two variable x and y with joint probability $p(x, y)$ there entropy is

$$S_{xy} = -\sum_{xy} p(x, y) \ln p(x, y) \tag{14.1.4}$$

and if they are independent $p(x,y) = p(x)p(y)$ so

$$S_{xy} = S_x + S_y \quad \text{independent variable} \tag{14.1.5}$$

and in general

$$S_{xy} \leq S_x + S_y \tag{14.1.6}$$

You can define the analogous entropy of a continuous distribution,

$$S[p] = -\int_{-\infty}^{\infty} dx\ p(x) \ln\left(p(x)\right) \tag{14.1.7}$$

This entropy has the important flaw that it is not coordinate invariant. If the coordinates are changed to $y = f(x)$ this should not change the amount of information contained in the distribution. We will see later how this can be resolved, but for now we will take this to be the entropy of a continuous distribution.

14.1.1 the maximum entropy principle for choosing a distribution

The maximum entropy principle, sometime abbreviated MaxEnt, holds that *the best distribution to use for a variable is the one that has the maximum entropy (or least information) subject to any prior constraints on the distribution.* The idea is that you should assume the least possible information beyond your constraints and this is quantified by the entropy, i.e. maximum ignorance.

The maximum entropy distribution with only the normalization constraint ($\sum_i^m p_i = 1$) is of course $p_i = 1/m$ and the maximum entropy is $S_{max} = \ln(m)$. This is the state of maximal ignorance or minimum information.

Now let us say that we have a constraint on the variance of a continues distribution, namely

$$Var_p[x] = \sigma^2 \tag{14.1.8}$$

where σ^2 is a constant. We can find the MaxEnt distribution using Lagrangian multipliers

$$\delta\left\{-\int_{-\infty}^{\infty} dx\ p(x)\ln p(x) - \lambda_o\left(\int_{-\infty}^{\infty} dx\ p(x) - 1\right) - \lambda_1\left(\int_{-\infty}^{\infty} dx\ x^2 p(x) - \sigma^2\right)\right\} = 0 \tag{14.1.9}$$

giving

$$-\ln p(x) - 1 - \lambda_o - \lambda_1 x^2 = 0 \tag{14.1.10}$$

or

$$p(x) = \exp(-1 - \lambda_o - \lambda_1 x^2) \tag{14.1.11}$$

$$= Ae^{-\lambda_1 x^2} \tag{14.1.12}$$

where A is a normalization constant. A and λ_1 are determined by the two constraints so

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{14.1.13}$$

The distribution with the maximum degree of ignorance subject to the constraint on the variance is the normal distribution. This is another reason to favor the normal distribution that has no apparent relation to the central limit theorem.

The entropy of a normal distribution is

$$S_{norm} = \frac{1}{2} + \frac{1}{2}\ln\left(2\pi\sigma^2\right). \tag{14.1.14}$$

Problem 58. *Say you have the constraint*

$$\sum_{i=0}^{\infty} ip_i = \mu \tag{14.1.15}$$

What is the MaxEnt distribution?

Problem 59. *Say you have the following constraints on a continuous distribution $p(x)$,*

$$p(x) = 0 \quad \forall x < 0 \quad \textit{and} \quad \langle x \rangle = \mu \tag{14.1.16}$$

What is the MaxEnt distribution?

14.2 Connection to Statistical Physics

Boltzman considered particles in a gas that can exist in momentum and position states. The number of ways to occupy m states with N particles

$$\Omega = \frac{N!}{n_1! \ldots n_m!} \tag{14.2.1}$$

with $\sum_i^m n_i = N$. We can recognize this as the normalization for the multinomial distribution that we have used many times. Using Sterling's approximation

$$\ln \Omega = N \ln N - N - \sum_i^m (n_i \ln n_i - n_i) \tag{14.2.2}$$

$$= N \ln N - \sum_i^m n_i \ln n_i \tag{14.2.3}$$

$$= N \left[\ln N - \sum_i^m \frac{n_i}{N} \left(\ln \frac{n_i}{N} + \ln N \right) \right] \tag{14.2.4}$$

$$= -N \sum_i^m \frac{n_i}{N} \ln \left(\frac{n_i}{N} \right) \tag{14.2.5}$$

$$= -N \sum_i^m p_i \ln p_i \tag{14.2.6}$$

if $p_i = n_i/N$. You can see the clear resemblance to Shannon's entropy.

Gibbs changed the interpretation of p_i to be the probability of a collective state i where the particles (spins, molecular species, etc.) are not necessarily independent and not the single particle occupation of single particle states (Gibbs, 1902). The constant in front of the entropy is a matter of convention. In statistical physics the entropy is often defined with Boltzmann's constant, k_B, in front. In information theory the base 2 logarithm is often used in which case the entropy has units of "bits".

We have a series of constraints on the expectation values of the form

$$\left\langle f^k \right\rangle = \sum_i^m f_i^k p_i = F^k \tag{14.2.7}$$

For example, F^k could be the average energy $\bar{E}$ in which case $f_i = \epsilon_i$ the energy of a specific state i. Or f_i^k is the number of particles or molecules of a certain type in state i then F^k will be the average number of those species. Or F^k could be the magnetic polarization, etc. The state of the system is labeled by the F^k's.

The canonical distribution is found by maximizing the entropy

$$-\sum_i^m p_i \ln p_i - \lambda_o \left(\sum_i^m p_i - 1 \right) - \lambda_k \left(\sum_i^m f_i^k p_i - F^k \right) = const. \tag{14.2.8}$$

$$-\ln p_i - 1 - \lambda_o - \lambda_k f_i^k = 0 \tag{14.2.9}$$

or

$$p_i = e^{-\lambda_o - 1} e^{-\lambda_k f_i^k} \tag{14.2.10}$$

$$p_i = \frac{e^{-\lambda_k f_i^k}}{Z} \qquad Z = \sum_i e^{-\lambda_k f_i^k} = e^{\lambda_o - 1} \tag{14.2.11}$$

You can recognize Z as the partition function and see that

$$\frac{\partial \ln Z}{\partial \lambda_k} = -F^k \tag{14.2.12}$$

The maximum entropy is

$$S_{max} = -\sum_i p_i \ln p_i \tag{14.2.13}$$

$$= -\sum_i \frac{e^{-\lambda_k f_i^k}}{Z} \left(-\lambda_k F^k - \ln Z\right) \tag{14.2.14}$$

$$= \lambda_k F^k + \ln Z \tag{14.2.15}$$

$$\frac{\partial S_{max}}{\partial F^k} = \frac{\partial \lambda_i}{\partial F^k} F^i + \lambda_k + \frac{\partial \ln Z}{\partial \lambda_k} \frac{\partial \lambda_i}{\partial F^k} \tag{14.2.16}$$

$$= \frac{\partial \lambda_i}{\partial F^k} F^i + \lambda_k - F^k \frac{\partial \lambda_i}{\partial F^k} \tag{14.2.17}$$

$$= \lambda_k \tag{14.2.18}$$

For the canonical ensemble F^k is the average energy of the system, $\overline{E}$, f_i^k is the energy of state i, ϵ_i and

$$\frac{\partial S_{max}}{\partial \overline{E}} \equiv \frac{1}{k_B T} = \lambda_\epsilon \tag{14.2.19}$$

so the Lagrangian multiplier associated with the internal energy is the inverse of the temperature. In this case equation (14.2.15) becomes the perhaps familiar

$$S_{max} = \frac{\overline{E}}{k_b T} + \ln Z \tag{14.2.20}$$

The energy $\overline{E}$ (and the ϵ_i's) will also depend on the volume

$$\frac{\partial S_{max}}{\partial V} = \frac{\partial S_{max}}{\partial \overline{E}} \frac{\partial \overline{E}}{\partial V} = -\frac{P}{k_B T} \tag{14.2.21}$$

Likewise, the Lagrangian multiplier associated with the average number of a chemical species is the the chemical potential, μ_k/k_BT.

From (14.2.11), the canonical distribution is then

$$p_i \propto \exp\left[-\frac{\epsilon_i}{k_bT}\right] \tag{14.2.22}$$

The thermodynamic "entropy of the system" is actually the maximum of the entropies that are consistent with the constraints. From an information theory prospective, the thermodynamic entropy is the negative of the amount of information that is needed to specify a microstate given a specified macrostate which corresponds to a fixed average energy, volume and number of particles.

14.3 Maximum Entropy as a method of inference

In chapter 5 we studied the Bayesian method for updating a probability from prior information about the parameters $\pi(\boldsymbol{\theta})$ to a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$ given additional information coming from new data $\boldsymbol{d}$. Now let us consider an even wider concept of updating our knowledge that can take into account not only data, but also information in the form of constraints on the expectation values.

Our prior knowledge is expressed through the prior distribution $\pi(x)$. After we take into account new information we will have a posterior distribution $p(x)$. How do we choose this distribution? If we have preferences for different distributions we can rank them with a real valued functional $S[p,q]$. The preferred distributions will have larger $S[p,q]$. We can reasonably impose these three requirements on the functional $S[p,q]$,

Caticha's Axioms:

- Axiom 1 : *Locality* In the absence of information about some domain D the probability should not change, $p(x|D) = \pi(x|D)$.
- Axiom 2: *Coordinate invariance* $S[p,q]$ should remain the same when the coordinates are changed.
- Axiom 3: *Consistency for independent systems subsystems* When a system is composed of subsystems that are known to be independent, it should not matter whether the inference procedure treats them separately or jointly.

(Caticha, 2008)

Amazingly, just these three axioms lead to a unique functional,

$$S[p,q] = -\int dx\ p(x)\ln\left(\frac{p(x)}{\pi(x)}\right) \tag{14.3.1}$$

This is called the **relative entropy**.

Maximum Entropy Updating: *Given prior knowledge $\pi(x)$ and some new information, the best posterior $p(x)$ is the one that maximizes the relative entropy while being consistent with the new information.*

You can see the MaxEnt method as being an application of this to choosing a distribution with a uniform prior $\pi(x) = const.$ because in this case the relative entropy differs form the Shannon's entropy by a constant. Statistical physics can be seen as another application of this principle as well.

14.3.1 Bayesian inference as a special case

Bayesian inference is a special case of maximum relative entropy inference. To see this let us say that an experiment/observation results in the specific data values $\boldsymbol{d}$ out of the possible data values $\boldsymbol{x}$. The constraint requires that the posterior for the data and the parameter have the property

$$p(\boldsymbol{x}) = \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}, \boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{d}) \tag{14.3.2}$$

or

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x})p(\boldsymbol{\theta}|\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{d})p(\boldsymbol{\theta}|\boldsymbol{d}) \tag{14.3.3}$$

The relative entropy of the posterior to the prior is

$$S[p, q] = -\int_{-\infty}^{\infty} d\boldsymbol{\theta} d\boldsymbol{x}\ p(\boldsymbol{x}, \boldsymbol{\theta}) \ln\left(\frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}, \boldsymbol{\theta})}\right) \tag{14.3.4}$$

Applying Lagrange's method the following must be stationary with respect to variations in the distribution function

$$\delta\left\{S[p, q] - \lambda_o\left[\int_{-\infty}^{\infty} d\boldsymbol{\theta} d\boldsymbol{x}\ p(\boldsymbol{x}, \boldsymbol{\theta}) - 1\right] - \int_{-\infty}^{\infty} d\boldsymbol{x}\ \lambda(\boldsymbol{x})\left[\int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{x}, \boldsymbol{\theta}) - \delta(\boldsymbol{x} - \boldsymbol{d})\right]\right\} = 0 \tag{14.3.5}$$

Note that that the constraint must hold at every $\boldsymbol{x}$ so there are an infinite number of Lagrange multipliers. This implies

$$-\ln p + \ln q - 1 - \lambda_o - \lambda(\boldsymbol{x}) = 0 \tag{14.3.6}$$

or

$$p(\boldsymbol{\theta}, \boldsymbol{x}) = \pi(\boldsymbol{\theta}, \boldsymbol{x})e^{-1-\lambda_o}e^{-\lambda(\boldsymbol{x})} \tag{14.3.7}$$

$$= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x})e^{-\lambda(\boldsymbol{x})}}{Z} \tag{14.3.8}$$

where Z is a normalization constant. Now applying the constraint

$$\int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{\theta},\boldsymbol{x}) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ \frac{\pi(\boldsymbol{\theta},\boldsymbol{x})e^{-\lambda(\boldsymbol{x})}}{Z} = \frac{\pi(\boldsymbol{\theta})e^{-\lambda(\boldsymbol{x})}}{Z} = \delta(\boldsymbol{x}-\boldsymbol{d}) \tag{14.3.9}$$

Substituting into (14.3.8) gives

$$p(\boldsymbol{\theta},\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta},\boldsymbol{x})}{\pi(\boldsymbol{x})}\delta(\boldsymbol{x}-\boldsymbol{d}) \tag{14.3.10}$$

Finally

$$p(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} d\boldsymbol{x}\ p(\boldsymbol{\theta},\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta},\boldsymbol{d})}{\pi(\boldsymbol{d})} = \frac{\pi(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{d})} \tag{14.3.11}$$

Note that the Bayesian updating is not the result of Bayes' theorem. In a strict sense all three distributions on the right hand side are *prior* distributions.

This new maximum entropy updating allows for more freedom in what can be considered new information. In the Bayesian method new information comes from new data and a likelihood function, but new information can now come in the form of a constraint on the data or the parameters.

14.4 relative entropy

The relative entropy is also called the **Kullback–Leibler divergence** or distance. It is used in many contexts and has different interpretations. In an important sense it is more fundamental than the entropy. An important property is

$$S[p|q] \geq 0 \tag{14.4.1}$$

and $S[p|q] = 0$ only when $p(x) = q(x)$. As required by the axioms, $S[p|q]$ is invariant under transformations of the random variables x.

The relative entropy is often used as a measure of how distant two distribution are from each other. It is not actually a true distance however because it is not symmetric, $S[p,q] \neq S[p,q]$.

$S[p,q]$ can be interpreted as the amount of information gained when the distribution is update from $q(x)$ to $p(x)$. For example, the information a new experiment adds to our knowledge of some parameters can be quantified in this way. This can be useful in planning an experiment when many parameters are being measured.

Another situation in which this comes up is when a new posterior is found, perhaps in a high dimensional parameters space, and one wants to quantify how much extra

information has been gained by including the latest data. The relative entropy of the posterior with respect to the prior can be used to quantify this. It is not always clear whether particular constraints are a result of the new data or the (sometimes somewhat arbitrary) prior especially when significant parameter degeneracies exist. The relative entropy between the marginal posterior and prior can be used to quantify how much information about a particular parameter or subset of parameters has been gained.

For reference and to establish some intuition the relative entropy of two Gaussians is

$$S[p,q] = \frac{1}{2}\left[\ln\left(\frac{|\boldsymbol{C}_q|}{|\boldsymbol{C}_p|}\right) + \mathrm{tr}\left[\boldsymbol{C}_p(\boldsymbol{C}_q^{-1} - \boldsymbol{C}_p^{-1})\right] + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T\boldsymbol{C}_q^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right] \quad (14.4.2)$$

In univariant case this is

$$S[p,q] = \frac{1}{2}\left[\ln\left(\frac{\sigma_q^2}{\sigma_p^2}\right) + \frac{\sigma_p^2}{\sigma_q^2} - 1 + \frac{(\mu_p - \mu_q)^2}{\sigma_q^2}\right] \quad (14.4.3)$$

$$= \frac{1}{2}\left[\ln\left(\frac{\sigma_q^2}{\sigma_p^2}\right) + \frac{(\sigma_p^2 - \sigma_q^2) + (\mu_p - \mu_q)^2}{\sigma_q^2}\right] \quad (14.4.4)$$

The first part expresses a change in the variances or constraining power of the distributions and the second comes from a mismatch in the means of the distributions.

The relative entropy of the posterior to the prior is sometimes called the **surprise**

$$S = S[p(\boldsymbol{\theta}|D), \pi(\boldsymbol{\theta})] = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|D)\ln\left[\frac{p(\boldsymbol{\theta}|D)}{\pi(\boldsymbol{\theta})}\right] \quad (14.4.5)$$

$$= \int d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|D)\ln\left[\frac{\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathcal{E}(\boldsymbol{D})\pi(\boldsymbol{\theta})}\right] \quad (14.4.6)$$

$$= \langle \ln\left[\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta})\right]\rangle_{p(\boldsymbol{\theta}|D)} - \ln\mathcal{E}(\boldsymbol{D}) \quad (14.4.7)$$

This term is also used for other measures of how different the posterior is from the prior. This is a measure of the information gain that comes from the data.

In the special case where the likelihood is Gaussian, the prior is uniform over a volume V_π and the likelihood is very small on all the borders of this prior volume so that integrals over the posterior are not effected by it, the surprise reduces to

$$S = -\frac{d}{2}\left(1 + \ln(2\pi)\right) - \frac{1}{2}\ln|\boldsymbol{C}| + \ln V_\pi \quad (14.4.8)$$

while the evidence is

$$\ln\mathcal{E} = \ln\mathcal{L}^{max} + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{C}| - \ln V_\pi \quad (14.4.9)$$

14.5 equivalence of maximum likelihood distribution & minimum relative entropy

Let us say we have an experiment that has a categorical outcome. There are k possible outcomes. The number of observed events in category i is n_i. We have a model that predicts the probability of outcome i as $p_i(\boldsymbol{\theta})$. The parameters of this model are $\boldsymbol{\theta}$.

The likelihood is a multinomial distribution

$$\mathcal{L}(\{n_i\}|\boldsymbol{\theta}) = \frac{N!}{\prod_i^k n_i!} \prod_i^k p(\boldsymbol{\theta})^{n_i} \tag{14.5.1}$$

We can find the maximum likelihood by taking the derivative

$$\begin{aligned}
\frac{\partial}{\partial\theta} \ln \mathcal{L}(\{n_i\}|\boldsymbol{\theta}) =& \frac{\partial}{\partial\theta} \left[\ln N! + \sum_i^k n_i \ln p_i(\boldsymbol{\theta}) - \sum_i^k n_i! \right] && (14.5.2)\\
=& \sum_i^k n_i \frac{\partial}{\partial\theta} \ln p_i(\boldsymbol{\theta}) && (14.5.3)\\
=& 0 && (14.5.4)
\end{aligned}$$

Now let's look at this problem differently. The empirical distribution is $p_i = n_i/N$. We could look for the distribution within the family of distributions parameterized by $\boldsymbol{\theta}$ that is "closest" to the empirical distribution. To define "closest" we might use the minimum relative entropy or Kullback–Leibler distance between the empirical and trial distributions.

$$S\left[\frac{n_i}{N}, p_i(\boldsymbol{\theta})\right] = -\sum_i^k \frac{n_i}{N} \ln\left(\frac{n_i}{N p_i(\boldsymbol{\theta})}\right) \tag{14.5.5}$$

Its minimum occurs at

$$\frac{\partial}{\partial\theta} S = \frac{1}{N} \sum_i^k n_i \frac{\partial}{\partial\theta} \ln p_i(\boldsymbol{\theta}) = 0 \tag{14.5.6}$$

which is clearly the same solution as for the maximum likelihood.

In machine learning the relative entropy is often used as a cost function. In this context one has a training set $\{\boldsymbol{y}, \boldsymbol{x}\}$ where $\boldsymbol{x}$ are the feature or independent variables and $\boldsymbol{y}$ are the dependent variables. The model gives a probability $p_i(\boldsymbol{\theta}, \boldsymbol{x})$. The relative entropy is minimized to find the best $\boldsymbol{\theta}$. We can see now that this is equivalent to finding the maximum likelihood solution.

I have presented this as for categorical dependent variables, but you can see that it works fine for the continuous variables as well. In this case all the n_i's are one. It is also sometimes said that the **cross-entropy** is used as a cost function. This comes from the decomposition

$$S[p,q] = -\int dx\ p(x)\ln[p(x)] + \int dx\ p(x)\ln[q(x)] \tag{14.5.7}$$

$$= -\langle \ln p \rangle + H[p,q] \tag{14.5.8}$$

$H[p,q]$ is called the cross-entropy. You can see that in this case the $\langle \ln p \rangle$ will not come into the minimization since it is not a function of the parameters $\boldsymbol{\theta}$ so minimizing the cross-entropy it will be equivalent to minimizing the relative entropy.

One thing to point out here is that there doesn't appear to be any deep reason why the correct distance is $S[n_i/N, p(x_i|\boldsymbol{\theta})]$ and not $S[p(x_i|\boldsymbol{\theta}), n_i/N]$. These are not equal. In the second case n_i might be zero for some x_i where $p(x_i|\boldsymbol{\theta})$ is not zero. This would make the relative entropy undefined which is a practical, but not theoretically motivated reason to prefer one over the other.

Appendix A

Selected Problem Solutions

[1] **Problem 1.**

1. This is Bayes' theorem

$$P(R|B) = \frac{P(B|R)P(R)}{P(B)} \tag{A.0.1}$$

2. This is the product rule

$$P(B, R) = P(B|R)P(R) \tag{A.0.2}$$

3. This is the extended sum rule

$$P(B||R) = P(B) + P(R) - P(B, R) \tag{A.0.3}$$
$$= P(B) + P(R) - P(B|R)P(R) \tag{A.0.4}$$

[2] **Problem 2.**
Let's say the C means the person has cancer and T means the person's test is positive.

1. From Bayes

$$p(C|T) = \frac{P(T|C)P(C)}{P(T)} \tag{A.0.5}$$

$P(T)$ can be calculated by summing over the possible states of having and not having cancer:

$$P(T) = P(T|\overline{C})P(\overline{C}) + P(T|C)P(C) = P(T|\overline{C})(1 - P(C)) + P(T|C)P(C) \tag{A.0.6}$$

$$= (1 - 0.90) \times (1 - 0.0001) + 0.90 \times 0.0001 \tag{A.0.7}$$

$$= 0.10008 \tag{A.0.8}$$

$$p(C|T) = \frac{0.9 \times 0.0001}{0.10008} = 0.0009 \tag{A.0.9}$$

This is a pretty useless test! Rare diseases require exceptionally accurate tests.

2. Again the product rule

$$p(F, C) = p(F|C)P(C) = [1 - p(T|C)]p(C) = (1 - 0.9) \times 0.0001 = 0.00001 \tag{A.0.10}$$

[3] **Problem 3.**

If the probability of getting a 6 is p_6 and the probability of getting a 5 is p_5 then the probability of getting any other number is $1 - p_6 - p_5$. Using the multinomial distribution distribution, the probability of getting two 6s and one 5 out of 6 is

$$P = \frac{6!}{2!1!3!}(p_6)^2 p_5 (1 - p_6 - p_5)^3 \tag{A.0.11}$$

If $p_6 = 2p_i$ for $i \neq 6$ then normalization requires that $p_6 = 2/7$ and $p_5 = 1/7$ and

$$P = \frac{6!}{2!1!3!}\left(\frac{2}{7}\right)^2 \frac{1}{7}\left(\frac{4}{7}\right)^3 \simeq 0.13 \tag{A.0.12}$$

[4] **Problem 4.**

I am going to ignore the fact that "years that are divisible by 100, but not by 400, do not contain a leap day". I will also ignore the possibility that the members of this group might be related in some way that makes it more likely that they were born within a few years of each other. Under these assumptions/approximations the chance of being born on a leap day, p_L, is one fourth of the chance of being born on any other day, i.e. $p_L = p_i/4$. Since there are 356 other days and the total probability for all the days must add up to 1, $p_L + 365p_i = 1$, it follows that

$$p_i = \frac{1}{365 + 1/4} \quad p_L = \frac{1}{4 \times 365 + 1} = \frac{1}{1425} \tag{A.0.13}$$

Let's say the number of birthdays on normal day i is n_i and the number of birthdays on the leap day is n_L. The probability of having any combination of birthdays is given by the multinomial distribution (section **??**)

$$P(n_1 \ldots n_{356}, n_L) = \frac{n!}{n_L! \prod_{i=1}^{356} n_i!} p_L^{n_L} \prod_{i=1}^{356} p_i^{n_i} \tag{A.0.14}$$

$$= \frac{n!}{n_L! \prod_{i=1}^{356} n_i!} p_L^{n_L} p_i^{\sum_i n_i} \tag{A.0.15}$$

where some of the n_i will have to be zero if $n < 367$. Just like for the no leap day case, we approach this problem by calculating the probability of no two birthdays being the same and then subtract it from 1. In this case there are two distinct cases that might occur: one is where no one has a leap day birthday and the other is where one person has one a leap day birthday. All the n's must be 1 or zero so the denominator above will always be 1.

For the case of no leap day birthday $\sum_i n_i = n$. We can choose these n distinct birthdays in $\binom{356}{n}$ ways that all have the same probability of happening so the probability is

$$P(\text{all normal days}) = \binom{356}{n} n! p_i^n \tag{A.0.16}$$

If one of the days is a leap day, $n_L = 1$ and $\sum_i n_i = n - 1$ and there are $\binom{356}{n-1}$ ways of picking the normal days so

$$P(1 \text{ leap day}) = \binom{356}{n-1} n! p_i^{n-1} p_L \tag{A.0.17}$$

The probability of having no two birthdays the same is then 1 minus the sum of these probabilities

$$P = 1 - \frac{n!}{356.25^n} \left[\binom{356}{n} + \frac{1}{4} \binom{356}{n-1} \right] \tag{A.0.18}$$

$$= 1 - \frac{n!}{356.25^n} \binom{356}{n} \left[1 + \frac{1}{4} \frac{n}{(357 - n)} \right] \tag{A.0.19}$$

This will decrease the probability slightly relative to the no leap year result.

[7] **Problem 7.**

Imagine a machine that scoops gelato automatically and the flavors all lined up in a row. The machine can do two actions. It can scoop and it can move to the next flavor. To make one bowl of gelato and get to the end of the flavors it must scoop

n times and move $f-1$ times so it does $n+f-1$ actions. Any combination of moves and scoops will make a valid bowl, but some combinations will make the same bowl. There are $\binom{n+f-1}{f-1}$ ways of choosing the $f-1$ moves which is the same thing as $\binom{n+f-1}{n}$, the number of ways to pick n scoops.

[8] **Problem 8.**

The probability of a star not being within a sphere of radius r is derived in the same way as in section 3.6. It is an exponential with $\nu = \eta V = \frac{4}{3}\pi r^3 \eta$. Then the probability of a star being in a shell between r and $r+dr$ is $4\pi\eta r^2 dr$. The probability of these both being true at the same time is the product,

$$p(r)dr = 4\pi r^2 \eta \exp\left[-\frac{4}{3}\pi\eta r^3\right]\, dr \tag{A.0.20}$$

$$= \left(\frac{r}{r_o}\right)^2 e^{-\frac{1}{3}\left(\frac{r}{r_o}\right)^3}\, \frac{dr}{r_o} \tag{A.0.21}$$

where $r_o \equiv (4\pi\eta)^{-1/3}$. You can verify that this is properly normalized by integrating it from 0 to ∞.

The average can be found by looking up the integral which gives

$$\langle r\rangle = 3^{1/3}\Gamma\left(\frac{4}{3}\right) r_o \simeq 1.2879\ r_o \tag{A.0.22}$$

[9] **Problem 9.**

In the rest frame of a gas atom the particle is moving at a speed $|v-v_t|$ where v_t is the velocity of the atom. For the particle to travel for a time t without coming within $\sqrt{\sigma}/\pi$ of an atom there must be a cylinder of volume $\sigma t|v-v_t|$ that contains no atoms. The number density of atoms with velocities between v_t and $v_t+\delta v_t$ is $\eta f(v_t)\delta v_t$ where $f(v_t)$ is the velocity distribution. The probability of none of these being in the cylinder is

$$p(0|v_t) = e^{-\eta\sigma t|v-v_t|f(v_t)\delta v_t} \tag{A.0.23}$$

We want the probability that no atoms of velocity v_t^1 **and** v_t^2, etc. to be in the cylinder so we can use the product rule

$$p(0) = p(0|v_t^1)p(0|v_t^2)\ldots \tag{A.0.24}$$

$$= e^{-\eta\sigma t|v-v_t^1|f(v_t^1)\delta v_t} e^{-\eta\sigma t|v-v_t^2|f(v_t^2)\delta v_t}\ldots \tag{A.0.25}$$

$$= e^{-\eta\sigma t\sum_i |v-v_t^i|f(v_t^i)\delta v_t} \tag{A.0.26}$$

$$\simeq e^{-\eta\sigma t\int dv_t |v-v_t^i|f(v_t^i)} \tag{A.0.27}$$

$$= e^{-\eta\sigma t\langle|v-v_t|\rangle} \tag{A.0.28}$$

At the end of the cylinder there must be an atom of velocity v_t^1 **or** v_t^2, etc. so we use the sum rule

$$p(1|t, t+\delta t)dt = \eta\sigma dt \sum_i |v - v_t^i| f(v_t^i)\delta v_t \tag{A.0.29}$$

$$= \eta\sigma\langle|v - v_t|\rangle dt \tag{A.0.30}$$

Putting these together along with the fact that the particle travels $l = vt$ during time t gives

$$P(l)dl = \eta\sigma\frac{\langle|v - v_t|\rangle}{v} e^{-\eta\sigma v\langle|v-v_t|\rangle l}\, dl \tag{A.0.31}$$

$$= e^{-\frac{l}{l_o}}\, \frac{dl}{l_o} \tag{A.0.32}$$

where $l_o = v/(\eta\sigma\langle|v - v_t|\rangle)$. It is an exponential distribution. The average is l_o.

[11] **Problem 11.**

The distribution for two independent normally distributed variables is

$$p_{xy}(x, y) = p_x(x)p_y(y) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{x^2 + y^2}{2\sigma^2}\right] \tag{A.0.33}$$

Let's define a new variable $z \equiv x/y$ and transform variable:

$$p_{z,y}(z, y) = p_{xy}(yz, y)\left|\frac{\partial x}{\partial z}\right| = p_{xy}(yz, y)|y| \tag{A.0.34}$$

$$= \frac{1}{2\pi\sigma^2}|y|\exp\left[-\frac{y^2(z^2+1)}{2\sigma^2}\right] \tag{A.0.35}$$

Now we can find $p_z(z)$ by marginalizing over y:

$$p_z(z) = \int_{-\infty}^{\infty} dy\; p_{zy}(z, y) \tag{A.0.36}$$

$$= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty} dy\; |y|\exp\left[-\frac{y^2(z^2+1)}{2\sigma^2}\right] \tag{A.0.37}$$

$$= \frac{1}{\pi(z^2+1)}\int_0^{\infty} dw\; we^{-w^2/2} \tag{A.0.38}$$

$$= \frac{1}{\pi(z^2+1)} \tag{A.0.39}$$

[12] **Problem 12.**

Characteristic function for a Poisson distribution is

$$C(k) = \exp\left[\lambda(e^{ik} - 1)\right] \tag{A.0.40}$$

The characteristic function for the sum of random variables is the product of their characteristic functions so

$$C_s(k) = C_1(k)C_2(k) = \exp\left[(\lambda_1 + \lambda_2)(e^{ik} - 1)\right] \tag{A.0.41}$$

which is the characteristic function for a Poisson distribution. So the sum is Poisson distributed with a a mean of $\lambda_1 + \lambda_2$.

[15] **Problem 15.**

The joint pdf is

$$p(x,y) = \begin{cases} \frac{1}{2}\delta^D(x^2 - y) & , \quad -1 < x < 1 \text{ and } 0 < y < 1 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{A.0.42}$$

Clearly this cannot be written in the form $p(x)p(y)$ so they are not independent. The correlation is

$$\begin{aligned} C_{xy} &= E[xy] - E[x]E[y] && \text{(A.0.43)} \\ &= \int_{-1}^{1} dx \int_{0}^{1} dy\ \frac{1}{2}\delta^D(x^2 - y) - 0 && \text{(A.0.44)} \\ &= \int_{-1}^{1} dx \frac{1}{2} && \text{(A.0.45)} \\ &= 0 && \text{(A.0.46)} \end{aligned}$$

The variables are not correlated.

[17] **Problem 17.**

This easily comes from the fact that the characteristic function for the some of two variables is the product of the characteristic functions of each.

$$\begin{aligned} \phi_s(t) &= \phi_x(t)\phi_{x'}(t) && \text{(A.0.47)} \\ &= (1 - 2it)^{-n/2}(1 - 2it)^{-m/2} && \text{(A.0.48)} \\ &= (1 - 2it)^{-(n+m)/2} && \text{(A.0.49)} \end{aligned}$$

[18] **Problem 18.**

The distribution for the velocity is

$$p(v_1, v_2, v_3) = \frac{1}{(2\pi)^{3/2}\sigma^3} e^{-\frac{1}{2}\sum_i \frac{v_i^2}{\sigma^2}} \tag{A.0.50}$$

We can recognize that

$$z = \sum_i \frac{v_i^2}{\sigma^2} \tag{A.0.51}$$

will be χ^2 distributed with 3 degrees of freedom so its distribution is

$$p(z)dz = \frac{2}{\sqrt{\pi}} z^{1/2} e^{-z/2} dz \tag{A.0.52}$$

The kinetic energy of a particle is

$$\epsilon = \frac{1}{2} m|v|^2 = \frac{m\sigma^2}{2} z \tag{A.0.53}$$

Changing variables from z to ϵ gives

$$p(\epsilon)d\epsilon = \frac{2}{\sqrt{\pi}} \frac{1}{(m\sigma^2)^{3/2}} \epsilon^{1/2} e^{-\frac{\epsilon}{m\sigma^2}} d\epsilon \tag{A.0.54}$$

The exponential can be recognized as the Boltzmann factor with the identification $m\sigma^2 = k_b T$, Boltzmann's constant times temperature. The $\epsilon^{1/2}$ factor represents the phase-space being larger for larger ϵ.

[19] **Problem 19 .**

The distribution of observed energies can be written

$$p(E) = f(E) + \delta^D(E - E_{\max}) \int_{E_{\max}}^{\infty} f(E) \tag{A.0.55}$$

$$= f(E) + \delta^D(E - E_{\max}) \left[1 - F(E_{\max})\right] \tag{A.0.56}$$

for $E \leq E_{\max}$ and zero for $E > E_{\max}$. The mean is thus

$$\langle E \rangle = \int_0^{E_{\max}} dE \; E f(E) + E_{\max} \left[1 - F(E_{\max})\right] \tag{A.0.57}$$

If we did not detect photons above this would be

$$\langle E \rangle = \frac{\int_0^{E_{\max}} dE \; E f(E)}{\left[1 - F(E_{\max})\right]} \tag{A.0.58}$$

[20] **Problem 20 .**

The dust particles are uniformly distributed so the distribution of the distance form the center of the balloon of radius R is

$$p(r)dr = \frac{3}{R^3} r^2 dr \tag{A.0.59}$$

implied the $r < R$. The cumulative distribution is

$$F(r) = \left(\frac{r}{R}\right)^3 \tag{A.0.60}$$

Using equation (4.5.1) we find the distribution for the most distant dust particle from the center to be

$$p(r_{\max}) = \frac{3n}{R}\left(\frac{r_{\max}}{R}\right)^{3n-1} \tag{A.0.61}$$

The distance between the skin and this particle is $R - r_{\max}$.

[21] **Problem 21 .**

1. The likelihood is

$$\mathcal{L}(\{t_i\}|\tau) = \prod_i \frac{1}{\tau} e^{-\frac{t_i}{\tau}} = \frac{1}{\tau^n} e^{-\frac{n\bar{t}}{\tau}}. \tag{A.0.62}$$

   Note that the average of the data $\bar{t}$ is a sufficient statistic in this case. The evidence is

$$\mathcal{E}(\{t_i\}) = \int_0^\tau d\tau \ \frac{1}{\tau^n} e^{-\frac{n\bar{t}}{\tau}} = (n\bar{t})^{1-n} \int_0^\infty dx \ x^{n-2} e^{-xn\bar{t}} = (n\bar{t})^{1-n}\Gamma(n-1) \tag{A.0.63}$$

   So the posterior is

$$p(\tau|\{t_i\})d\tau = \left(\frac{n\bar{t}}{\tau}\right)^n \frac{e^{-\frac{n\bar{t}}{\tau}}}{\Gamma(n-1)} \frac{d\tau}{(n\bar{t})} \tag{A.0.64}$$

2. Taking the log of the posterior gives

$$\ln(p(\tau|\{t_i\})) = -\frac{n\bar{t}}{\tau} - n\ln(\tau) + \ldots \tag{A.0.65}$$

   which is maximized at $\tau = \bar{t}$. Note that the mean of the exponential distribution is τ so this is the same as the most obvious estimator for the mean.

[22] **Problem 22.**

In this case

$$f(m) = -\frac{(1+\alpha)}{m_{min}}\left(\frac{m}{m_{min}}\right)^\alpha \alpha < -1 \tag{A.0.66}$$

The log of the likelihood is

$$\ln \mathcal{L} = n \ln(-1-\alpha) + \alpha \sum_{i=0}^{n} \ln\left(\frac{m_i}{m_{min}}\right) - n \ln\left(m_{min}\right) \tag{A.0.67}$$

Taking its derivative and setting it to zero gives the solution

$$\hat{\alpha} = -\left[1 + \frac{1}{\frac{1}{n}\sum_{i=0}^{n} \ln\left(\frac{m_i}{m_{min}}\right)}\right] \tag{A.0.68}$$

[23] **Problem 23 .**

If the probability for a ν_e is p then the probability of a not ν_e is $1-p$. With the selection function probability are

$$p_e = \frac{pS_e}{pS_e + (1-p)S_{-e}} \tag{A.0.69}$$

$$p_{-e} = \frac{(1-p)S_{-e}}{pS_e + (1-p)S_{-e}} \tag{A.0.70}$$

The likelihood for the whole data set is

$$\mathcal{L}(n_e, n_{-e}|p) = (p_e)^{n_e}(p_{-e})^{n_{-e}} = \frac{\left[pS_e\right]^{n_e}\left[(1-p)S_{-e}\right]^{n_{-e}}}{\left[S_e p + S_{-e}(1-p)\right]^N} \tag{A.0.71}$$

The maximum of this likelihood can be found by taking the derivative of its log and setting it to zero giving,

$$\hat{p} = \frac{n_e S_{-e}}{\left[n_e S_{-e} + n_{-e} S_e\right]}. \tag{A.0.72}$$

[24] **Problem 24.**

The normalized mass function is

$$p(M|\alpha, M_*) = \frac{\left(\frac{M}{M_*}\right)^{\alpha} e^{-M/M_*}}{M_* \Gamma\left(\alpha+1, \frac{M_{\min}}{M_*}\right)} \tag{A.0.73}$$

where $\Gamma(\beta, x)$ is the incomplete gamma function. The likelihood for all the data is

$$\ln \mathcal{L}(\{M_i\}|\alpha, M_*) = \ln \prod_i p(M_i|\alpha, M_*) \tag{A.0.74}$$

$$= \alpha \sum_i^n \ln M_i - (1+\alpha) n \ln M_* - n\frac{\overline{M}}{M_*} - n \ln \Gamma\left(\alpha+1, \frac{M_{\min}}{M_*}\right) \tag{A.0.75}$$

[26] **Problem 26 .**

The posterior in section 5.8 does not contain any constraint on the normalization of the spectrum just its shape. Any normalization would show up in the likelihood and evidence and drop cancel out. Expected number of observed stars is

$$N = V \int_o^\infty dl \; f(l) \tag{A.0.76}$$

where it is assumed that $f(l)$ is normalized to be the number of stars per volume. V is the volume surveyed. The distribution of the actual number of stars is Poisson so we can multiply the likelihood by

$$p(N_{ob}|N) = \frac{N^{N_{ob}}}{N_{ob}!} e^{-N} \tag{A.0.77}$$

And then the posterior will give both a constraint on the shape and normalization of the luminosity function.

[27] **Problem 27 .**

$$\left\langle \hat{f}(x)\hat{f}(z) \right\rangle = \frac{\sum_{ij}^n \langle y_i y_j \rangle \; \mathrm{K}_h(x, x_i)\, \mathrm{K}_h(z, x_j)}{[\sum_{i=1}^n \mathrm{K}_h(x, x_i)]\,[\sum_{i=1}^n \mathrm{K}_h(z, x_i)]} \tag{A.0.78}$$

$$= \frac{\sum_{ij}^n (C_{ij} + f(x_i)f(x_j)) \; \mathrm{K}_h(x, x_i)\, \mathrm{K}_h(z, x_j)}{[\sum_{i=1}^n \mathrm{K}_h(x, x_i)]\,[\sum_{i=1}^n \mathrm{K}_h(z, x_i)]} \tag{A.0.79}$$

So

$$\hat{C}_{xz} = \frac{\sum_{ij}^n C_{ij} \; \mathrm{K}_h(x, x_i)\, \mathrm{K}_h(z, x_j)}{[\sum_{i=1}^n \mathrm{K}_h(x, x_i)]\,[\sum_{i=1}^n \mathrm{K}_h(z, x_i)]} \tag{A.0.80}$$

If the noise is uncorrelated for each data point

$$C_{xz} = \frac{\sum_{i}^n \sigma_i^2 \; \mathrm{K}_h(x, x_i)\, \mathrm{K}_h(z, x_i)}{[\sum_{i=1}^n \mathrm{K}_h(x, x_i)]\,[\sum_{i=1}^n \mathrm{K}_h(z, x_i)]} \tag{A.0.81}$$

[28] **Problem 28 .**

1. The maximum likelihood solution for $\boldsymbol{\theta}$ was given in the text.

$$\boldsymbol{\theta} = \left[\boldsymbol{W}^T \boldsymbol{N}^{-1} \boldsymbol{W}\right]^{-1} \boldsymbol{W}^T \boldsymbol{N}^{-1} d \tag{A.0.82}$$

2. The log of the posterior is

$$\ln(\mathcal{L}) = -\frac{1}{2}(\boldsymbol{W\theta} - d)^T \boldsymbol{N}^{-1}(\boldsymbol{W\theta} - d) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{A}^{-1}\boldsymbol{\theta} + ... \tag{A.0.83}$$

$$= -\frac{1}{2}(W_{il}\boldsymbol{\theta}_l - d_i)N_{ij}^{-1}(W_{jk}\theta_k - d_j) - \frac{1}{2}\theta_i A_{ij}^{-1}\theta_j + ... \tag{A.0.84}$$

Einstein summation implied and normalization constants ignored. We take the derivative to find the maximum.

$$\frac{\partial \ln(\mathcal{L})}{\partial \theta_p} = -\frac{1}{2}\left[W_{ip}N_{ij}^{-1}(W_{jk}\theta_k - d_j) + (W_{il}\boldsymbol{\theta}_l - d_i)N_{ij}^{-1}W_{jp}\right] - \frac{1}{2}\left[A_{pj}^{-1}\theta_j + \theta_i A_{ip}^{-1}\right] \tag{A.0.85}$$

$$= -[W_{ip}N_{ij}^{-1}W_{jk} + A_{pk}^{-1}]\theta_k + W_{ip}N_{ij}^{-1}d_j \tag{A.0.86}$$

where the fact that $\boldsymbol{A}$ and $\boldsymbol{N}$ must be symmetric has been used. Setting this equal to zero and solving for θ_k gives the result

$$\boldsymbol{\theta} = \left[\boldsymbol{A}^{-1} + \boldsymbol{W}^T\boldsymbol{N}^{-1}\boldsymbol{W}\right]^{-1}\boldsymbol{W}^T\boldsymbol{N}^{-1}\boldsymbol{d} \tag{A.0.87}$$

In some contexts this solution is called a **Wiener filter**. When $\boldsymbol{d}$ are the pixels of a noisy signal or the pixels of a noisy image this can be used to "denoise" it into a smoother signal or image as long as one knows the expected covariance of the signal $\boldsymbol{A}$ or is willing to estimate it. $\boldsymbol{W}$ could be the psf of a telescope that blurs the image or some other response function. This technique is used to fill in

[30] **Problem 30.**

In this case

$$t_n = \frac{1}{n}\sum_i^n x_i \tag{A.0.88}$$

$$t_{n-1}^{(i)} = \frac{1}{n-1}\sum_{j\neq i} x_j = \frac{1}{n-1}\left[\sum_j^n x_j - x_i\right] = \frac{n}{n-1}t_n - \frac{1}{n-1}x_i \tag{A.0.89}$$

and

$$t_{n-1}^J = \frac{1}{n}\sum_i^n t_{n-1}^{(i)} = \frac{1}{n}\sum_i^n\left[\frac{n}{n-1}t_n - \frac{1}{n-1}x_i\right] \tag{A.0.90}$$

$$= \frac{n}{n-1}t_n - \frac{1}{n-1}t_n = t_n \tag{A.0.91}$$

$$t_n^J = nt_n + (1-n)\bar{t}_{n-1}^J = t_n \tag{A.0.92}$$

So the jackknife estimate of the mean is just the mean and since the mean is unbiased so is the jackknife mean.

$$Var^J[t_n] = \frac{n-1}{n}\sum_{i=1}^{n}\left[t_{n-1}^{(i)} - \bar{t}_{n-1}^J\right]^2 \tag{A.0.93}$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}\left[t_{n-1}^{(i)} - t_n\right]^2 \tag{A.0.94}$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}\left[(t_{n-1}^{(i)})^2 - 2t_{n-1}^{(i)}t_n + t_n^2\right] \tag{A.0.95}$$

$$= \frac{n-1}{n}\left[\sum_{i=1}^{n}(t_{n-1}^{(i)})^2\right] - (n-1)t_n^2 \tag{A.0.96}$$

$$= \frac{n-1}{n}\left[\sum_{i=1}^{n}(\frac{n}{n-1}t_n - \frac{1}{n-1}x_i)^2\right] - (n-1)t_n^2 \tag{A.0.97}$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(nt_n - x_i)^2\right] - (n-1)t_n^2 \tag{A.0.98}$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(n^2t_n^2 - 2nt_nx_i + x_i^2)\right] - (n-1)t_n^2 \tag{A.0.99}$$

$$= \frac{1}{n(n-1)}\left[n^3t_n^2 - 2n^2t_n^2 + \sum_{i=1}^{n}x_i^2\right] - (n-1)t_n^2 \tag{A.0.100}$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}x_i^2 - nt_n^2\right] \tag{A.0.101}$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}(x_i^2 - t_n)^2\right] \tag{A.0.102}$$

So in this case the jackknife estimate of the variance of the mean is the same as the variance in the mean that we already know.

[31] **Problem 31 .**

In this case

$$t_n = \frac{1}{n}\sum_{i}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left[\sum_{i}^{n}x_i^2 - n\overline{x}^2\right] \tag{A.0.103}$$

So the jackknife sampled statistics are

$$t_{n-1}^{(i)} = \frac{1}{n-1}\sum_{j\neq i}(x_j - \overline{x}^{(i)})^2 \tag{A.0.104}$$

$$= \frac{1}{n-1}\left[\sum_{j\neq i} x_j^2 - (n-1)\left(\overline{x}^{(i)}\right)^2\right] \tag{A.0.105}$$

$$= \frac{1}{n-1}\left[\sum_{j\neq i} x_j^2 - (n-1)(\frac{1}{n-1}\sum_{j\neq i} x_j)^2\right] \tag{A.0.106}$$

$$= \frac{1}{n-1}\left[\sum_{j}^{n} x_j^2 - x_i^2 - \frac{1}{(n-1)}(n\overline{x} - x_i)\right] \tag{A.0.107}$$

Averaging this over i,

$$\bar{t}_{n-1}^{J} = \frac{1}{n}\sum_i t^(i)_{n-1} \tag{A.0.108}$$

$$= \frac{1}{n(n-1)}\left[n\sum_j x_j^2 - \sum_i x_i^2 - \frac{1}{n-1}\left(n^3\overline{x}^2 - 2n^2\overline{x}^2 + \sum_i x_i^2\right)\right] \tag{A.0.109}$$

$$\vdots \tag{A.0.110}$$

$$= \frac{n-2}{(n-1)^2}\left[\sum_i x_i^2 - n\overline{x}^2\right] \tag{A.0.111}$$

So the jackknife bias corrected estimator is

$$t_n^J = nt_n + (1-n)\bar{t}_{n-1}^J \tag{A.0.112}$$

$$= \left[\sum_i^n x_i^2 - n\overline{x}^2\right] - \frac{n-2}{(n-1)}\left[\sum_i x_i^2 - n\overline{x}^2\right] \tag{A.0.113}$$

$$= \frac{1}{n-1}\left[\sum_i^n x_i^2 - n\overline{x}^2\right] = \frac{1}{n-1}\sum_i^n (x_i - \overline{x})^2 \tag{A.0.114}$$

So the bias is correctly removed.

[34] **Problem 34.**

The null hypothesis is that the Hubble constants are the same. If the distribution of each measurement is Gaussian then the distribution of their difference , $\Delta H_o = H_o^{(1)} - H_o^{(2)}$, is Gaussian with a mean of zero and a variance of $\sigma_{\Delta H_o}^2 = \sigma_1^2 + \sigma_2^2$. In this case $\Delta H_o = 10$ km/s/Mpc and $\sigma_{\Delta H_o} = 8.6$ km/s/Mpc so $\Delta H_o/\sigma_{\Delta H_o} = 1.162$. So

there is a little more than a "one sigma" disagreement. The the probability of getting a value larger than this is 0.1226 (in Python $1 -$ scipy.stats.norm.cdf(1.162)). This would be a one tail test. In this case a two tailed test would be more appropriate. The probability that $|\Delta H_o/\sigma_{\Delta H_o}| > 1.162$ is 0.25 (twice the one tail). So the hypothesis that these measurements agree can be ruled out with 75 % confidence which would not be considered high enough to conclude that there is any inconsistency between them.

If this confidence level was high (say larger than 99%) you would conclude that these measurements are not consistent. Either one or both of the experiments have made an error (perhaps underestimated the systematics) or the physical assumptions made by both experiments is incorrect in which case they are measuring different things and/or there are other physical parameters that were not included in the analysis.

[35] **Problem 35.**

$$f = Var[\rho_\alpha^A]/Var[\rho_\alpha^B] = 2.612 \tag{A.0.115}$$

In Python code for this calculation would be
A = np.array([97, 90, 95, 90, 101, 99, 99, 107,102, 95])
B = np.array([101, 94, 93, 96,94 ,97, 94, 98, 98, 90, 90, 95])

f = np.var(A)/np.var(B)
df1 = len(A) - 1
df2 = len(B) - 1

print "p_value =", 1-scipy.stats.f.cdf(f, df1, df2)

The result is p_value = 0.068 so, as long as the distributions are Gaussian (big if), we can be reasonable confident that they to not have the same variances. Since the variance of A is larger you would want to use company B. Note that we could have used f = np.var(B)/np.var(A) which would be less than one and we would use p_value = scipy.stats.f.cdf(f, b, a) which would give us the same result.

[37] **Problem 37.**

$$\bar{\boldsymbol{P}}\bar{\boldsymbol{P}} = (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{I} - \boldsymbol{P} - \boldsymbol{P} + \boldsymbol{P}\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{P} = \bar{\boldsymbol{P}} \tag{A.0.116}$$

If $\boldsymbol{x}_e$ is any eigenvector with eigenvalue a_e

$$\bar{\boldsymbol{P}}\bar{\boldsymbol{P}}\boldsymbol{x}_e = a_e\bar{\boldsymbol{P}}\boldsymbol{x}_e = a_e^2\boldsymbol{x}_e \tag{A.0.117}$$

so $a_e^2 = a_e$ or $a_e = 0$ or 1.

The cyclic property of the trace implies

$$\mathrm{tr}\,[\boldsymbol{P}] = \mathrm{tr}\left[\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\right] = \mathrm{tr}\left[\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)\right] = \mathrm{tr}\,[\boldsymbol{I}] = k \tag{A.0.118}$$

Remember that $\boldsymbol{M}$ is a n-by-k matrix so $\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}$ is a k-by-k matrix.

$$\mathrm{tr}\left[\bar{\boldsymbol{P}}\right] = \mathrm{tr}\,[\boldsymbol{I} - \boldsymbol{P}] = \mathrm{tr}\,[\boldsymbol{I}] - \mathrm{tr}\,[\boldsymbol{P}] = n - k \tag{A.0.119}$$

Since all the eigenvalues are either zero or one, the trace must be equal to the number of non zero eigenvalues which is equal to the number of eigenvectors the span the subspace of $\bar{\boldsymbol{P}}$'s range.

[40] **Problem 40.**

First

$$\sum_i (X_i - Y_i)^2 = \sum_i (X_i - \overline{X} - Y_i + \overline{Y})^2 \qquad \overline{X} = \overline{Y} \tag{A.0.120}$$

$$= \sum_i \left[(X_i - \overline{X})^2 + (Y_i - \overline{Y})^2 - 2(X_i - \overline{X})(Y_i - \overline{Y})\right] \tag{A.0.121}$$

$$= 2nV_X - 2\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) \tag{A.0.122}$$

Then using

$$r_s = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{nV_X} \tag{A.0.123}$$

if follows that

$$r_s = 1 - \frac{1}{2nV_X}\sum_i (X_i - Y_i)^2 \tag{A.0.124}$$

$$= 1 - \frac{6}{n(n^2-1)}\sum_i (X_i - Y_i)^2 \tag{A.0.125}$$

using (9.4.5).

[42] **Problem 42 .**

In this case the likelihood is

$$\mathcal{L} = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right] \tag{A.0.126}$$

The maximum likelihood estimates of μ and σ^2 are

$$\hat{\mu} = \frac{1}{n}\sum_i x_i \qquad \hat{\sigma^2} = \Delta^2 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i - \hat{\mu}\right]^2 \tag{A.0.127}$$

The likelihood evaluated at its maximum is

$$\ln \mathcal{L}_{max} = -\frac{1}{2\Delta^2}\sum_i (x_i - \mu)^2 - \frac{n}{2}\ln \Delta^2 \tag{A.0.128}$$

$$= -\frac{n}{2} - \frac{n}{2}\ln \Delta^2 \tag{A.0.129}$$

And so

$$\mathrm{BIC} = 2\ln n + n\left(\ln \Delta^2 + 1\right) \tag{A.0.130}$$

[48] **Problem 48.**

$$\langle A \rangle = \frac{a}{N}\sum_i^N \langle x_i \rangle = a\mu \tag{A.0.131}$$

$$\left\langle (A-\mu)^2 \right\rangle = \left\langle A^2 \right\rangle - 2\mu\langle A\rangle + \mu^2 \tag{A.0.132}$$

$$= \left(\frac{a}{N}\right)^2 \sum_{ij}^N \langle x_i x_j \rangle - 2a\mu^2 + \mu^2 \tag{A.0.133}$$

$$= \left(\frac{a}{N}\right)^2 \left(\sum_{ij}^N \langle x_i x_j \rangle\right) + (1-2a)\mu^2 \tag{A.0.134}$$

$$= \left(\frac{a}{N}\right)^2 \left(\sum_{i}^N \langle x_i^2 \rangle + \sum_{i\neq j}^N \langle x_i x_j \rangle\right) + (1-2a)\mu^2 \tag{A.0.135}$$

$$= \left(\frac{a}{N}\right)^2 \left(N(\sigma^2+\mu^2) + \sum_{i\neq j}^N \langle x_i\rangle\langle x_j\rangle\right) + (1-2a)\mu^2 \tag{A.0.136}$$

$$= \left(\frac{a}{N}\right)^2 \left(N(\sigma^2+\mu^2) + N(N-1)\mu^2\right) + (1-2a)\mu^2 \tag{A.0.137}$$

Minimizing this gives:

$$\frac{2a}{N}\left((\sigma^2+\mu^2) + (N-1)\mu^2\right) - 2\mu^2 = 0 \tag{A.0.138}$$

so the solution is

$$a = \frac{\mu^2 N}{((\sigma^2 + \mu^2) + (N-1)\mu^2)}. \tag{A.0.139}$$

You can see that for large N $a \to 1$ and this estimator becomes the sample mean. You can also see that for any finite N, A is biased, i.e. $\langle A \rangle \neq \mu$.

[49] **Problem 49 .**

Let us first find the Fisher information

$$\ln \mathcal{L} = -\frac{1}{2}\sum_i^n \frac{(x_i - \mu)^2}{\sigma^2} - \frac{n}{2}\ln(2\pi\sigma^2) \tag{A.0.140}$$

$$\frac{\partial}{\partial\sigma^2}\ln\mathcal{L} = \frac{1}{2[\sigma^2]^2}\sum_i^n (x_i-\mu)^2 - \frac{n}{2\sigma^2} \tag{A.0.141}$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}\ln\mathcal{L} = -\frac{1}{[\sigma^2]^3}\sum_i^n (x_i-\mu)^2 + \frac{n}{2\sigma^4} \tag{A.0.142}$$

$$F_{\sigma^2\sigma^2} = -\left\langle \frac{\partial^2}{\partial(\sigma^2)^2}\ln\mathcal{L}\right\rangle = \frac{n}{2\sigma^4} \tag{A.0.143}$$

Now we know that $\left[\frac{n-1}{\sigma^2}S_n^2\right] \sim \chi^2_{n-1}$. We also know that the variance of a χ^2_n distribution is $2n$ so

$$Var\left[S_n^2\right] = \frac{2\sigma^4}{n-1} \tag{A.0.144}$$

So the efficiency is $F_{\sigma^2\sigma^2}/Var\left[S_n^2\right] = n/(n-1)$. For large n this makes no difference, but for small n it is significantly above the limit.

[53] **Problem 53.**

The log of the likelihood is

$$\ln\mathcal{L}(\boldsymbol{d}) = -\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{d}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{n}{2}\ln[2\pi] \tag{A.0.145}$$

Taking the first derivative with respect to a parameter, α, gives

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln\mathcal{L}(\boldsymbol{d}) =& \frac{1}{2}(\boldsymbol{\mu}_{,\alpha})^T\boldsymbol{C}^{-1}(\boldsymbol{d}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{,\alpha}) \\ & - \frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}{}_{,\alpha}(\boldsymbol{d}-\boldsymbol{\mu}) - \frac{1}{2}\frac{d}{d\alpha}\ln|\boldsymbol{C}|\end{aligned} \tag{A.0.146}$$

where the subscript with commas are derivatives. Using these formulas

$$\frac{d}{d\beta}\ln|\boldsymbol{C}| = \mathrm{tr}\left[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\right] \tag{A.0.147}$$

we can change the last term

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln\mathcal{L}(\boldsymbol{d}) =& \frac{1}{2}(\boldsymbol{\mu}_{,\alpha})^T\boldsymbol{C}^{-1}(\boldsymbol{d}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{,\alpha}) \\ &- \frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}{}_{,\alpha}(\boldsymbol{d}-\boldsymbol{\mu}) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\right]\end{aligned} \tag{A.0.148}$$

Now we know that $\langle(\boldsymbol{d}-\boldsymbol{\mu})\rangle = 0$ so when we take another derivative and average all the terms that are linear in this will be zero,

$$\left\langle\frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathcal{L}(\boldsymbol{d})\right\rangle = -\frac{1}{2}(\boldsymbol{\mu}_{,\alpha})^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{,\beta}) - \frac{1}{2}(\boldsymbol{\mu}_{,\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{,\alpha}) - \frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}{}_{,\alpha\beta}(\boldsymbol{d}-\boldsymbol{\mu}) \tag{A.0.149}$$

$$-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}{}_{,\alpha}\boldsymbol{C}_{,\beta} + \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta\alpha}\right] \tag{A.0.150}$$

The first two terms are the same because $\boldsymbol{C}$ is symmetric and and using $\left\langle(\boldsymbol{d}-\boldsymbol{\mu})(\boldsymbol{d}-\boldsymbol{\mu})^T\right\rangle = \boldsymbol{C}$ in the third term

$$\mathcal{F}_{\alpha\beta} = -\left\langle\frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathcal{L}(\boldsymbol{d})\right\rangle = \boldsymbol{\mu}_{,\alpha}^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_{,\beta} + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}{}_{,\alpha\beta}\boldsymbol{C}\right] + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}{}_{,\alpha}\boldsymbol{C}_{,\beta} + \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta\alpha}\right] \tag{A.0.151}$$

Using

$$\boldsymbol{C}^{-1}{}_{,\beta} = -\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1} \tag{A.0.152}$$

and the chain rule

$$\boldsymbol{C}^{-1}{}_{,\alpha\beta} = \boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1} + \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta\alpha}\boldsymbol{C}^{-1} \tag{A.0.153}$$

Canceling some terms out and using the fact the trace of a product of matrices does not depend on the order of the product one gets equation (12.5.1).

[55] **Problem 55.**

We can solve this by looking at the joint probability of $\boldsymbol{x}_{n+1}$, $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n-1}$ and applying the product rule

$$p(\boldsymbol{x}_{n+1},\boldsymbol{x}_n,\boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n,\boldsymbol{x}_{n-1}) \tag{A.0.154}$$

$$= p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})p(\boldsymbol{x}_{n-1}) \tag{A.0.155}$$

$$= p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})p(\boldsymbol{x}_{n-1}) \tag{A.0.156}$$

The last step comes from the requirement that a Markov chain's transition kernel be expressible as only dependent on the previous state. As we are showing here, this does not mean that a transition probability that skips one or more generations cannot be written down and that it is not dependent on the state $\boldsymbol{x}_{n-1}$.

We can get the joint probability of $\boldsymbol{x}_{n+1}$ and $\boldsymbol{x}_{n-1}$ by marginalizing over $\boldsymbol{x}_n$,

$$p(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n-1}) = \int_{-\infty}^{\infty} d\boldsymbol{x}_n \; p(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n, \boldsymbol{x}_{n-1}) \tag{A.0.157}$$

$$= p(\boldsymbol{x}_{n-1}) \int_{-\infty}^{\infty} d\boldsymbol{x}_n \; p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n) p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \tag{A.0.158}$$

From the product rule we know $p(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_{n-1}) p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{n-1})$ so

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{n-1}) = \int_{-\infty}^{\infty} d\boldsymbol{x}_n \; p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n) p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \tag{A.0.159}$$

So to get from $\boldsymbol{x}_{n-1}$ to $\boldsymbol{x}_{n+1}$ we need to account for all possible intermediate states, $\boldsymbol{x}_n$. If we continued this to more steps we would find the transition by "propagating" through more intermediate states. This starts to remind one of path integrals and Feynman diagrams and indeed there is a connection.

[56] **Problem 56.**

The acceptance probability for a Gibbs step will be

$$\alpha(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n) = \frac{q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})}{q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)} \frac{f(\boldsymbol{x}_{n+1})}{f(\boldsymbol{x}_n)} \tag{A.0.160}$$

$$= \frac{f(x_n^{(i)}|\boldsymbol{x}_{n+1}^{(i-)})}{f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})} \frac{f(\boldsymbol{x}_{n+1}^{(i)}, \boldsymbol{x}_{n+1}^{(i-)})}{f(\boldsymbol{x}_n^{(i)}, \boldsymbol{x}_n^{(i-)})} \tag{A.0.161}$$

$$= \frac{f(x_n^{(i)}|\boldsymbol{x}_n^{(i-)})}{f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})} \frac{f(\boldsymbol{x}_{n+1}^{(i)}, \boldsymbol{x}_n^{(i-)})}{f(\boldsymbol{x}_n^{(i)}, \boldsymbol{x}_n^{(i-)})} \qquad \boldsymbol{x}_{n+1}^{(i-)} = \boldsymbol{x}_n^{(i-)} \tag{A.0.162}$$

$$= \frac{f(x_n^{(i)}|\boldsymbol{x}_n^{(i-)})}{f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})} \frac{f(\boldsymbol{x}_n^{(i-)}) f(\boldsymbol{x}_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})}{f(\boldsymbol{x}_n^{(i-)}) f(\boldsymbol{x}_n^{(i)}|\boldsymbol{x}_n^{(i-)})} \tag{A.0.163}$$

$$= 1 \tag{A.0.164}$$

A.1 Matrix basics

$$(\boldsymbol{ABC}\dots)^T = \dots\boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T \tag{A.1.1}$$
$$(\boldsymbol{ABC}\dots)^{-1} = \dots\boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \tag{A.1.2}$$
$$(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T \tag{A.1.3}$$
$$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T \tag{A.1.4}$$

Some properties of the determinant

$$|\boldsymbol{A}| = \prod_i \lambda_i \quad \text{where } \lambda_i \text{ are the eigenvalues} \tag{A.1.5}$$
$$|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}| \tag{A.1.6}$$
$$|\boldsymbol{BA}| = |\boldsymbol{B}||\boldsymbol{A}| \tag{A.1.7}$$
$$|c\boldsymbol{A}| = c^n|\boldsymbol{A}| \tag{A.1.8}$$
$$|\boldsymbol{A}^T| = |\boldsymbol{A}| \tag{A.1.9}$$

Some properties of the trace

$$\mathrm{tr}(\boldsymbol{A}) = \sum_i A_{ii} \tag{A.1.10}$$
$$\mathrm{tr}(\boldsymbol{A}) = \sum_i \lambda_i \quad \text{where } \lambda_i \text{ are the eigenvalues} \tag{A.1.11}$$
$$\mathrm{tr}(\boldsymbol{A}^T) = \mathrm{tr}(\boldsymbol{A}) \tag{A.1.12}$$
$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}) \tag{A.1.13}$$
$$\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B}) \tag{A.1.14}$$

derivatives of matrices

$$\frac{d}{d\beta}\boldsymbol{C}^{-1} = -\boldsymbol{C}^{-1}\left[\frac{\partial \boldsymbol{C}}{\partial \beta}\right]\boldsymbol{C}^{-1} \tag{A.1.15}$$
$$\tag{A.1.16}$$

$$\frac{d}{d\beta}\ln|\boldsymbol{C}| = \frac{d}{d\beta}\ln\left(\prod_i \lambda_i\right) = \frac{d}{d\beta}\sum_i \ln\lambda_i = \sum_i \frac{1}{\lambda_i}\frac{d\lambda_i}{d\beta} = \mathrm{tr}\left[\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \beta}\right] \tag{A.1.17}$$

$\boldsymbol{A}$ is an **orthogonal matrix** if and only if

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} \tag{A.1.18}$$

An orthogonal matrix has the following properties

$$\boldsymbol{A}^T = \boldsymbol{A}^{-1} \tag{A.1.19}$$

$$|\boldsymbol{A}| = \pm 1 \tag{A.1.20}$$

The $|\lambda_i| = 1$ for all eigenvalues and the magnitude of all eigenvactors are 1.

$\boldsymbol{C}$ is a **positive definite matrix** if

$$\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} > 0 \ \ \forall \boldsymbol{x}. \tag{A.1.21}$$

It has the following properties

- all eigenvalues are positive
- $\text{tr}(\boldsymbol{C}) > 0$
- all diagonal elements are positive, $\boldsymbol{C}_{ii} > 0, \forall i$
- $\boldsymbol{C}$ is invertible

The covariance matrix is always positive definite.

A.2 Matrix decompositions

Eigenvalue decomposition

If $\boldsymbol{A}$ is a $N \times N$ matrix with linear independent columns the it can be decomposed as

$$\boldsymbol{A} = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{M}^{-1} \tag{A.2.1}$$

where $\boldsymbol{\Lambda}$ is diagonal and Λ_{ii} is the ith eigenvalue and the ith column of $\boldsymbol{M}$ is the corresponding eigenvalue.

Single-value decomposition

If $\boldsymbol{A}$ is a $M \times N$ matrix it can be factorized as

$$\boldsymbol{A} = \boldsymbol{S}\boldsymbol{V}\boldsymbol{D}^\dagger \tag{A.2.2}$$

where

- $\boldsymbol{S}$ is a unitary (orthogonal if real) $M \times M$ matrix, i.e. $\boldsymbol{S}\boldsymbol{S}^\dagger = \boldsymbol{S}^\dagger\boldsymbol{S} = \boldsymbol{I}$
- $\boldsymbol{V}$ is a diagonal matrix $M \times N$ with non-negative real entries
- $\boldsymbol{D}$ is a unitary(orthogonal if real) $N \times N$ matrix

$\boldsymbol{D}^\dagger$ is the Hermitian conjugate of $\boldsymbol{D}$. In the case a real matrix $\boldsymbol{D}^\dagger = \boldsymbol{D}^T$. The diagonal elements of $\boldsymbol{V}$ are called the **singular values** of $\boldsymbol{A}$. The columns of $\boldsymbol{D}$ are called the **right-singular vectors**. They are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\dagger$. The columns of $\boldsymbol{S}$ the **left-singular vectors** and are the eigenvectors of $\boldsymbol{A}^\dagger\boldsymbol{A}$.

"A and B"	A, B
"A or B"	$A \| \| B$
continuous random variables	x,y,x_i,y_i
vector of random variables	$\boldsymbol{x}$ or $\vec{x}$
discrete numbers, sometimes random	n,m
parameters	θ_α or p_α
estimator of parameter θ_α	$\tilde{\theta}_\alpha$
maximum likelihood solution for parameter θ_α	$\hat{\theta}_\alpha$
data	$\boldsymbol{D}$ or d_i
indexes data or for multiple random numbers	i, j
statistical and/or theoretical model	M
Gaussian or Normal pdf	$\mathcal{G}\left(\boldsymbol{x}\,\middle\vert\boldsymbol{\mu},\boldsymbol{C}\right)$
$\boldsymbol{x}$ is normally distributed	$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$
x is χ^2 distributed with n degrees of freedom	$x \sim \chi^2_n$
arithmetic mean of N samples	$\bar{x}_N$
likelihood of data given model	$\mathcal{L}(\boldsymbol{D}\|M_i)$ or $P(\boldsymbol{D}\|M_i)$
Bayesian evidence of data	$\mathcal{E}(\boldsymbol{D})$
Heaviside function, 1 when B is true, 0 otherwise	$\Theta(B)$
factorial	$N! = N(N-1)(N-2)\ldots 1$
double factorial	$N!! = N(N-2)(N-4)\ldots$
expectation value of $f(x)$	$\langle f(x)\rangle$ or $E[f(x)]$

Table A.1: notation

A.3 Notation

Notation may vary but in general I follow the guide in table A.1

A.4 Some useful integrals and mathematical definitions

A.4.1 Gaussian integrals

$$\int_{-\infty}^{\infty} dx\ e^{-\frac{x^2}{2}} = \sqrt{2\pi} \tag{A.4.1}$$

$$\begin{aligned}\int_{-\infty}^{\infty} dx\ e^{-(ax^2+bx+c)} &= e^{-c}\int_{-\infty}^{\infty} dx\ e^{-\left(\sqrt{a}x+\frac{b}{2\sqrt{a}}\right)^2+\frac{b^2}{4a}} = e^{-c+\frac{b^2}{4a}}\int_{-\infty}^{\infty}\frac{dy}{\sqrt{a}}\ e^{y^2} \\ &= \sqrt{\frac{\pi}{a}}e^{-c+\frac{b^2}{4a}}\end{aligned} \tag{A.4.2}$$

$$\int_0^{\infty} dx\ x^n e^{-\frac{1}{2}Ax^2} = 2^{\frac{n-1}{2}}A^{-\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right) \qquad n > -1 \tag{A.4.3}$$

A.4.2 Stirling's approximation

$$\ln N! \simeq N\ln N - N \text{ for } N \gg 1 \tag{A.4.4}$$

or more accurately

$$N! \simeq \sqrt{2\pi N}\left(\frac{N}{e}\right)^N \text{ for } N \gg 1 \tag{A.4.5}$$

A.4.3 The Gamma function

$$\begin{array}{lll}\int_0^{\infty} dx\ x^n e^{-x^2} & = \frac{1}{2}\Gamma\left(\frac{n+1}{2}\right) & \\ \int_0^{\infty} dx\ x^{z-1}e^{-x} & = \Gamma(z) & \\ \Gamma(n) & = (n-1)! & n = 1, 2, \ldots \\ \Gamma\left(\frac{1}{2}+n\right) & = \frac{(2n)!}{4^n n!}\sqrt{\pi} & n = 0, 1, 2, \ldots\end{array} \tag{A.4.6}$$

The incomplete gamma function for positive real a and b is

$$\Gamma(z, a, b) = \int_a^b dx\ x^{z-1}e^{-x} \tag{A.4.7}$$

A.4.4 Error function

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2}du = \frac{1}{\sqrt{\pi}}\int_{-z}^{z} e^{-u^2}du \tag{A.4.8}$$

$$\frac{1}{\sqrt{2\pi}\sigma}\int_b^a dx\ e^{-\frac{x^2}{2\sigma}} = \frac{1}{2}\left[\mathrm{erf}\left(\frac{a}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right] \tag{A.4.9}$$

$$\mathrm{erf}(\infty) = 1 \qquad \mathrm{erf}(-x) = -\mathrm{erf}(x) \tag{A.4.10}$$

The cumulative distribution for a standard normal distribution is

$$F(z) = \frac{1}{2}\left[1 + \mathrm{erf}(z)\right]. \tag{A.4.11}$$

A.4.5 Beta function

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 dx\ x^{p-1}(1-x)^{q-1} \tag{A.4.12}$$

There are also many integral forms of this beta function.

A.4.6 Miscellaneous

$$\lim_{N\to\infty}\left[1 + \frac{t^2}{2N}\right]^N = e^{\frac{t^2}{2}} \tag{A.4.13}$$

Sum of arithmetic progression

$$\sum_{i=0}^{n} i = \frac{n(n+1)}{2} \tag{A.4.14}$$

Sum of geometric progression

$$\sum_{i=0}^{n-1} a^i = \frac{1-a^n}{1-a} \tag{A.4.15}$$

A.5 Data Whitening

White noise refers to noise that is not correlated and equal variance in all components. Is $\boldsymbol{y}$ has white noise then its covariance matrix is proportional to the identity matrix

$$< \boldsymbol{y}\boldsymbol{y}^T >= \lambda \boldsymbol{\mathcal{I}} \tag{A.5.1}$$

Note that the Fourier transform of white noise will also be white noise. In an image this would be homogenious, uncorrelated noise.

In some cases it is useful to transform the data into a form that has the covariance of white noise with $\lambda = 1$. This is called **whitening** or pre-whitening the data. This can sometimes make algebra easier, proofs simpler and it can be used to transform a correlated χ^2 into a least-squared problem for example.

First some preliminaries. We can define a standardized data vector

$$\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} \tag{A.5.2}$$

where $\boldsymbol{\Sigma}^{-1/2}$ is the diagonal matrix with $\boldsymbol{\Sigma}_{ii}^{-1/2} = 1/\sqrt{\boldsymbol{C}_{ii}}$ or in other words $1/\sigma_1, \ldots, 1/\sigma_n$. This is the multivariate version of the standardized variable introduced in section 3.2.

The covariance of this data vector, $\left\langle \boldsymbol{z}\boldsymbol{z}^T \right\rangle$ will be the **correlation matrix**,

$$\boldsymbol{P} = \left\langle \boldsymbol{w}\boldsymbol{w}^T \right\rangle = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{C}\boldsymbol{\Sigma}^{-1/2} \tag{A.5.3}$$

This matrix will have ones on the diagonal and be scale invariant in the sense that it will not depend on the units used for the data. This is the matrix of Person's correlation coefficients which is useful for assessing how correlated the variables are.

The eigenvalue decomposition of the covariance matrix is

$$\boldsymbol{C} = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{M}^T \tag{A.5.4}$$

were $\boldsymbol{\Lambda}$ is a diagonal matrix with the variances of each variable on the diagonal, $\sigma_1^2, \ldots, \sigma_n^2$. he components of $\boldsymbol{y}$ will have variances equal to one, but they will still be correlated with each other – their covariance will not be diagonal.

The whitened data vector, $\boldsymbol{w}$, is related to the original data through the whitening matrix $\boldsymbol{W}$,

$$\boldsymbol{w} = \boldsymbol{W}\boldsymbol{x} \tag{A.5.5}$$

If

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{C}^{-1} \tag{A.5.6}$$

then the requirement the $< \boldsymbol{w}\boldsymbol{w}^T >= \boldsymbol{\mathcal{I}}$ will be satisfied. This does not fully specify the whitening matrix however. If $\boldsymbol{R}$ is a rotation matrix such that $\boldsymbol{R}^T = \boldsymbol{R}^{-1}$ then all matrices $\boldsymbol{R}\boldsymbol{W}$ will be valid whitening matrices if $\boldsymbol{W}$ is. Several common choices for $\boldsymbol{W}$ are used.

One choice is to use the matrix

$$\boldsymbol{W} = \boldsymbol{C}^{-1/2} = \boldsymbol{M}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{M}^T \tag{A.5.7}$$

It can be shown that the components of $\boldsymbol{w}$ will be the closest to the original variables as possible in this case.

Another option is to use the PCA or eigenvector bases in which case

$$\boldsymbol{W} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{M}^T \tag{A.5.8}$$

This can be useful when the data is being compressed by removing modes with low signal-to-noise.

Another choice is to use the **Cholesky decomposition** of the covariance matrix

$$\boldsymbol{C}^{-1} = \boldsymbol{L}\boldsymbol{L}^T \tag{A.5.9}$$

where $\boldsymbol{L}$ is lower triangular. This is a unique decomposition. The whitening matrix is

$$\boldsymbol{W} = \boldsymbol{L}^T \tag{A.5.10}$$

Because $\boldsymbol{L}$ is triangular one of the new variable will just be a re-scaling of the original variable, the next one will be composed of just two of the original variable, etc. This might be useful if there is a particular order to the variables that are being analyzed. Kessy et al. (2018) give an interesting summery of whitening options and their properties.

Consider the case where the covariance is a contribution from the a signal, $\boldsymbol{S}$, and correlated noise $\boldsymbol{N}$ which is known so that the total covariance is

$$\boldsymbol{C} = \boldsymbol{S} + \boldsymbol{N} \tag{A.5.11}$$

We can transform this with $\boldsymbol{N}^{-1/2}$

$$\boldsymbol{N}^{-1/2}\boldsymbol{C}\boldsymbol{N}^{-1/2} = \boldsymbol{N}^{-1/2}\boldsymbol{S}\boldsymbol{N}^{-1/2} + \boldsymbol{I} \tag{A.5.12}$$

Let us do a further eigen-decomposition of the whitened signal

$$\boldsymbol{N}^{-1/2}\boldsymbol{S}\boldsymbol{N}^{-1/2} = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{M}^T \tag{A.5.13}$$

which because $\boldsymbol{M}$ is orthogonal implies

$$\boldsymbol{N}^{-1/2}\boldsymbol{C}\boldsymbol{N}^{-1/2} = \boldsymbol{M}\left(\boldsymbol{\Lambda} + \boldsymbol{I}\right)\boldsymbol{M}^T \tag{A.5.14}$$

Thus the modes

$$\boldsymbol{x}' = \boldsymbol{M}^T\boldsymbol{N}^{-1/2}\boldsymbol{x} \tag{A.5.15}$$

have the covariance

$$\left\langle \boldsymbol{x}'\boldsymbol{x}'^{T}\right\rangle = \boldsymbol{\Lambda} + \boldsymbol{I} \tag{A.5.16}$$

These are call **Karhunen-Loéve modes** or **signal-to-noise modes**. They are often used for data compression. The modes with variances close to one will be dominated be noise and those with larger variances will be dominated by signal. In some circumstances it is convenient for computational or data transport reasons to throw away low signal to noise modes and one can do this without significant lose of information. Note that finding the Karhunen-Loéve modes can be very computationally intensive (you need to find $\boldsymbol{N}^{-1/2}$ and the the eigen-decomposition of $\boldsymbol{N}^{-1/2}\boldsymbol{C}\boldsymbol{N}^{-1/2}$) so even if the remainder of the data analysis can proceed with greatly compressed data there is not automatically a savings in processing time.

Index

Bibliography

Akeret J., Refregier A., Amara A., Seehars S., Hasner C., 2015, Journal of Cosmology and Astro-Particle Physics, 2015, 043

Babu G., Rao C., 2004, Sankhyā, 66, 63

Caticha A., 2008, arXiv e-prints

Feigelson E. D., Babu G. J., 2012, Modern Statistical Methods for Astronomy. Cambridge University Press

Feroz F., Hobson M. P., 2008, MNRAS, 384, 449

Gibbs J., 1902, "Elementary Principles in Statistical Michanics". Yale U. Press

Gregory P., 2006, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press

Handley W. J., Hobson M. P., Lasenby A. N., 2015, MNRAS, 453, 4384

Jaynes E., 2003, Probability Theory - The Logic of Science

Kacprzak T., Herbel J., Amara A., Réfrégier A., 2018, Journal of Cosmology and Astroparticle Physics, 2018, 042

Kessy A., Lewin A., Strimmer K., 2018, The American Statistician, 72, 309

Neal R., , 1998, Probabilistic Inference using Markov Chain Monte Carlo Methods

Neal R., 2001, Statistics and Computing, 11, 125

Norisen T., 2017, Foundations of Quantum Mechanics: An Exploration of the Physical Meaning of Quantum Theory. Springer International Publishing,

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 2007, Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3 edn. Cambridge University Press, New York, NY, USA

Protassov R., van Dyk D. A., Connors A., Kashyap V. L., Siemiginowska A., 2002, ApJ, 571, 545

Schwartz G., 1978, Annals of Statistics, 6, 461

Silvia D., Skilling J., 2006, Data Analysis a Bayesian Tutorial, 2nd edn. Oxford University Press

Skilling J., 2004a, in AIP Conf. Proc. Vol. 735 Bayesian inference and maximum entropy methods in science and engineering. Am. Inst. Phys., Melville, NY, p. 395

Skilling J., 2004b, BayeSys and MassInf, http://www.inference.org.uk/bayesys/manual.ps

Surfling R. J., 1980, Approximate Theorems of Mathimatical Statistics. Wiley & Son

Tokdar S., Kass R., 2010, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 54