Notes: Practical Statistics for Physics & Astronomy

R. Benton Metcalf

Alma Mater Studiorum - Universitá di Bologna

May 24, 2018

Contents

1 What is Probability?

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 repeated in the same way or circumstances an *infinite* number of times.

$$P(A) = \lim_{N\to\infty} \frac{\text{number of trials where A is true}}{N(\text{total number of trials})} \tag{1.1}$$

This is the traditional definition of probability as formally stated by Laplace in 1774 and almost universally used for centuries despite no one ever having done anything *exactly* the same 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 repatriation 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 flipping 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 a 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.

Lets be a bit more practical. Lets say we don't need an infinite number of trials, but just a very *large number* of them. Lets 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 classical interpretation of probability

The classical interpretation of probability relies on identifying events that are equally likely or probable. 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 probability of macro states which will not necessarily be equal.

The biggest criticism of this interpretation is that it doesn't really say what probability is, it just tells you how to calculate it in a restricted class of problems. What does it mean that two states are equally probable? What does the probability of a macro-state mean? Another problem is that not all events that we commonly apply probability to can reduced in this way to a collections of equally probable mutually exclusive events.

1.3 Subjective or Bayesian interpretation of probability

Thomas Bayes (1701 - 1761) (and initially by 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).

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 has a probability function 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 Boolean 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 and show that they are complete without ever mentioning randomness or repetition of experiments.

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 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.
   (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 $B|C$

(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.

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

Table 1: The truth table for binary logical expressions.

1.4 Quantum mechanical probability

Probability in standard 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 variable" that are different every trial. In quantum mechanics it can be proven that such hidden variables do not exist or do not take on deterministic values for example with Bell's inequalities. When a measurement is made the square of the wave function gives the probability of an observation, but up to that point the outcome was impossible to determine, not just difficult to determine. This makes probability a property of physical systems and not solely a property of the observer. This seems to imply an intimate connection between physical laws and knowledge.

This is obviously a subject for a different course (or a *Star Trek* episode) so I will go no further.

1.5 the rules of probability

Suppose the 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 if both are false. $(A \cup B)$ 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 $\overline{A \cup B} = \overline{A}, \overline{B}$ and $\overline{A, B} = \overline{A} \cup \overline{B}$ 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 table 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. $P(A \cup B)$ 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 irrelavent and are left out . Some might be relevant but it is 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, in this case flipping a coin.

The two fundamental rules of probability theory are

$$\begin{array}{ll} P(A, B) = P(A)P(B|A) & \textbf{product rule} \\ P(A) + P(\overline{A}) = 1 & \textbf{sum rule} \end{array} \tag{1.2}$$

These rules are actually derivable from some basic requirements or "desidariata" of how probabilities should behave, but for our purposes we can take them to be 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

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \textbf{Bayes' theorem} \tag{1.3}$$

Applying the sum rule to $(A \cup B)$ gives

$$\begin{aligned}
P(A \cup B) &= 1 - P(\overline{A \cup B}) && (1.4)\\
&= 1 - P(\overline{A}, \overline{B}) && (1.5)\\
&= 1 - P(\overline{A})P(\overline{B}|\overline{A}) && (1.6)\\
&= 1 - P(\overline{A})\left[1 - P(B|\overline{A})\right] && (1.7)\\
&= 1 - P(\overline{A}) - P(\overline{A})P(B|\overline{A}) && (1.8)\\
&= P(A) + P(\overline{A})P(B|\overline{A}) && (1.9)\\
&= P(A) + P(\overline{A}, B) && (1.10)\\
&= P(A) + P(B)P(\overline{A}|B) && (1.11)\\
&= P(A) + P(B)\left[1 - P(A|B)\right] && (1.12)\\
&= P(A) + P(B) - P(B)P(A|B) && (1.13)\\
P(A \cup B) &= P(A) + P(B) - P(B, A) \qquad \textbf{extended sum rule} && (1.14)
\end{aligned}$$

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

$$P(A, B) = P(A)P(B) \quad \text{independent events} \tag{1.15}$$

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

$$P(A \cup B) = P(A) + P(B) \quad \text{mutually exclusive events} \tag{1.16}$$

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 because, $\frac{1}{18}$, 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 $\frac{1}{3}$. 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

$$\begin{aligned}
1 &= P(A_1 \cup A_2 \cup A_3 \cup \ldots|B) + P(\overline{A_1 \cup A_2 \cup A_3 \cup \ldots}|B) && (1.17)\\
&= P(A_1|B) + P(A_2 \cup A_3 \cup \ldots|B) + 0 && (1.18)\\
&= P(A_1|B) + P(A_2|B) + P(A_3 \cup \ldots|B) && (1.19)\\
&= \sum_i P(A_i|B) && (1.20)
\end{aligned}$$

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

$$\sum_i P(B|A_i)P(A_i) = \sum_i P(B, A_i) = \sum_i P(A_i|B)P(B) = P(B)\sum_i P(A_i|B) = P(B) \tag{1.21}$$

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?*

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 position each atoms 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.

2.1 Rolling Dice

Say we roll a die 10 times. Lets consider the following questions:

What is the probability of getting at least one 6? This is an "or" question - What is the probability of the first roll being 6 or the second one being six or ... Lets call the event that the ith roll is a 6 A_i. The sum rule (1.14) applies, but since these are not mutually exclusive events the sum rule 1.16 does not. These are independent events since the outcome of any one does not effect the outcome of any other. We could successive apply the extended sum rule (1.14 and the product rule (1.15) to $P(A_1 \cup A_2 \cup \cdots \cup A_{10})$ to break it down into $P(A_i)$'s which we know is 1/6. However, a quicker way to the answer is to realize that the probability of at least one being 6 is 1 minus the probability that non are 6. This follows from the logical requirement that $\overline{A_1 \cup A_2 \cup \cdots \cup A_{10}} = \overline{A_1}, \overline{A_2}, \ldots, \overline{A_{10}}$. Using the original sum rule (1.2) we get symbolically

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_{10}) &= 1 - P(\overline{A_1 \cup A_2 \cup \cdots \cup A_{10}}) && (2.1)\\ &= 1 - P(\overline{A_1}, \overline{A_2}, \ldots, \overline{A_{10}}) && (2.2)\\ &= 1 - P(\overline{A_1})P(\overline{A_2}) \ldots P(\overline{A_{10}}) && (2.3)\\ &= 1 - P(\overline{A})^{10} && (2.4)\\ &= 1 - \left(\frac{5}{6}\right)^{10} && (2.5)\\ &= 0.838\ldots && (2.6) \end{aligned}$$

We could also solve this problem by counting. How many combinations of rolls are there? The first roll has 6 possibilities, the second one 6, etc. so there are 6^{10} combinations. There are 5^{10} combinations with no 6s. So the fraction of the cases that have no 6s is $\left(\frac{5}{6}\right)^{10}$ so the probability of having 1 or more is $\left(\frac{5}{6}\right)^{10}$.

What is the probability of getting exactly one 6? Lets first try to solve this problem by pure symbolic logic and the rules of probability. The proposition could be stated as roll one is a 6 *and* all the others are not *or* roll two is a 6 *and* all the other are not *or* etc. Symbolically this is represented as

$$B_1 = (A_1, \overline{A_2}, .., \overline{A_{10}}) \cup (\overline{A_1}, A_2, .., \overline{A_{10}}) \cup \cdots \cup (\overline{A_1}, \overline{A_2}, .., A_{10}) \tag{2.7}$$

Each of the propositions in the parenthesis are mutually exclusive so the sum rule (1.16) can be applied to B_1 to break it up into a sum

$$P(B_1) = P(A_1, \overline{A_2}, .., \overline{A_{10}}) + P(\overline{A_1}, A_2, .., \overline{A_{10}}) + \ldots \tag{2.8}$$

Since each of the rolls are identical, the probabilities for reach of situation must be the same and

each term must be the same

$$P(B_1) = 10P(A_1, \overline{A_2}, .., \overline{A_{10}}) \tag{2.9}$$
$$= 10P(A_1)P(\overline{A_2}, .., \overline{A_{10}}) \tag{2.10}$$
$$= 10\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^9 \tag{2.11}$$
$$= 0.323\ldots \tag{2.12}$$

where we use the same logic that got us from equation (2.2) to line (2.5) in the previous problem.

Now lets do this again by counting. There are 6^{10} possible combinations. If one roll is a 6 the other nine need to be less than 6. There are 5^9 combinations of nine numbers between 1 and 5. The 6 can come up on any of 10 rolls so there are in total 10^9 ways of rolling 10 times and getting one 6.

What is the probability of getting exactly four 6s? This can be confusing, but if we just look at it from a symbolic point of view we can avoid some common misunderstandings. Here we must find all the combinations of four As and six $\overline{A}$. The first A can go in one of ten slots and the second in one of the remaining 9, etc. giving $10 \times 9 \times 8 \times 7 = 10!/(10-4)!$. We have over counted here though because the order in which we place the As in the slots should not matter, it gives the same logical statement. How many orderings are there? For each selection of 4 slots there are four choices for the first one, three choices etc. - 4! orderings or **permutations**. So there are $\frac{10!}{4!(10-4)!}$ ways of having four As and six $\overline{A}$. The probability of all these combinations are equal and mutually exclusive (a roll cannot be both A and $\overline{A}$) so we can add their probabilities

$$P(B_4) = \frac{10!}{4!(10-4)!}P(A_1, A_2, A_3, A_4, \overline{A_5}, .., \overline{A_{10}}) \tag{2.13}$$
$$= \frac{10!}{4!(10-4)!}P(A_1, A_2, A_3, A_4)P(\overline{A_5}, .., \overline{A_{10}}) \tag{2.14}$$
$$= \frac{10!}{4!(10-4)!}P(A)^4P(\overline{A})^6 \tag{2.15}$$
$$= \frac{10!}{4!(10-4)!}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6 \tag{2.16}$$
$$= 0.05\ldots \tag{2.17}$$

A confusion with this problem often arises because it is often stated or implied that all the permutations of the 6s must be considered one combination because they are indistinguishable. This might lead one to consider any two repeated numbers that are not 6s as indistinguishable and try not to over count them. This quickly becomes a very complex calculation. Although it is true that the 6s are indistinguishable this misses the point. For the purposes of this problem each roll has a binary outcome. It is either a 6 or not a 6. 6s are indistinguishable, but so are not 6s. We could have considered a different problem – "What is the probability of getting 4 rolls that are more than 4?". The calculation would be exactly the same except that the probabilities $P(A)$ and $P(\overline{A})$ would be different, $\frac{1}{3}$ and $\frac{2}{3}$ instead of $\frac{1}{6}$ and $\frac{5}{6}$.

These dice throwing problems are a special case of the **binomial distribution** which we will discuss later in more detail.

2.2 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

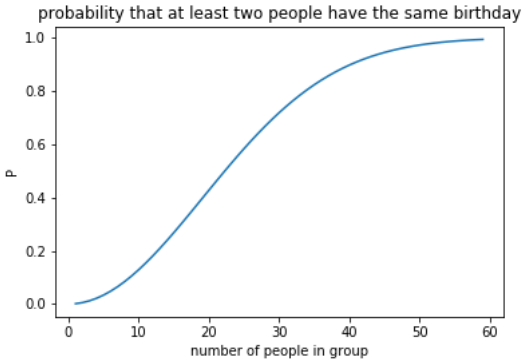

Figure 1: Probability of more than one person having the same birthday.

probability that two of them will have the same birthday.

Lets 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 of combinations with repeat birthdays lets 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

$$P(\text{at least two the same}) = 1 - P(\text{no two the same}) = 1 - \frac{365!}{365^n (365 - n)!}. \tag{2.18}$$

If you try to calculate this number in your directly with a computer you will find that some of these numbers are too big to store. The 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.

$$\ln\left(\frac{N!}{N^n (N-n)!}\right) = \ln N! - \ln(N-n)! - n \ln N \tag{2.19}$$

$$= N \ln N - N - (N-n)\ln(N-n) - (N-n) - n \ln N \tag{2.20}$$

$$= (N-n)\ln N - (N-n)\ln(N-n) - n \tag{2.21}$$

$$= (N-n)\ln\left(\frac{N}{(N-n)}\right) - n \tag{2.22}$$

We can then take the exponential of this to get

$$P(\text{at least two the same}) \simeq 1 - \left(\frac{N}{N-n}\right)^{N-n} e^{-n} \tag{2.23}$$

This is plotted in figure 1. For a group of 23 people there is a 50% chance that at least 2 of them will have the same birthday.

2.3 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 $(13-1)/(52-1)$.

$$P(\text{flush}) = \frac{4}{4}\frac{12}{51}\frac{11}{50}\frac{10}{49}\frac{9}{48} = 0.00198\ldots \tag{2.24}$$

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

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 below of 10 or lower or there wont be enough cards of higher value. That probability is $4 \times 9/52$. Then there are 4 cards of one higher value out of 51 remaining cards, etc.. Then for each case there are 5! permutations.

$$P(\text{straight}) = 5!\frac{36}{52}\frac{4}{51}\frac{4}{50}\frac{4}{49}\frac{4}{48} = 0.003546\ldots \tag{2.25}$$

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).

Lets do this one a little differently. Lets 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. Lets 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 $N \cdot (N-1)... \cdot (N-n+1)$ 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 $n!$. So the number of these distinct sets is

$$\binom{N}{n} \equiv \frac{N!}{n!(N-n)!} \tag{2.26}$$

This is the **binomial coefficient**. In English this is often spoken as "N choose n." for obvious reasons. Lets 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 need to be different than the pair so there are $12 \cdot \binom{4}{3}$ of them. So the probability of a full house is

$$P(\text{full house}) = \frac{\binom{4}{2} \cdot \binom{4}{3} \cdot 13 \cdot 12}{\binom{52}{5}} = 0.00144\ldots \tag{2.27}$$

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 long correspondence in the 17th century in which they developed basic probability theory.

Problem 2. **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 3. *If you roll a die 10 times what is the probability of getting one 1, two 2s, three 3s and four 4s?*

Problem 4. *You have a bag of 100 blue and yellow balls. 60 of them are blue and 40 of them are yellow.*

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

$$\binom{n+r-1}{n} \tag{2.28}$$

combinations of flavors you could order.

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 others 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.

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.

$$p(x) \geq 0 \tag{3.1}$$

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 (1.17) to these bins. In the infinitesimal limit the some becomes an integral

$$\int_{-\infty}^{\infty} dx\ p(x) = 1 \tag{3.2}$$

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 $< x$:

$$F(x) = \int_{-\infty}^{x} dx' p(x') \tag{3.3}$$

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

- **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

$$E\left[f(x)\right] = \langle f(x) \rangle = \begin{cases} \sum_x p(x)\ f(x) \\ \int_{-\infty}^{\infty} dx\ p(x)\ f(x) \end{cases} \tag{3.4}$$

- **mode** - A point where a distribution has a maximum. **Unimodal** distributions have one mode and **multimodal** distributions have more than one.
- **median** - The point in the distribution where $F(x) = 1/2$. 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.
- **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 peak 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

$$X = \frac{(x-\mu)}{\sigma}. \tag{3.5}$$

  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:

$$p(x) = \frac{\gamma}{\pi\left[(x-x_o)^2+\gamma^2\right]} \quad \text{Cauchy-Lorentz distribution.} \tag{3.6}$$

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 (Try proving this.). 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!).

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.

Note that

$$Var[x] = E\left[(x-\bar{x})^2\right] = E\left[x^2 - 2x\bar{x} + \bar{x}^2\right] \tag{3.7}$$
$$= E\left[x^2\right] - 2E\left[x\right]\bar{x} + \bar{x}^2 \tag{3.8}$$
$$= E\left[x^2\right] - \bar{x}^2. \tag{3.9}$$

3.3 moment generating function

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

$$m_x(t) = \langle e^{tx} \rangle = \begin{cases} \sum_x e^{tx} p(x) \\ \int_{-\infty}^{+\infty} dx\ e^{tx} p(x) \end{cases} \tag{3.10}$$

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

$$\left.\frac{d^n m_x(t)}{dt^n}\right|_{t=0} = \langle x^n \rangle \tag{3.11}$$

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. If t is replaced with it it is the Fourier transform.

3.4 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 differentiated. Then for a change dy, x changes by $dx = \left[\frac{d}{dy} f(y)\right] dy$. The probability of being within this range of should not depend on which variable is used to measure the range so it must be that

$$p(x)dx = p\left(f(y)\right) \frac{df}{dy} dy \tag{3.12}$$

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

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

$$p(x_1, x_2, \dots)dx_1 dx_2 \cdots = p(y_1, y_2, \dots) \left|\frac{\partial x}{\partial y}\right| dy_1 dy_2 \dots \tag{3.13}$$

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

$$p\left(x = r\sin(\theta)\cos(\phi), y = r\sin(\theta)\cos(\phi), z = r\cos(\theta)\right)\ r^2 \sin(\theta) dr d\theta d\phi. \tag{3.14}$$

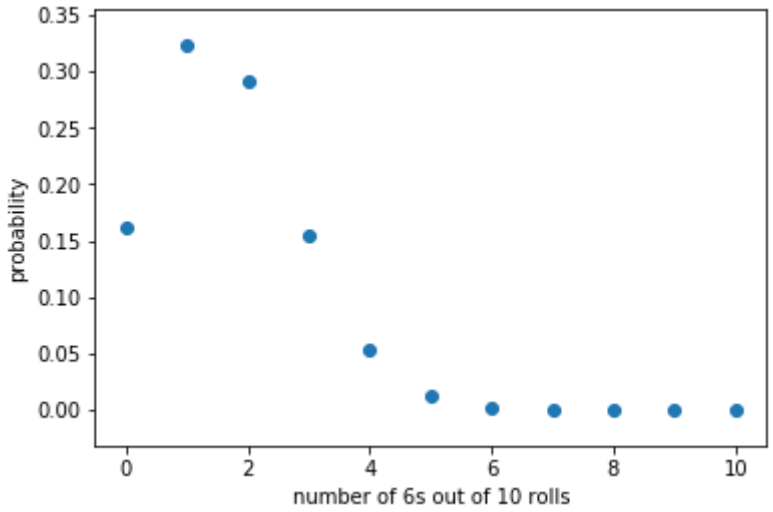

Figure 2: The binomial distribution for the number of 6s in ten rolls of a die or one roll of ten dice. $N = 10$, $k = 0 \ldots 10$, $p = 1/6$

3.5 Binomial and Bernoulli

Say there is some experiment or observation and for each trial the probability of having some outcome A is p. The probability of not having this outcome will be $1 - p$. Each trial is statistically independent. In N trials what is the probability of having n A's?

Using the product rule for independent events we know that the probability of getting n A's in a row is p^n and the probability of having the other be not A is $(1-p)^{N-n}$. This is just like for the dice rolls we discussed earlier. This is the probability each set of N with n A's. Now we need to count how many combinations there are. Since we are not concerned with the order the number is our friend the **binomial coefficient**

$$\binom{N}{n} \equiv \frac{N!}{n!(N-n)!} \tag{3.15}$$

So we get the final result

$$p(n|N) = \binom{N}{n} p^n (1-p)^{N-n} \qquad n \leq N \tag{3.16}$$

which is called the binomial distribution. The case of $N = 10$ and $p = 1/6$ is shown in figure 2. We can now calculate the number of getting any number of 6s out of any number of dice rolls.

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

$$p(n) = \begin{cases} p & , \quad n = 1 \\ (1-p) & , \quad n = 0 \end{cases} \tag{3.17}$$

, an almost trivial case, but perhaps the first probability distribution written down.

The binomial distribution is important for calculating the distribution of any finite sample of observations and comes up a lot in statistics as we will see.

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

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k} \tag{3.18}$$

Using this expansion we can find the moment generating function

$$m_x(t) = \sum_{n=0}^{\infty} e^{tn} \binom{N}{n} p^n (1-p)^{N-n} \tag{3.19}$$

$$= \sum_{n=0}^{\infty} \binom{N}{n} (e^t p)^n (1-p)^{N-n} \tag{3.20}$$

$$= \left(e^t p + 1 - p\right)^N \tag{3.21}$$

From this we can find the mean and variance

$$\langle n \rangle = Np \tag{3.22}$$

$$\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = Np(1-p) \tag{3.23}$$

3.5.1 drawing without replacement, the hypergeometric distribution

Let us briefly consider the case where there are a finite number of objects of two types, we select them at random and we do not replace them 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 of red ones and $N-R$ blue ones. The probability of getting r red ones out of n tries *without replacement* is

$$p(r|n,N,R) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \tag{3.24}$$

Note that $p(r|1,N,R) = R/N$ and $p(r|N,N,R) = \delta_{Rr}$ as they should. The probability of a flush in 5 card stud would be $4\times p(5|5,52,13)$ and in 7 card stud $4\times[p(5|7,52,13) + p(6|7,52,13) + p(7|7,52,13)]$ (see §2.3).

3.6 Poisson distribution

Lets 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 lets find the probability of *no* events happening within a finite range, t_o to $t+dt$. Lets 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

$$p(0|t_o, t+dt) = p(0|t_o, t)\,[1 - rdt] \tag{3.25}$$

Rearranging this we can obtain the differential equation

$$\frac{p(0|t_o, t+dt) - p(0|t_o, t)}{dt} = \frac{d}{dt} p(0|t_o, t) = -p(0|t_o, t) r \tag{3.26}$$

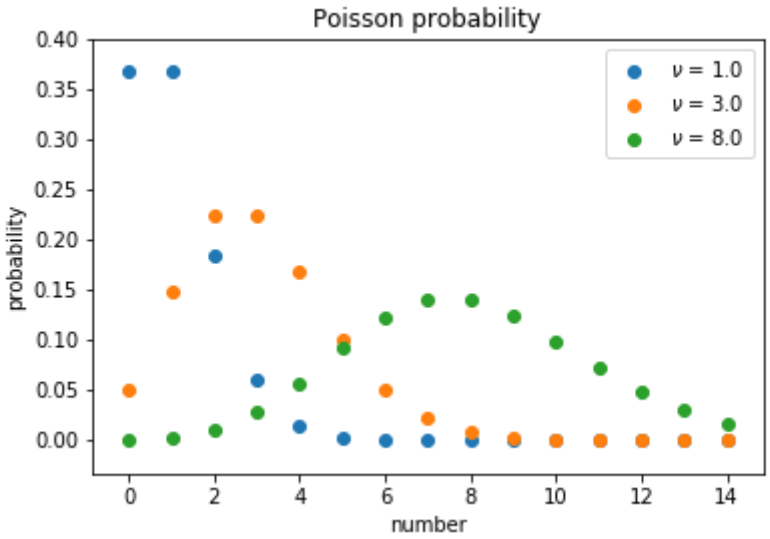

Figure 3: The Poisson distribution for several rates ν.

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,

$$p(0|t_o, t) = e^{-r(t-t_o)} \tag{3.27}$$

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

$$p(0 < t_1 < t_2 < \cdots < t_n < t) = p(0|0, t_1) r dt_1 \Theta(t_1 < t_2) p(0|t_1, t_2) r dt_2 \Theta(t_2 < t_3) \ldots p(0|t_n, t) dt_n \Theta(t_n < t) \tag{3.28}$$

$$= r^n e^{-rt} \prod_{i=1}^{n} dt_i \Theta(t_i < t_{i+1}) \tag{3.29}$$

where

$$\Theta(x < y) = \begin{cases} 1 & , \quad x \leq y \\ 0 & , \quad x > y \end{cases} \tag{3.30}$$

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.

$$p(n|r,t) = \prod_i \int_0^{t_i} p(0 < t_1 < t_2 < \cdots < t_n < t) \tag{3.31}$$

$$= r^n e^{-rt} \int_0^t dt_n \ldots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1 \tag{3.32}$$

$$= r^n e^{-rt} \int_0^t dt_n \ldots \int_0^{t_3} dt_2 t_2 \tag{3.33}$$

$$= r^n e^{-rt} \int_0^t dt_n \ldots \int_0^{t_4} dt_3 \frac{t_3^2}{2} \tag{3.34}$$

$$= \frac{(rt)^n}{n!} e^{-rt} \tag{3.35}$$

$$= \frac{(\nu)^n}{n!} e^{-\nu} \quad \text{Poisson Distribution} \tag{3.36}$$

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

$$E[n] = \nu \tag{3.37}$$

$$Var[n] = \nu \tag{3.38}$$

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.

3.6.1 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 one galaxy being in v equal to $p = \frac{v}{V}$.

Now lets 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 highest order.

$$\binom{N}{n} p^n (1-p)^{N-n} = \binom{N}{n} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \tag{3.39}$$

$$= \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \qquad \text{using } \frac{N!}{(N-n)!} \simeq N^n \tag{3.40}$$

$$\simeq \frac{\nu^n}{n!} e^{-\nu} \tag{3.41}$$

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 is when $\nu \gg 1$ to treat n as continuous and replace $n!$ with the gamma function

$$p(n|\nu) \simeq \frac{\nu^n}{\Gamma(x+1)} e^{-\nu} \qquad \nu \gg 1 \tag{3.42}$$

3.7 Gaussian and normal

Gaussian and normal 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

$$p(x|\sigma,\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{3.43}$$

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 (3.43). I will use $\mathcal{G}(x\,|\mu,\sigma)$ to signify this Gaussian function.

The *cumulative distribution function* is

$$F(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \tag{3.44}$$

with the **error function** defined as

$$\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du \tag{3.45}$$

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

The *moment generating function* is

$$m_{x-\mu}(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ e^{tx} e^{-\frac{x^2}{2\sigma^2}} \tag{3.46}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ \exp\left[-\left(\frac{x}{\sqrt{2}\sigma} - \frac{t\sigma}{\sqrt{2}}\right)^2 + \frac{t^2\sigma^2}{2}\right] \tag{3.47}$$

$$= e^{\frac{1}{2}\sigma^2 t^2} \tag{3.48}$$

The moments are

$$\mu_n = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ x^n e^{-\frac{x^2}{2\sigma^2}} = \begin{cases} \sigma^n (n-1)!! & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \tag{3.49}$$

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

$$!!n = n \cdot (n-2) \cdot (n-4) \ldots 1 \tag{3.50}$$

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

$$p(\mu - n\sigma \leq x \leq \mu + n\sigma) = 1 - F(\mu - n\sigma) - [1 - F(\mu + n\sigma)] \tag{3.51}$$

$$= \frac{1}{2}\left[\text{erf}\left(\frac{n}{\sqrt{2}}\right) - \text{erf}\left(-\frac{n}{\sqrt{2}}\right)\right] \tag{3.52}$$

$$= \text{erf}\left(\frac{n}{\sqrt{2}}\right) \tag{3.53}$$

some specific values for this are

$$p(\mu - \sigma \leq x \leq \mu + \sigma) = 0.683 \tag{3.54}$$

$$p(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.954 \tag{3.55}$$

$$p(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.997 \tag{3.56}$$

$$p(\mu - 4\sigma \leq x \leq \mu + 4\sigma) = 0.999937 \tag{3.57}$$

3.8 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.

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

$$z_i = \frac{x_i - \mu}{\sigma}. \tag{3.58}$$

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

$$Y = \frac{1}{\sqrt{N}} Z = \frac{1}{\sqrt{N}} \sum_i z_i. \tag{3.59}$$

This will again have mean zero and variance 1. Now lets find the moment generating function for

Y,

$$m_Y(t) = \langle \exp(tY) \rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} \sum_i z_i\right) \right\rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} z_i\right) \right\rangle^N \tag{3.60}$$

$$= \left\langle 1 + \frac{t}{\sqrt{N}} z_i + \frac{t^2}{N} \frac{z_i^2}{2} + \frac{t^3}{N^{3/2}} \frac{z_i^3}{3!} + \dots \right\rangle^N \tag{3.61}$$

$$= \left[1 + \frac{t}{\sqrt{N}} \langle z_i \rangle + \frac{t^2}{N} \frac{\langle z_i^2 \rangle}{2} + \frac{t^3}{N^{3/2}} \frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N \tag{3.62}$$

$$= \left[1 + \frac{t^2}{2N} + \frac{t^3}{N^{3/2}} \frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N \tag{3.63}$$

$$\simeq \lim_{N\to\infty} \left[1 + \frac{t^2}{2N} \right]^N \tag{3.64}$$

$$= e^{\frac{t^2}{2}} \tag{3.65}$$

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.

3.8.1 The distribution of the sum of independent random variables

Lets 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. Lets 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

$$p(S) = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \; \delta(S - \sum_i x_i) \; p_1(x_1) \dots p_n(x_n) \tag{3.66}$$

$$= \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \; \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} \exp\left[-ik(S - \sum_i x_i) \right] \; p_1(x_1) \dots p_n(x_n) \tag{3.67}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \int_{-\infty}^{\infty} dx_i e^{+ikxi} p_i(x_i) \tag{3.68}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \tilde{p}_i(k) \tag{3.69}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \tilde{p}_S(k) \tag{3.70}$$

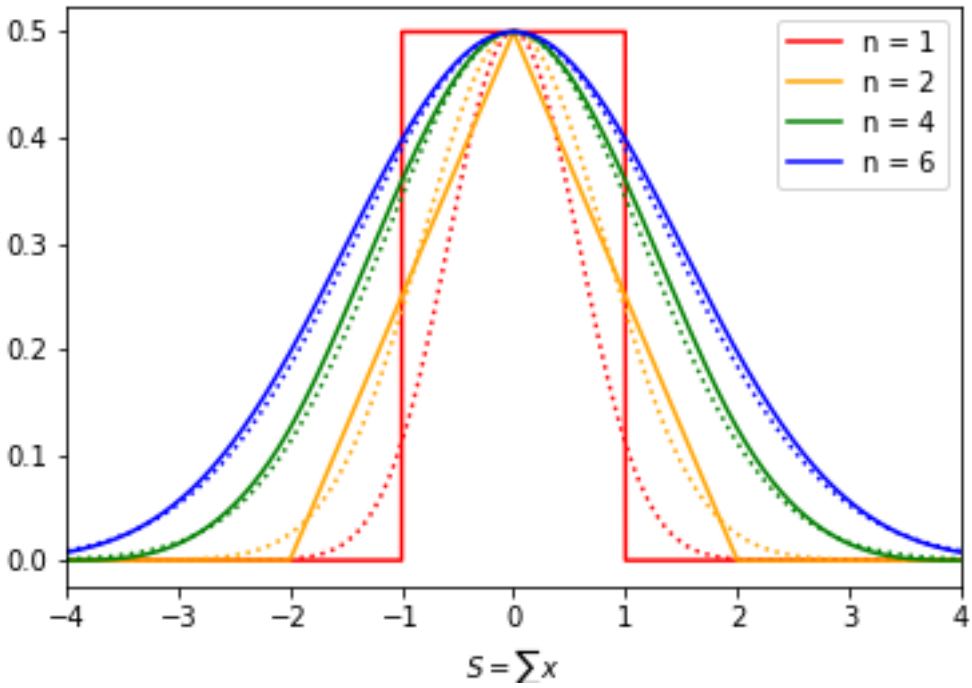

Figure 4: Probability distribution for the sum of n random variables that are uniformly distributed between -1 and 1. The normalizations have been changed so that there maximum is 0.5 in all cases. The dotted curves are for Gaussians with the same variance. You can see that the distribution converges to Gaussian remarkably quickly even for a very non Gaussian initial distribution.

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. 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 $\tilde{p}(0) = 1$.

Lets look at a uniform distribution between $-L/2$ and $L/2$. The Fourier transform of this distribution is

$$\tilde{p}(k) = \frac{1}{L}\int_{-L/2}^{L/2} dx\ e^{+ikx} = \frac{2}{Lk}\sin\left(\frac{kL}{2}\right) = \text{sinc}\left(\frac{kL}{2}\right). \tag{3.71}$$

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

$$p_n(S) = \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS}\text{sinc}^n\left(\frac{kL}{2n}\right). \tag{3.72}$$

Figure 4 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 (Discrete Fourier Transformation) and inverse DFT will do the 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.

3.9 connection between Poisson and Gaussian distributions

You can see from the figure 3 of the Poisson distribution that as the average gets larger the Poisson pdf gets more symmetric and looks more Gaussian. Lets make this connection more precise. The Poisson distribution is

$$p(n|\nu) = \frac{(\nu)^n}{n!}\, e^{-\nu} \tag{3.73}$$

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

$$n! \to \sqrt{2\pi n} e^{-x} n^n \tag{3.74}$$

Making this substitution we get the probability

$$p(n) = \frac{\nu^{\nu(1+\delta)} e^{-\nu}}{\sqrt{2\pi} e^{-\nu(1+\delta)} \left[\nu(1+\delta)\right]^{\nu(1+\delta)+1/2}} \tag{3.75}$$

$$= \frac{e^{\nu\delta}(1+\delta)^{-\nu(1+\delta)-1/2}}{\sqrt{2\pi\nu}} \tag{3.76}$$

Lets look at the lowest order terms of the numerator

$$\ln\left[(1+\delta)^{-\nu(1+\delta)-1/2}\right] = -(\nu(1+\delta)+1/2)\ln(1+\delta) \tag{3.77}$$

$$= -\left(\nu+\nu\delta+1/2\right)\left(\delta - \frac{\delta^2}{2} + \ldots\right) \qquad \nu \gg 1 \tag{3.78}$$

$$\simeq -\left(\nu+\nu\delta\right)\left(\delta - \frac{\delta^2}{2} + \ldots\right) \tag{3.79}$$

$$\simeq -\nu\delta - \frac{\nu\delta^2}{2} + \ldots \tag{3.80}$$

Putting this back into the above

$$p(\delta) = p(n)\nu \tag{3.81}$$

$$\simeq \sqrt{\frac{\nu}{2\pi}}\, e^{-\frac{\nu\delta^2}{2}}. \tag{3.82}$$

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.

3.10 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

$$p(y)dy = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right\} \frac{dy}{y} & , \quad y > 0 \\ 0 & , \quad y \leq 0 \end{cases} \tag{3.83}$$

Some of its properties are

$$E[y] = \exp(\mu + \frac{\sigma^2}{2}) \tag{3.84}$$

$$\text{median[y]} = \exp(\mu) \tag{3.85}$$

$$\text{mode[y]} = \exp(\mu - \sigma^2) \tag{3.86}$$

$$Var[y] = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2) \tag{3.87}$$

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 multivariant Gaussian distribution (section 3.14). This is sometimes also used as a model for density fluctuations in the Universe.

3.11 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. 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

$$p(x|x_{\min}, x_{\max}, \alpha) = x^\alpha \times \begin{cases} 0 & , \quad x < x_{\min} \\ (\alpha+1)\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]^{-1} & , \quad x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \ln\left(\frac{x_{\max}}{x_{\min}}\right)^{-1} & , \quad x_{\min} < x < x_{\max}, \alpha = -1 \\ 0 & , \quad x > x_{\max} \end{cases} \tag{3.88}$$

The cumulative distribution is easily worked out

$$F(x|x_{\min}, x_{\max}, \alpha) = \begin{cases} 0 & , \quad x < x_{\min} \\ \frac{\left[x^{\alpha+1} - x_{\min}^{\alpha+1}\right]}{\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]} & , \quad x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \frac{\ln\left(\frac{x}{x_{\min}}\right)}{\ln\left(\frac{x_{\max}}{x_{\min}}\right)} & , \quad x_{\min} < x < x_{\max}, \alpha = -1 \\ 1 & , \quad x > x_{\max} \end{cases} \tag{3.89}$$

as is the mean and variance.

3.12 multivariant distributions

A multivariant distribution is the probability distribution for the joint probability of two or more random variables. Lets 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 $f(x_1, x_2, \ldots, x_k)$

$$E[f(x_1, x_2, \ldots, x_k)] = \int \ldots \int dx_1 \ldots dx_k \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \tag{3.90}$$

$$= \prod_{i=1}^{k} \int dx_i \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \tag{3.91}$$

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

$$C_{ij} = Cov[x_i x_j] \equiv E[(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)] \tag{3.92}$$

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 B).

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,

$$\rho_{xy} \equiv \frac{C_{xy}}{\sigma_x \sigma_y} \tag{3.93}$$

$Cov[xy]$ 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**

$$|Cov[xy]|^2 \leq Var[x]Var[y] \tag{3.94}$$

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

Another important relation is

$$C_{xy} = E[xy] - \bar{x}\bar{y} \tag{3.95}$$

which is an extension to the relation we already saw for the variance (3.9).

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**.

3.13 multinomial distributions

The binomial distribution can be extended to the case where there are multiple possible outcomes of each trial. The probabilities are p_1, p_2, ... p_k and these are all the possible outcomes so $\sum_i p_i = 1$. The occurrence of each of theses is x_1, x_2, ... x_k. The probability of any sequence of these will be $\prod_i p_i^{x_i}$ (Look back at the dice throwing example again.). There are $N!$ such sequences for N trials, but for each one with x_i there are $x_i!$ permutations that are the same. Thus

$$P(x_1, x_2, x_3, \ldots, x_k | N, \{p_i\}) = \frac{N!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} = \frac{N!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i} \tag{3.96}$$

The mean and variance of the distribution are

$$E[x_i] = N p_i \tag{3.97}$$

$$Var[x_i] = N p_i (1 - p_i) \tag{3.98}$$

And the covariance is

$$Cov[x_i x_j] = -N p_i p_j \tag{3.99}$$

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.

3.14 multivariant gaussian

The mutivariant Gaussian or normal distribution is by far the most often used multivariant distribution. It is a good approximation to many natural phenomena and is often used even when it is not. It is also very useful when trying to understand some statistical argument or principle to put in a multivariant 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 multivariant 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 multivariant Gaussian is a generalization of the one dimensional Gaussian pdf.

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right] \tag{3.100}$$

$$\equiv \mathcal{G}(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.101}$$

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 $\mathcal{G}(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{C})$. 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.

Theorem 3.1 *The means of the multivariant Gaussian are*

$$E[x_i] = \mu_i \quad or \quad E[\boldsymbol{x}] = \boldsymbol{\mu} \tag{3.102}$$

Theorem 3.2 *And the covariances of the multivariant Gaussian are*

$$Cov[x_i, x_j] = E[(x_i - \mu_i)(x_j - \mu_j)] = C_{ij} \quad or \quad Cov[\boldsymbol{x}, \boldsymbol{x}] = E[(\boldsymbol{x} - \boldsymbol{\mu})^T (\boldsymbol{x} - \boldsymbol{\mu})] = \boldsymbol{C} \tag{3.103}$$

So $\boldsymbol{C}$ is the correlation matrix as the choice of notation suggests. $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

For the **special case of a diagonal covariance matrix**, the diagonal elements are the σ^2's. The covariance matrix will take the form

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 & \dots \\ 0 & \sigma_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.104}$$

In this case there are no correlations between different variables.

PROOF OF MEAN: (theorem 3.1)

Lets calculate the means first

$$E[x_i] = \int_{-\infty}^{\infty} dx_i \dots \int_{-\infty}^{\infty} dx_i \dots \int_{-\infty}^{\infty} dx_n \; x_i \; p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.105}$$

$$\tag{3.106}$$

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 $mu'_i = 0$,

$$E[x_i] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx'_i \dots \int_{-\infty}^{\infty} dx_n \; (\mu_i + x'_i) \; p(\boldsymbol{x'}|\boldsymbol{\mu'}, \boldsymbol{C}) \tag{3.107}$$

$$= \mu_i \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx'_i \dots \int_{-\infty}^{\infty} dx_n \; p(\boldsymbol{x'}|\boldsymbol{\mu'}, \boldsymbol{C}) + \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx'_i \dots \int_{-\infty}^{\infty} dx_n \; x'_i \; p(\boldsymbol{x'}|\boldsymbol{\mu'}, \boldsymbol{C}) \tag{3.108}$$

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 3.2)

$$Corr[\boldsymbol{x}, \boldsymbol{x}] = \int_{-\infty}^{\infty} d^n x \; (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T \; p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.109}$$

$$= \int_{-\infty}^{\infty} d^n y \; \boldsymbol{y}^T \boldsymbol{y} \; p(\boldsymbol{y}|0, \boldsymbol{C}) \qquad\qquad \boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu} \tag{3.110}$$

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

$$\boldsymbol{C} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^{-1} \tag{3.111}$$

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

$$\mathbf{M}^T = \mathbf{M}^{-1} \tag{3.112}$$

$$|\mathbf{M}| \equiv \det(\mathbf{M}) = 1 \tag{3.113}$$

The columns of $\boldsymbol{M}$ are the eigenvectors of $\boldsymbol{C}$.

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

$$e^{\frac{1}{2}\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x}} \; d^n x = e^{\frac{1}{2}\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{x}} \; d^n x = e^{\frac{1}{2}(\boldsymbol{M}^T\boldsymbol{x})^T\boldsymbol{\Sigma}(\boldsymbol{M}^T\boldsymbol{x})} \; d^n x = e^{\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}\boldsymbol{y}} \; |\boldsymbol{M}| d^n y = e^{\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}\boldsymbol{y}} \; d^n y \tag{3.114}$$

$$Corr[\boldsymbol{x},\boldsymbol{x}] = \int_{-\infty}^{\infty} d^n x\ \boldsymbol{x}\boldsymbol{x}^T\ p(\boldsymbol{x}|0,\boldsymbol{C}) \tag{3.115}$$

$$= \int_{-\infty}^{\infty} d^n y\ (\boldsymbol{M}\boldsymbol{y})(\boldsymbol{M}\boldsymbol{y})^T\ p(\boldsymbol{y}|0,\boldsymbol{\Sigma}) \tag{3.116}$$

$$= \int_{-\infty}^{\infty} d^n y\ \boldsymbol{M}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{M}^T\ p(\boldsymbol{y}|0,\boldsymbol{\Sigma}) \tag{3.117}$$

$$= \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T \tag{3.118}$$

$$= \boldsymbol{C} \tag{3.119}$$

The transformed or rotated variables $\boldsymbol{y} = \mathbf{M}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ will be uncorrelated. This is the basis of the **principle components** or **PCA** decomposition of the data that we will get to later.

3.14.1 conditional Gaussian distribution

Lets break the parameters, $\boldsymbol{x}$, into two set, $\boldsymbol{y}$ and $\boldsymbol{z}$. We will fix the parameters $\boldsymbol{z}$ and ask what the psf 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.

Lets 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, $\mathbf{C}_{xy}$.

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{zy} \\ \mathbf{C}_{zy}^T & \mathbf{C}_{zz} \end{bmatrix} \tag{3.120}$$

The conditional pdf is then

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}'_y, \boldsymbol{\Sigma}_{yy}\right) \qquad \begin{cases} \boldsymbol{\mu}'_y = \boldsymbol{\mu}_y + \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{yy} = \boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T \end{cases} \tag{3.121}$$

which means

$$p(\boldsymbol{y}|\boldsymbol{z}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}_y - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z)\right)^T \left(\boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}_y - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z)\right)\right] \tag{3.122}$$

3.14.2 marginalized Gaussian distribution

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

$$p(\boldsymbol{y}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{x}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{y},\boldsymbol{z}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}) \tag{3.123}$$

Using the same definitions (without proof) this is

$$p(\boldsymbol{y}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}_y, \boldsymbol{C}_{yy}\right) \tag{3.124}$$

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

The proof for the conditional and marginal distributions in the general case are rather long algebraicly. 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

$$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \mathbf{b}^T\boldsymbol{x} = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\mathbf{b}\right)^T \boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\mathbf{b}\right) - \frac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b} \tag{3.125}$$

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

3.14.3 combining two multivariant Gaussians

$$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_1, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_2, \boldsymbol{C}_2\right) = \mathcal{G}\left(\boldsymbol{\mu}_1\,|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) \tag{3.126}$$

$$\boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2 \tag{3.127}$$

$$\boldsymbol{\mu}_c = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{C}_1\boldsymbol{\mu}_1 + \boldsymbol{C}_2\boldsymbol{\mu}_2\right) \tag{3.128}$$

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

Theorem 3.3 *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* $\boldsymbol{s} \sim \mathcal{N}(0, \boldsymbol{C}_1 + \boldsymbol{C}_2)$.

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

$$\begin{aligned}
p(\boldsymbol{s}) &= \int_{-\infty}^{\infty} d^n x \int_{-\infty}^{\infty} d^n x'\; p(\boldsymbol{x}, \boldsymbol{x}')\delta({}^D\boldsymbol{s} - \boldsymbol{x} - \boldsymbol{x}') && (3.129)\\
&= \int_{-\infty}^{\infty} d^n x\; p(\boldsymbol{x}, \boldsymbol{s} - \boldsymbol{x}) && (3.130)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|0, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{s} - \boldsymbol{x}\,|0, \boldsymbol{C}_2\right) && (3.131)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|0, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{s}, \boldsymbol{C}_2\right) && (3.132)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{s}\,|0, \boldsymbol{\Sigma}\right) && (3.133)\\
&= \mathcal{G}\left(\boldsymbol{s}\,|0, \boldsymbol{\Sigma}\right)\int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) && (3.134)\\
&= \mathcal{G}\left(\boldsymbol{s}\,|0, \boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2\right) && (3.135)
\end{aligned}$$

In particular if

$$\boldsymbol{C}_1 = \sigma_1^2 \quad \text{and} \quad \boldsymbol{C}_1 = \sigma_2^2 \tag{3.136}$$

then

$$\begin{aligned}
&\boldsymbol{C}_1^{-1} = \tfrac{1}{\sigma_1^2} \quad \text{and} \quad \boldsymbol{C}_1^{-1} = \tfrac{1}{\sigma_2^2}\\
&\boldsymbol{\Sigma} = \sigma_1^2 + \sigma_2^2\\
&\boldsymbol{\Sigma}^{-1} = (\sigma_1^2 + \sigma_2^2)^{-1}\\
&\boldsymbol{\mu}_c = \tfrac{\mu_1\sigma_1^2 + \mu_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}
\end{aligned} \tag{3.137}$$

3.15 χ^2 distribution

The χ^2 distribution is not a multivariant distribution, but is closely related to the multivariate Gaussian. Consider a multivariate Gaussian distribution with uncorrelated variable, or equivalently

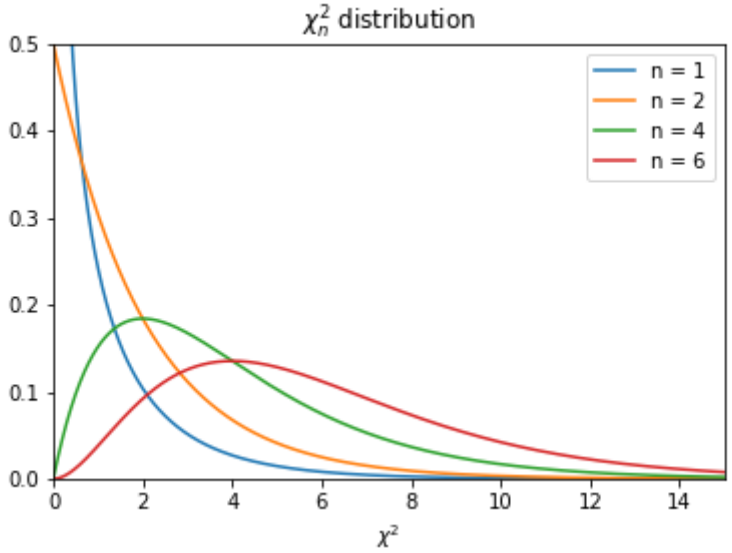

Figure 5: χ^2_n distribution for some different degrees of freedom, n.

a diagonal covariance. Lets define a new variables

$$z = \sum_i^N \frac{(x_i - \mu_i)^2}{\sigma_i^2}. \tag{3.138}$$

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 is

$$p(x_1, x_2, \ldots x_N)dx_1 \ldots dx_N = \frac{1}{(2\pi)^{N/2}\prod_i \sigma_i} e^{-\frac{1}{2}\sum_i^N \frac{(x_i-\mu_i)^2}{\sigma_i^2}} dx_1 \ldots dx_N \tag{3.139}$$

$$= \frac{1}{(2\pi)^{N/2}\prod_i \sigma_i} e^{-\frac{1}{2}z} dx_1 \ldots dx_N \tag{3.140}$$

$$\frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}z} dx'_1 \ldots dx'_N \qquad x' = \frac{x-\mu}{\sigma} \tag{3.141}$$

z can be seen as the square of the radial coordinate in N dimensional space

$$dx'_1 \ldots dx'_N = r^{n-1} dr d\theta_1 d\theta_3 \cdots = z^{n/2-1} dz d^n\Omega \tag{3.142}$$

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

$$p(z = \chi^2|n) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{3.143}$$

where the **gamma function** is defined as

$$\Gamma(x) \equiv \int_0^\infty dt\ e^{-t} t^{x-1}. \tag{3.144}$$

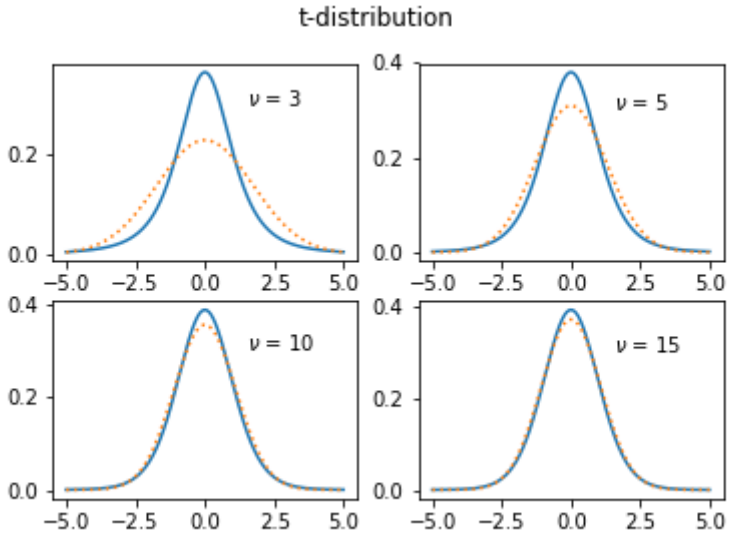

Figure 6: Student's t distribution for some different degrees of freedom, ν. The dotted curves are Gaussians with the same variances for comparison.

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 5.

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

$$F(x|2) = 1 - e^{-x/2} \tag{3.145}$$

Theorem 3.4 *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 $\sim \chi^2$.

Problem 6. *Prove theorem 3.4.*

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

$$p_t(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{3.146}$$

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 6.

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. They are used extensively in frequentist statistics. 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 only practical statistics.

In this chapter we will look at some of the basic properties of a finite sample drawn from a random 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. Lets say there are N samples denotes $x_1,\dots x_N$ and they are all independent draws from the distribution.

The **arithmetic mean** of this data is

$$\bar{x}_N \equiv \frac{1}{N}\sum_{i=0}^{N} x_i \tag{4.1}$$

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, these are distinct concepts. $E[x]$ is a sum over all possible values of x weighted by the pdf and $\bar{x}_N$ is an unweighted sum over a finite sample.

We can take the expectation value of the arithmetic mean

$$\langle \bar{x}_N \rangle = \frac{1}{N}\sum_{i=0}^{N} \langle x_i \rangle \tag{4.2}$$

$$= \frac{1}{N}\sum_{i=0}^{N} \mu \tag{4.3}$$

$$= \mu \tag{4.4}$$

This means that the arithmetic mean of a sample is an estimate 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,

$$Var[\bar{x}_N] = \left\langle [\bar{x}_N - \mu]^2 \right\rangle \tag{4.5}$$

$$= \left\langle [\text{Mean}(\{x\})]^2 \right\rangle - 2\mu\langle \text{Mean}(\{x\})\rangle + \mu^2 \tag{4.6}$$

$$= \left\langle [\text{Mean}(\{x\})]^2 \right\rangle - \mu^2 \tag{4.7}$$

$$= \left\langle \left[\frac{1}{N} \sum_{i=0}^{N} x_i \right]^2 \right\rangle - \mu^2 \tag{4.8}$$

$$= \frac{1}{N^2} \sum_{i=0}^{N} \sum_{j=0}^{N} \langle x_i x_j \rangle - \mu^2 \tag{4.9}$$

$$= \frac{1}{N^2} \left[\sum_{i=0}^{N} \langle x_i^2 \rangle + \sum_{i \neq j} \langle x_i x_j \rangle \right] - \mu^2 \tag{4.10}$$

$$= \frac{1}{N^2} \left[\sum_{i=0}^{N} (\sigma^2 + \mu^2) + \sum_{i \neq j} \langle x_i \rangle \langle x_j \rangle \right] - \mu^2 \tag{4.11}$$

$$= \frac{1}{N^2} \left[N(\sigma^2 + \mu^2) + N(N-1)\mu^2 \right] - \mu^2 \tag{4.12}$$

$$= \frac{\sigma^2}{N} \tag{4.13}$$

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. Of course to calculate this variance we need to know the underlying variance, σ^2, which we sometimes do not know.

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.

$$\text{if} \quad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \sigma) \quad \text{then} \quad \text{Mean}(\{x\}) \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\sigma}{\sqrt{N}}\right) \tag{4.14}$$

It often happens that one is making repeated measurement 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. Neither the less the thing you want to know, the luminosity of the star, should be constant. The arithmetic mean (4.2) 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

$$\hat{\theta} = \sum_i w_i x_i \tag{4.15}$$

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

$$\sum_i w_i = 1. \tag{4.16}$$

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

$$\sigma_\theta^2 = \langle \theta^2 \rangle - \mu^2 \tag{4.17}$$

$$= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 \tag{4.18}$$

$$= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 \tag{4.19}$$

$$= \sum_i w_i^2 \langle x_i^2 \rangle + \sum_{i \neq j} w_i w_j \langle x_i \rangle \langle x_j \rangle - \mu^2 \tag{4.20}$$

$$= \sum_i w_i^2 \left[\sigma_i^2 + \mu^2 \right] + \mu^2 \sum_{i \neq j} w_i w_j - \mu^2 \tag{4.21}$$

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

$$F(\boldsymbol{w}) = \sigma_\theta^2(\boldsymbol{w}) + \lambda \left(1 - \sum_i w_i \right) \tag{4.22}$$

that is

$$\frac{\partial F}{\partial w_k} = \frac{\partial \sigma_\theta^2}{\partial w_k} - \lambda = 0 \tag{4.23}$$

The derivative of the variance is

$$\frac{\partial \sigma_\theta^2}{\partial w_k} = 2 w_k \left[\sigma_k^2 + \mu^2 \right] + 2\mu^2 \sum_{i \neq k} w_i \tag{4.24}$$

$$= 2 w_k \left[\sigma_k^2 + \mu^2 \right] + 2\mu^2 \left[\sum_{i=0}^{N} w_i - w_k \right] \tag{4.25}$$

$$= 2 w_k \left[\sigma_k^2 + \mu^2 \right] + 2\mu^2 \left[1 - w_k \right] \qquad \text{use constraint } \sum_i w_i = 1 \tag{4.26}$$

$$= 2 w_k \sigma_k^2 + 2\mu^2 \tag{4.27}$$

putting this into (4.23) gives

$$w_k = \frac{\lambda - 2\mu}{2\sigma_k^2} \tag{4.28}$$

Plugging this into the constraint (4.16) and solving for

$$\lambda = 2\mu + 2 \left[\sum_k \frac{1}{\sigma_k^2} \right]^{-1} \tag{4.29}$$

so

$$w_k = \left[\sum_i \frac{1}{\sigma_i^2} \right]^{-1} \frac{1}{\sigma_k^2} \tag{4.30}$$

So the estimator (4.15) is

$$\hat{\theta} = \frac{1}{\left[\sum_i \frac{1}{\sigma_i^2}\right]} \sum_i \frac{x_i}{\sigma_i^2}. \tag{4.31}$$

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

Lets 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 characteristic 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

$$S_N^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 \tag{4.32}$$

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

Unkown mean: In most cases one does not know the average ahead of time. In this case the best estimator is

$$S_N^2 = \frac{1}{N-1}\sum_i \left(x_i - \bar{x}_N\right)^2. \tag{4.33}$$

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

$$\begin{aligned}
\left\langle S_N^2 \right\rangle &= \frac{1}{N-1}\sum_i \left\langle \left(x_i - \bar{x}_N\right)^2 \right\rangle && (4.34)\\
&= \frac{1}{N-1}\left[\sum_i \left\langle x_i^2 \right\rangle - 2\left\langle \sum_i x_i \bar{x}_N \right\rangle + \sum_i \left\langle (\bar{x}_N)^2 \right\rangle\right] && (4.35)\\
&= \frac{1}{N-1}\left[\sum_i (\sigma^2 + \mu^2) - 2N\left\langle (\bar{x}_N)^2 \right\rangle + N\left\langle (\bar{x}_N)^2 \right\rangle\right] && (4.36)\\
&= \frac{1}{N-1}\left[\sum_i (\sigma^2 + \mu^2) - N\left\langle (\bar{x}_N)^2 \right\rangle\right] && (4.37)\\
&= \frac{1}{N-1}\left[N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)\right] \qquad \text{using (4.13)} && (4.38)\\
&= \sigma^2 && (4.39)
\end{aligned}$$

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 (4.33) then $\left\langle s_N^2 \right\rangle = (N-1)\sigma/N$ 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 4.1 *If $x_i \sim \mathcal{N}(\mu,\sigma)$ and S_N is given by (4.33) then $z = \frac{(N-1)S_N^2}{\sigma^2}$ is χ^2_{N-1} distributed.*

Proof:

$$\frac{(N-1)S_N^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_i (x_i - \bar{x})^2 \tag{4.40}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i-\mu) - (\bar{x}-\mu)\right]^2 \tag{4.41}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i-\mu)^2 - 2(x_i-\mu)(\bar{x}-\mu) - (\bar{x}-\mu)^2\right] \tag{4.42}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i-\mu)^2\right] - 2N(\bar{x}-\mu)(\bar{x}-\mu) - N(\bar{x}-\mu)^2 \tag{4.43}$$

$$= \sum_i \frac{(x_i-\mu)^2}{\sigma^2} - \frac{N(\bar{x}-\mu)^2}{\sigma^2} \tag{4.44}$$

This is the difference of two χ^2 distributed quantities: $(\bar{x}-\mu)^2/(\sigma^2/N) \sim \chi^2_1$ and $\sum_i (x_i-\mu)^2 \sim \chi^2_N$. By theorem 3.4 the sum of the χ^2 distributed is χ^2 distributed. QED

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

$$\begin{array}{lll} \left\langle \frac{(N-1)}{\sigma^2} S_N^2 \right\rangle = N-1 & \Rightarrow & \left\langle S_N^2 \right\rangle = \sigma^2 \\ Var\left[\frac{(N-1)}{\sigma^2} S_N^2\right] = \left\langle \left(\frac{(N-1)}{\sigma^2} S_N^2\right)^2 \right\rangle - \left\langle \frac{(N-1)}{\sigma^2} S_N^2 \right\rangle^2 = 2(N-1) & \Rightarrow & Var\left[S_N^2\right] = \frac{2\sigma^4}{(N-1)} \end{array} \tag{4.45}$$

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

$$P\left(\frac{\sigma^2}{(N-1)} z_1 < S_N^2 < \frac{\sigma^2}{(N-1)} z_2\right) = F_{\chi^2_{N-1}}(z_2) - F_{\chi^2_{N-1}}(z_1) \tag{4.46}$$

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

$$z = (\bar{x}-\mu)\sqrt{\frac{n}{S_n^2}} \tag{4.47}$$

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

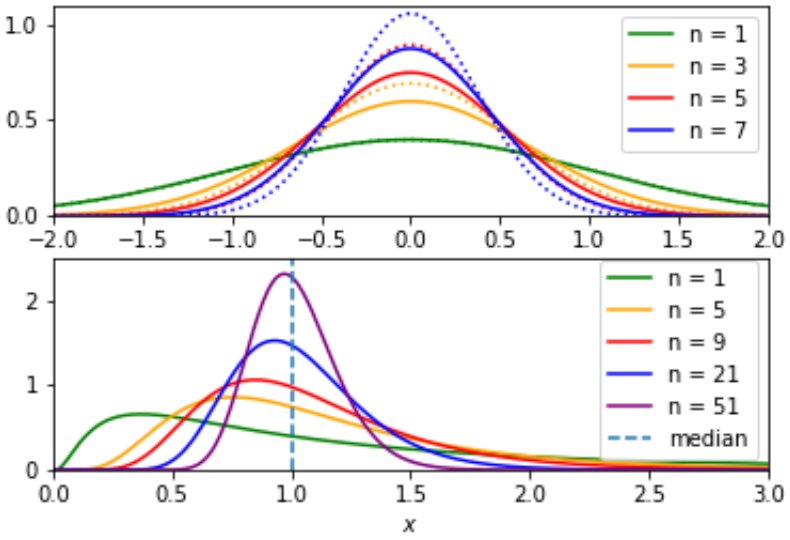

Figure 7: The probability of the sample median for normal (above) and lognormal (below) distributions. The $n = 1$ case is the original distribution. The dotted curves in the normal case are the distributions of the sample means based on the same n's.

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

$$p(\mu - \delta x < \bar{x} < \mu + \delta x) = \int_{-\delta x/\sqrt{\frac{n}{S_n^2}}}^{+\delta x/\sqrt{\frac{n}{S_n^2}}} dt\ p_t(t|\nu = n-1) \tag{4.48}$$

$$= \sqrt{\frac{n}{S_n^2}} \int_{-\delta x}^{+\delta x} dx'\ p_t\left(x'\sqrt{\frac{n}{S_n^2}}\middle|\ \nu = n-1\right) \tag{4.49}$$

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.

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. Lets 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

$$p_m(x_{\rm med}|N) = Np(x_{\rm med})P_{\rm binom}\left(\frac{(N-1)}{2}\middle| F(x_{\rm med}), N-1\right) \tag{4.50}$$

$$= N\binom{N-1}{\frac{N-1}{2}}p(x_{\rm med})F(x_{\rm med})^{\frac{N-1}{2}}\left[1-F(x_{\rm med})\right]^{\frac{N-1}{2}} \tag{4.51}$$

$$= N\binom{2n}{n}p(x_{\rm med})F(x_{\rm med})^{n}\left[1-F(x_{\rm med})\right]^{n} \tag{4.52}$$

where $N=2n+1$.

Lets find what the mode of this distribution. The log of the probability is

$$\ln p_m(x) = \ln p(x) + n\ln F(x) + n\ln[1-F(x)] + \mathcal{C}. \tag{4.53}$$

Taking the derivative of this and setting it to zero gives

$$\frac{1}{p(\hat{x})}\frac{\partial p(\hat{x})}{\partial x} + n\frac{p(\hat{x})}{F(\hat{x})} - n\frac{p(\hat{x})}{1-F(\hat{x})} = 0 \tag{4.54}$$

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

$$F(\hat{x}) = \frac{1}{2}. \tag{4.55}$$

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

Lets expand the log-pdf for the median, (4.53) around the mode $\hat{x}$,. First

$$\ln\left[F(x)\right] \simeq \ln\left[F(\hat{x})\right] + \frac{1}{F(\hat{x})}F'(\hat{x})(x-\hat{x}) + \frac{1}{2}\left(\frac{1}{F(\hat{x})}F''(\hat{x}) - \frac{1}{F(\hat{x})^2}(F'(\hat{x}))^2\right)(x-\hat{x})^2 + \ldots \tag{4.56}$$

$$= -\ln[2] + 2p(\hat{x})(x-\hat{x}) + \left(p'(\hat{x}) - 2(p(\hat{x}))^2\right)(x-\hat{x})^2 + \ldots \tag{4.57}$$

and

$$\ln\left[1-F(x)\right] \simeq -\ln[2] + 2p(\hat{x})(x-\hat{x}) - \left(p'(\hat{x}) + 2(p(\hat{x}))^2\right)(x-\hat{x})^2 + \ldots \tag{4.58}$$

so

$$\ln p_m(x) \simeq \ln p_m(\hat{x}) - 4n(p(\hat{x}))^2(x-\hat{x})^2 + \mathcal{C} \tag{4.59}$$

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

$$p_m(x) = \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(x-\hat{x})^2}{2\sigma^2}} \tag{4.60}$$

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

$$Var[x_{\rm med}] = \sigma_m^2 \simeq \frac{1}{8np(x_{\rm med})^2} \simeq \frac{1}{4Np(x_{\rm med})^2} \quad N \gg 1 \tag{4.61}$$

For $x \sim \mathcal{N}(\mu, \sigma)$ the sample mean has a smaller variance than the sample median by a factor of $\sim \frac{2}{\pi}\frac{(N+2)}{N}$. 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

$$\begin{aligned} p_{\rm max}(x|N) &= Np(x)P_{\rm binom}\left(N-1|F(x), N-1\right) && (4.62)\\ &= Np(x)P_{\rm binom}\left(0|1-F(x), N-1\right) && (4.63)\\ &= Np(x)F(x)^{N-1} && (4.64) \end{aligned}$$

4.6 quintile estimation

The q-**quintiles** 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-quintile. The nth q-quintile 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-quintile. 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.

$$\begin{aligned} p(x_n|N) &= Np(x_n)P_{\rm binom}\left(r-1|F(x_n), N-1\right) && (4.65)\\ p(x_n|N) &= Np(x_n)P_{\rm binom}\left(\frac{nN}{q} - \frac{1}{2}\middle| F(x_n), N-1\right) && (4.66) \end{aligned}$$

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 quintile range the distribution or estimate a range that contains some fixed probability, say 68% or 95%. One might use (4.65) with an estimate of the true pdf to judge how well the range can be estimated.

Problem 7. *Find the minimum variance weighting for an estimator of the the variance in the form*

$$S_w^2 = \sum_i w_i (x_i - \bar{x})^2 \tag{4.67}$$

where the variance, σs, of each measurement are not equal.

Problem 8. *Prove that the variance of the quintile distribution is*

$$Var[x_q] \simeq \frac{p(1-p)}{Np(x_q)^2} \tag{4.68}$$

where $p = n/q$. For example the 95% lower bound would have $p = 1/20 = 0.05$ and the 99% upper bound would have $p = 99/100$. This will be use when doing Monte Carlo tests later.

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 its 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 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 buy only the values of a model's parameters (the planet has unknown mass m). Lets let I represent everything else in the Universe that we will take to be fixed or irrelevant to our experiment (existence of the apperatise, the day of the week, the phase of the moon on a distant planet). We apply Bayes's theorem to this situation

$$P(M_i|\boldsymbol{D}, I) = \frac{P(\boldsymbol{D}|M_i, I)P(M_i|I)}{P(\boldsymbol{D}|I)} \tag{5.1}$$

$$= \frac{P(\boldsymbol{D}|M_i, I)P(M_i|I)}{\sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)} \tag{5.2}$$

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 confidence 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, is 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.

- $P(\boldsymbol{D}) = \sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)$ 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 subtlety is usually ignored by scientists.

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

$$p(M|\boldsymbol{D}_1) = \frac{p(M)p(\boldsymbol{D}_1|M)}{p(\boldsymbol{D}_1)} \quad p(M|\boldsymbol{D}_2) = \frac{p(M)p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_2)} \tag{5.3}$$

Now lets look at the posterior for both data sets,

$$p(M|\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{p(M)p(\boldsymbol{D}_1, \boldsymbol{D}_2|M)}{p(\boldsymbol{D}_1, \boldsymbol{D}_2)} \tag{5.4}$$

$$= \frac{p(M)p(\boldsymbol{D}_1|M)p(\boldsymbol{D}_2|\boldsymbol{D}_1, M)}{p(\boldsymbol{D}_1, \boldsymbol{D}_2)} \qquad \text{product rule} \tag{5.5}$$

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

$$p(M|\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{p(M)p(\boldsymbol{D}_1|M)p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_1)p(\boldsymbol{D}_2)} \tag{5.6}$$

$$= \frac{p(M)p(\boldsymbol{D}_1|M)}{p(\boldsymbol{D}_1)} \frac{p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_2)} \tag{5.7}$$

$$= p(M|\boldsymbol{D}_1)\frac{p(\boldsymbol{D}_2|M)}{p(\boldsymbol{D}_2)} \qquad \text{using 5.3} \tag{5.8}$$

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**. More on this later perhaps.

5.3 Parameter estimation

The most common use for Bayesian inference is parameter estimation. 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

$$P(\theta_1, \theta_2, \ldots | \boldsymbol{D}) = \frac{\mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots) p(\theta_1, \theta_2, \ldots)}{\left[\int d^n\theta \; \mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots) p(\theta_1, \theta_2, \ldots)\right]} \tag{5.9}$$

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

5.3.1 example: Poisson radiation

Lets 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 hard in many cases).

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

$$p(n|r) = \frac{(r\delta t)^n}{n!} e^{-r\delta t} \tag{5.10}$$

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,

$$p(r) = \frac{\Theta(0 < r < r_{max})}{r_{max}} \tag{5.11}$$

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.

$$\mathcal{E}(n) = \int_{-\infty}^{\infty} dr\ p(n|r)p(r) = \frac{1}{r_{max}} \int_0^{r_{max}} dr\ \frac{(r\delta t)^n}{n!}\, e^{-r\delta t} \tag{5.12}$$

$$= \frac{\delta t^{-1}}{n! r_{max}} \int_0^{\delta t r_{max}} dx\ x^n\, e^{-x} \qquad x = r\delta t \tag{5.13}$$

$$\simeq \frac{\delta t^{-1}}{n! r_{max}} \int_0^{\infty} dx\ x^n\, e^{-x} \qquad r_{max} \gg 1/\delta t \tag{5.14}$$

$$= \frac{\delta t^{-1}}{n! r_{max}} \Gamma(n+1) \tag{5.15}$$

$$= \frac{1}{\delta t r_{max}} \qquad \text{because } \Gamma(n+1) = n! \tag{5.16}$$

So combining (5.16), (5.16), and (5.16) the posterior for the rate is

$$p(r|n) = \frac{\delta t}{n!}(\delta t r)^n e^{-r\delta t} \tag{5.17}$$

The normalization of the prior, r_{max}, drops out.

The average of this distribution is

$$\langle r \rangle = \int_o^{\infty} dr\ r p(r|n) = \frac{\delta t}{n!} \int_o^{\infty} dr\ r(\delta t r)^n e^{-r\delta t} = \frac{1}{\delta t n!} \int_o^{\infty} dx\ x^{n+1} e^{-x} \tag{5.18}$$

$$= \frac{(n+1)!}{\delta t n!} = \frac{(n+1)}{\delta t} \tag{5.19}$$

and the variance is

$$Var[r] = \frac{(n+1)}{\delta t^2} \tag{5.20}$$

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

$$\frac{\partial}{\partial r} \ln p(r|n) = \frac{\partial}{\partial r} \left([\ln(r\delta t) - r\delta t - \ln(\delta t/n!)] \right) \tag{5.21}$$

$$= \frac{n}{r} - \delta t \tag{5.22}$$

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

$$r_{mode} = \frac{n}{\delta t}. \tag{5.23}$$

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

5.3.2 example: estimating mean

Lets 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

$$d_i = \theta + n_i, \tag{5.24}$$

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

$$\mathcal{L}(d|\theta) = \mathcal{G}\left(d\,|\theta,\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d-\theta)^2}{2\sigma^2}} \tag{5.25}$$

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

$$p(\theta) = \begin{cases} \frac{1}{\theta_{\max}-\theta_{\min}} & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.26}$$

$$= \mathcal{C}\Theta(\theta_{\min} < \theta < \theta_{\max}) \qquad \mathcal{C} \equiv \frac{1}{\theta_{\max} - \theta_{\min}} \tag{5.27}$$

You might be concerned that the parameters $\theta_{\max}$ 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 lets 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

$$\mathcal{L}(\boldsymbol{d}|\theta) = \mathcal{G}\left(d_1\,|\theta,\sigma\right) \times \mathcal{G}\left(d_2\,|\theta,\sigma\right) \times \ldots \tag{5.28}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2}\sum_i \frac{(d_i-\theta)^2}{\sigma^2}\right) \tag{5.29}$$

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

$$\mathcal{L}(\boldsymbol{d}|\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_i \left(d_i^2 - 2d_i\theta + \theta^2\right)\right) \tag{5.30}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_i d_i^2 - 2\sum_i d_i\theta + n\theta^2\right]\right) \tag{5.31}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left[n\overline{d^2} + n(\theta-\bar{d})^2 - n(\bar{d})^2\right]\right) \tag{5.32}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{n}{2\sigma^2}\left[\overline{d^2} - (\bar{d})^2\right]\right) \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{d})^2\right) \tag{5.33}$$

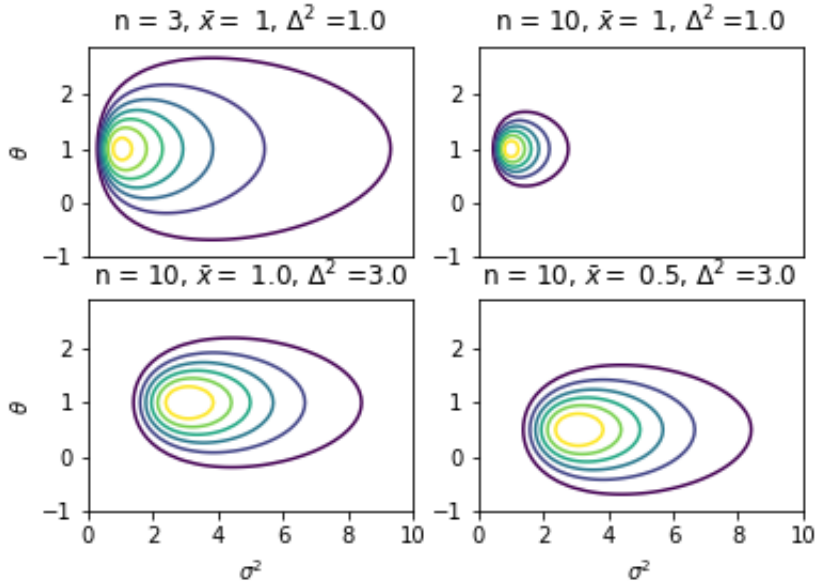

Figure 8: The posterior distributions for the mean and variance based on a sample of Gaussian distributed measurement with the number, sample mean and sample variance given above each one.

where

$$\bar{d} \equiv \frac{1}{n}\sum_i d_i \qquad \overline{d^2} \equiv \frac{1}{n}\sum_i d_i^2. \tag{5.34}$$

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

$$\mathcal{E}(\boldsymbol{d}) = \frac{1}{(2\pi)^{(n-1)/2}\sigma^{n-1}\sqrt{n}} \exp\left(-\frac{n}{2\sigma^2}\left[\overline{d^2} - (\bar{d})^2\right]\right) \tag{5.35}$$

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

$$P(\theta|\boldsymbol{d}) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{d})^2\right) = \mathcal{G}\left(\theta\,|\bar{d}, \sigma^2/n\right). \tag{5.36}$$

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 $\langle\theta\rangle = \boldsymbol{d}$. No surprise here.

5.3.3 example: estimating mean and variance

Lets 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:

$$d_i = \theta + x_i + n_i \tag{5.37}$$

We can think of the likelihood as the probability of the actual alcohol is x and then the probability of the alcohol level x being measured as d,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{d}|\theta,\sigma_n^2,\sigma_a^2) &= \int_{-\infty}^{\infty} d^n x \; P(\boldsymbol{d},\boldsymbol{x}|\theta,\sigma_a^2) && (5.38)\\
&= \int_{-\infty}^{\infty} d^n x \;\; \left[\mathcal{G}\left(d_1\,|x_1,\sigma_n^2\right)\mathcal{G}\left(d_2\,|x_2,\sigma_n^2\right)\dots\right]\left[\mathcal{G}\left(x_1\,|\theta,\sigma_a^2\right)\mathcal{G}\left(x_2\,|\theta,\sigma_a^2\right)\dots\right] && (5.39)\\
&= \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{d}\,|\boldsymbol{x},\sigma_n^2\right)\mathcal{G}\left(\boldsymbol{x}\,|\theta,\sigma_a^2\right) && (5.40)\\
&= \mathcal{G}\left(\boldsymbol{d}\,|\theta,\sigma_n^2+\sigma_a^2\right) && (5.41)
\end{aligned}$$

where we are using the results of section 3.14.3 to combine Gaussian pdfs. This is the same likelihood as we got in the first example except σ^2 is replaced with $\sigma_n^2 + \sigma_a^2$,

$$\mathcal{L}(\boldsymbol{d}|\theta,\sigma_n^2,\sigma_a^2) = \frac{1}{\sqrt{(2\pi)^{n/2}(\sigma_n^2+\sigma_a^2)^n}} \exp\left(-\frac{n\left[\overline{d^2}-(\bar{d})^2\right]}{2(\sigma_n^2+\sigma_a^2)}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2(\sigma_n^2+\sigma_a^2)}\right) \tag{5.42}$$

To make things simpler lets make the following substitutions

$$\begin{aligned}
\Delta^2 &\equiv \overline{d^2} - (\bar{d})^2 && (5.43)\\
\sigma^2 &\equiv \sigma_n^2 + \sigma_a^2 && (5.44)
\end{aligned}$$

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

$$\mathcal{L}(\boldsymbol{d}|\theta,\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.45}$$

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

$$\begin{aligned}
P(\theta,\sigma^2) &= \frac{\Theta(\theta_{\max}<\theta<\theta_{\min})}{(\theta_{\max}-\theta_{\min})}\frac{\Theta(0<\sigma^2<\sigma^2_{\max})}{\sigma^2_{\max}} && (5.46)\\
&= \mathcal{C}\Theta(\theta_{\max}<\theta<\theta_{\min})\Theta(0<\sigma^2<\sigma^2_{\max}) && (5.47)
\end{aligned}$$

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

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

$$\mathcal{E}(\boldsymbol{d}) = \mathcal{C}\int_0^{\sigma^2_{\max}} d\sigma^2 \int_{\theta_{\min}}^{\theta_{\max}} d\theta\ \mathcal{L}(\boldsymbol{d}|\theta,\sigma^2) \tag{5.48}$$

$$\simeq \mathcal{C}\int_0^{\sigma^2_{\max}} d\sigma^2 \int_{-\infty}^{\infty} d\theta\ \mathcal{L}(\boldsymbol{d}|\theta,\sigma^2) \tag{5.49}$$

$$= \frac{\mathcal{C}}{(2\pi)^{n/2}}\int_0^{\sigma^2_{\max}} d\sigma^2 \int_{\theta_{\min}}^{\theta_{\max}} d\theta\ \frac{1}{\sigma^n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.50}$$

$$= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}}\int_0^{\sigma^2_{\max}} d\sigma^2\ \frac{1}{\sigma^{n-1}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \tag{5.51}$$

In doing this we have taken the the range of the θ integral to go to infinity. This is justifiable if $|\sigma^2_{\max}| = |\sigma^2_{\min}| \gg \sigma^2$. 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 lets make the change of variables to

$$y = \sqrt{\frac{n\Delta^2}{2\sigma^2}} \quad \text{so} \quad d\sigma^2 = \frac{n\Delta^2}{y^3}dy \tag{5.52}$$

$$\mathcal{E}(\boldsymbol{d}) = \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}}\left(\frac{n\Delta^2}{2}\right)^{\frac{3-n}{2}}\int_{\sqrt{\frac{n\Delta^2}{2\sigma^2_{\max}}}}^{\infty} dy\ y^{n-4}e^{-y^2} \tag{5.53}$$

$$\simeq \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}}\left(\frac{n\Delta^2}{2}\right)^{\frac{3-n}{2}}\int_0^{\infty} dy\ y^{n-4}e^{-y^2} \tag{5.54}$$

$$= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}}\left(\frac{n\Delta^2}{2}\right)^{\frac{3-n}{2}}\Gamma\left(\frac{n-3}{2}\right) \tag{5.55}$$

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

$$P(\theta,\sigma^2|\boldsymbol{d}) = \frac{1}{\sqrt{2\pi}\Gamma\left(\frac{n-3}{2}\right)}\left(\frac{\Delta^2}{2}\right)^{\frac{n-3}{2}}\left(\frac{n}{\sigma^2}\right)^{\frac{n}{2}}\frac{1}{n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.56}$$

This posterior is plotted in figure 8 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.

$$\ln P(\theta,\sigma^2|\boldsymbol{d}) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2\sigma^2}\left[\Delta^2 + (\theta-\bar{d})^2\right] + \text{constant terms} \tag{5.57}$$

$$\frac{\partial}{\partial\theta}\ln P(\theta,\sigma^2|\boldsymbol{d}) = -\frac{n}{\sigma^2}(\theta-\bar{d}) \tag{5.58}$$

$$\frac{\partial}{\partial\sigma^2}\ln P(\theta,\sigma^2|\boldsymbol{d}) = \frac{n}{2\sigma^2}\left(-1 + \frac{\Delta^2}{\sigma^2} + \frac{(\theta-\bar{d})^2}{\sigma^2}\right) \tag{5.59}$$

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 $(N-1)^{-2}$ factor that we saw was needed to make the estimator unbiased has a been replaced with N^{-2}.

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,

$$\frac{\partial}{\partial \sigma^2} \ln P(\theta, \sigma^2 | \boldsymbol{d}) = \frac{1}{2\sigma} \frac{\partial}{\partial \sigma} \ln P(\theta, \sigma | \boldsymbol{d}) = \frac{1}{4\sigma^3} \frac{\partial}{\partial \sigma^{1/2}} \ln P(\theta, \sigma^{1/2} | \boldsymbol{d}) \tag{5.60}$$

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, $\langle \sigma^2 \rangle \neq \langle \sigma \rangle^2$.

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.

Lets 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.

$$\begin{aligned} P(\alpha_1, \alpha_2, \ldots | \boldsymbol{D}) &= \int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots \; P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots | \boldsymbol{D}) \\ &= \frac{\int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots \; P(\boldsymbol{D} | \alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots) P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots)}{\mathcal{E}(\boldsymbol{D})} \end{aligned} \tag{5.61}$$

5.4.1 example: the mean without the variance

As a simple example lets say we have the posterior (5.56). We are interested in the parameter θ, but we are not interested in the "noise" parameter σ^2. Lets 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. Lets

make the substitution $A = n\Delta^2 + n(\theta - \bar{d})^2$ in which case the relevant parts of the posterior are

$$P(\theta|\Delta^2, \bar{d}) = \int_0^\infty d\sigma^2 \; P(\theta, \sigma^2|\Delta^2, \bar{d}) \tag{5.62}$$

$$\propto \int_0^\infty d\sigma^2 \; \frac{e^{-\frac{A}{2\sigma^2}}}{\sigma^n} \tag{5.63}$$

$$\propto -2\int_\infty^0 dx \; x^{n-3} e^{-\frac{A}{2}x^2} \qquad x = \frac{1}{\sigma} \tag{5.64}$$

$$\propto 2^{\frac{n-3}{2}} A^{-\left(\frac{n-2}{2}\right)} \Gamma\left(\frac{n-2}{2}\right) \qquad \text{integral in Appendix E} \tag{5.65}$$

$$\propto \left[\Delta^2 + \left(\theta - \bar{d}\right)^2\right]^{\frac{2-n}{2}} \tag{5.66}$$

$$\propto \left[1 + \frac{\left(\theta - \bar{d}\right)^2}{\Delta^2}\right]^{\frac{2-n}{2}} \tag{5.67}$$

If we compare this to the t-distribution((3.146) 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

$$P(\theta|\Delta^2, \bar{d}) = \frac{\Gamma\left(\frac{n-2}{2}\right)}{\sqrt{(n-3)\pi}\,\Gamma\left(\frac{n-3}{2}\right)} \left[1 + \frac{(\theta - \bar{d})^2}{\Delta^2}\right]^{\frac{2-n}{2}} \tag{5.68}$$

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

$$\left\langle x^2 \right\rangle = (n-3)\frac{\left\langle(\theta - \bar{d})^2\right\rangle}{\Delta^2} = \frac{\nu}{\nu - 2} = \frac{n-3}{n-5} \tag{5.69}$$

so

$$\left\langle(\theta - \bar{d})^2\right\rangle = \frac{\Delta^2}{(n-5)}. \tag{5.70}$$

From this example we learn that if we model a series of observation 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. That would be the frequentist method of putting a constraint on the distribution mean μ. The number of degrees of freedom are different! More on this later when we talk about the choice of prior.

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 us a uniform 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, but a uniform prior on one choice will not be uniform for the other.

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

$$p(\alpha) = \frac{1}{\ln(\alpha_{max}/\alpha_{min})} \begin{cases} 1/\alpha & \alpha_{min} < \alpha < \alpha_{max} \\ 0 & \text{otherwise} \end{cases} \tag{5.71}$$

This prior gives equal weight to equal logorithmic ranges of α ($d\ln\alpha = d\alpha/\alpha$). If parameter, α, is replaced with parameter $\beta = \alpha^\gamma$ for any γ this prior will not change since $d\ln\alpha^\gamma = \gamma d\ln\alpha \propto d\alpha/\alpha$. 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. A location parameter can be shifted by a constant without fundamentally changing the problem. 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. 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. We will compare and contrast these methods further later.

5.5.1 example: Jeffreys prior

Going back the alcohol in wine example, we can now recagnise 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 (5.56) will change to

$$P(\theta, \sigma^2|\boldsymbol{d}) \propto \left(\frac{1}{\sigma^2}\right) \frac{1}{\sigma^n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.72}$$

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

$$P(\theta, \sigma^2|\boldsymbol{d}) = \frac{n^{n/2}}{\sqrt{2^n\pi}\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{\Delta^2}{2}\right)^{\frac{n-1}{2}} \frac{1}{\sigma^{n+2}} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.73}$$

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

$$P(\theta|\boldsymbol{d}) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\Delta} \left[1 + \frac{(\theta-\bar{d})^2}{\Delta^2}\right]^{-\frac{n}{2}} \tag{5.74}$$

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 averages should be equal.

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.

5.5.2 example: radiation with Jeffreys prior

Going back to the example given in section 5.3.1 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 (5.17) to

$$p(r|n) = \frac{\delta t}{(n-1)!}(\delta t r)^{n-1} e^{-r\delta t} \tag{5.75}$$

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

$$\langle r \rangle = \frac{n}{\delta t} \qquad Var\,[r] = \frac{n}{\delta t^2} \tag{5.76}$$

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

$$r_{mode} = \frac{n-1}{\delta t} \tag{5.77}$$

Figure 9 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.

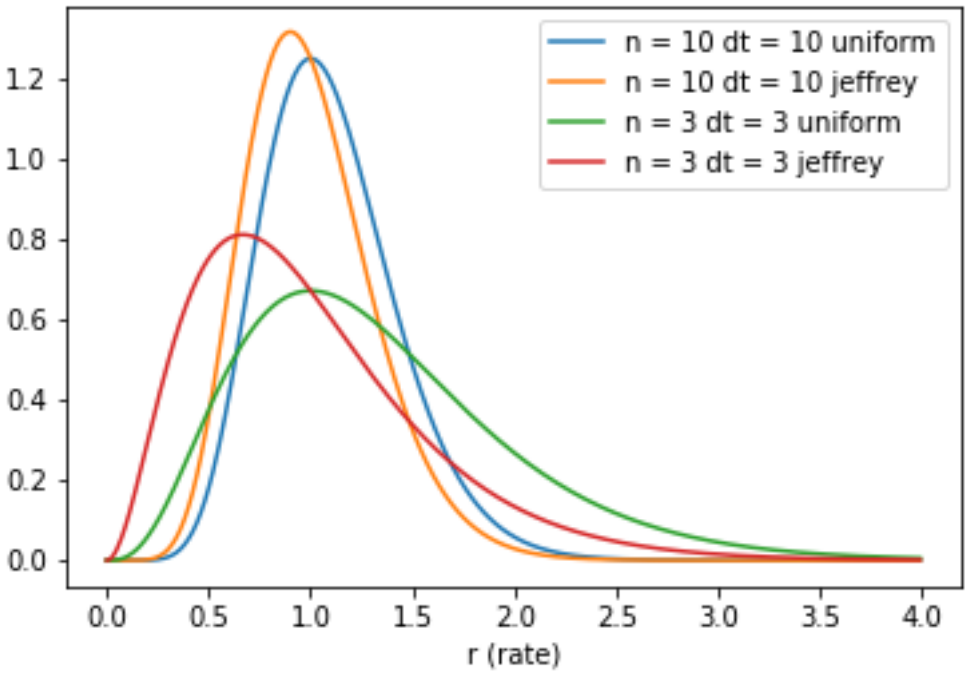

Figure 9: Posteriors for the rate r given n counts in δt time units using uniform and Jeffrey priors on the rate.

5.6 A comment

A final important point about Bayesian parameter estimation: **Bayesian analysis is always relative.** You always compare one model to another or many others. A corollary to this is that **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 meant 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 Model selection

Lets consider a somewhat different problem than parameter estimation. Lets 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

$$P(M_i|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|M_i)P(M_i)}{P(\boldsymbol{D})} = \frac{P(\boldsymbol{D}|M_i)P(M_i)}{\sum_i P(\boldsymbol{D}|M_i)P(M_i)}. \tag{5.78}$$

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

$$O_{1,2} = \frac{P(M_1|\boldsymbol{D})}{P(M_2|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|M_1)}{P(\boldsymbol{D}|M_2)}\frac{P(M_1)}{P(M_2)} = B_{1,2}\frac{P(M_1)}{P(M_2)}. \tag{5.79}$$

$O_{1,2}$ 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 N_2.

How can we calculate $P(\boldsymbol{D}|M)$? In the parameter estimation problem we stayed within one 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

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)}{P(\boldsymbol{D}|M)}. \tag{5.80}$$

We can now see that $P(\boldsymbol{D}|M)$ is actually the evidence:

$$P(\boldsymbol{D}|M_i) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\ P(\boldsymbol{D}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i) = \mathcal{E}(\boldsymbol{D}|M_i) \tag{5.81}$$

where the integral is over all of parameter space within model M_i. Bayes' factor is the ratio of evidences for two models.

5.7.1 Occam's factor

Situation often arises where one has a standard model that explains the data and an extension to the model that includes some additional parameters. For example, the standard ΛCDM 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. The extended model has mixing regulated with an additional parameter.

In these situations 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 that additional parameters fixed to some value (perhaps 0 or for in the dark energy case $w = -1$). Lets 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. Lets 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

$$\mathcal{L}_{ex}(\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta}) \geq \mathcal{L}_{st}(\hat{\boldsymbol{\theta}}_{st}) \tag{5.82}$$

Because of this one might be drawn to the conclusion that more complicated models are always as good or better than less complicated ones. This violates Occam's principle, or razor, that the best model is the simplest one that is consistent with the observations (William of Ackham ~ 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, a parabola, to the data 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. But if your model includes random noise in the data you would not expect the correct model to go through all the points perfectly. "Any theory that fits all the data is wrong, because some of the data is wrong." (I don't know who said this.) 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 Occam's razor to answer these questions. To demonstrate this lets consider an extended model with on extra parameter β. The prior on this parameter will be $\mathcal{N}(\beta_o, \sigma_\beta)$. The standard model will be the extended one with $\beta = \beta_o$. We will take the priors on the models to be equal ($P(M_1) = P(M_2)$). The odds between the models is

$$O_{2,1} = B_{2,1} = \frac{\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta) P(\boldsymbol{\theta}) P(\beta)}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o) P(\boldsymbol{\theta})} \tag{5.83}$$

Now lets consider two extreme cases. The first is where the prior, $P(\beta)$, is very broad compared to the likelihood so that $P(\beta) = \exp(-(\beta - \beta_o)^2/2\sigma_\beta^2)/\sqrt{2\pi\sigma_\beta^2} \simeq 1/\sqrt{2\pi\sigma_\beta^2}$ everywhere $\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta)$ is significant so

$$B_{2,1} \simeq \frac{1}{\sqrt{2\pi}\sigma_\beta} \frac{\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta) P(\boldsymbol{\theta})}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o) P(\boldsymbol{\theta})} \tag{5.84}$$

Now lets express the integrals here as the products of two factors

$$\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta) P(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta}) \mathcal{V}_{\theta,\beta} \tag{5.85}$$

where $\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta})$ is the maximum likelihood and $\mathcal{V}_{\theta,\beta}$ is a measure of the volume in parameter space to which the parameters are constrained by the data in the extended model. Likewise $\mathcal{V}_\theta$ is a measure of the parameter space volume in the standard model. Writing them like this the Bayes' factor becomes

$$B_{2,1} \simeq \left[\frac{1}{\sqrt{2\pi}\sigma_\beta}\frac{\mathcal{V}_{\theta,\beta}}{\mathcal{V}_\theta}\right]\frac{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta})}{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{st}, \beta_o)} \tag{5.86}$$

The part in square brackets is sometimes called Occam's factor. The ratio $\frac{\mathcal{V}_{\theta,\beta}}{\mathcal{V}_\theta}$ can be interpreted as a measure of the width in parameter space of the posterior in the β dimension in the extended model. So Occam's factor is small if the data constrains the parameter β to a much smaller range than was allowed by the prior ($\sim \sigma_\beta$). The other factor is the ratio of the likelihood at its maximum with and without the extra parameters. This factor will always be larger than one and thus favor the extended model. Occam's factor will always be smaller than one in this example where we have taken the prior to be very broad. 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.

Another extreme example 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

$$O_{2,1} = B_{2,1} = \frac{\int d\boldsymbol{\theta}\int d\beta\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta)P(\boldsymbol{\theta})P(\beta)}{\int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})} \tag{5.87}$$

$$\simeq \frac{\int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})\int d\beta\ P(\beta)}{\int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})} \tag{5.88}$$

$$\simeq \frac{\int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})}{\int d\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})} \tag{5.89}$$

$$\simeq 1 \tag{5.90}$$

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 maybe a bit 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.

One criticism of Bayesian model selection is that it depends on you 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 at them. For this reason you can use a uniform or Jeffreys prior that extends to infinity for example. 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. We will find later

that frequentist hypothesis testing offers an alternative to model selection that is more satisfying for many.

5.7.2 example: detection with Gaussian errors

Lets consider the problem of detecting a source. Here we have two models in one there is no source, M_0, and in the second, M_1, there is a source. There is a background flux and Gaussian noise. The noise might be instrumental noise or it might be instrumental noise plus background fluctuations. We will assume the noise in the pixels is uncorrelated. We don't know the noise and background, but there are many pixels that we can model as having only background in them. From these we can find the posterior for the average background flux, θ and the background variance σ. This is the same calculation as we did in example 5.5.1. We will use the posterior for θ and σ as the prior for these parameters in this problem. Ignoring the normalization that will drop out later this is

$$Pr(\theta,\sigma|\bar{d},\Delta) \propto \frac{1}{\sigma^{n+2}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.91}$$

$$\propto \frac{1}{\sigma^{n+1}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\mathcal{G}\left(\theta\left|\bar{d},\frac{\sigma}{\sqrt{n}}\right.\right) \tag{5.92}$$

Now consider the likelihood for the pixel that might have a source in it. Lets say the flux measured in this pixel is f. In model M_0 the likelihood would be

$$\mathcal{L}_0(f|\theta,\sigma) = \mathcal{G}\left(f\,|\theta,\sigma\right) = \mathcal{G}\left(\theta\,|f,\sigma\right) \tag{5.93}$$

We have a Gaussian in the prior and in the likelihood. We can use the Gaussian multiplication rule we learned in section 3.14.3 to simplify the algebra:

$$\mathcal{G}\left(\theta\,|f,\sigma\right)\mathcal{G}\left(\theta\left|\bar{d},\frac{\sigma}{\sqrt{n}}\right.\right) = \mathcal{G}\left(f\left|\bar{d},\sigma\sqrt{1+\frac{1}{n}}\right.\right)\mathcal{G}\left(\theta\left|\frac{nf+\bar{d}}{n+1},\sigma\sqrt{1+\frac{1}{n}}\right.\right) \tag{5.94}$$

So the likelihood times the prior for model M_0 is

$$\mathcal{L}_0(f|\theta,\sigma)Pr(\theta,\sigma|\bar{d},\Delta) \propto \frac{1}{\sigma^{n+1}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\mathcal{G}\left(f\left|\bar{d},\sigma\sqrt{1+\frac{1}{n}}\right.\right)\mathcal{G}\left(\theta\left|\frac{nf+\bar{d}}{n+1},\sigma\sqrt{1+\frac{1}{n}}\right.\right) \tag{5.95}$$

Now we can integrate over θ which is simple because the Gaussian is normalized,

$$\int d\theta\ \mathcal{L}_0(f|\theta,\sigma)Pr(\theta,\sigma|\bar{d},\Delta) \propto \frac{1}{\sigma^{n+1}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\mathcal{G}\left(f\left|\bar{d},\sigma\sqrt{1+\frac{1}{n}}\right.\right) \tag{5.96}$$

Integrating over σ^2 is done just like it was done in the example in section 5.4.1

$$\mathcal{E}_0(\bar{d},\Delta,f) = \int d\theta \int d\sigma^2\ \mathcal{L}_0(f|\theta,\sigma)Pr(\theta,\sigma|\bar{d},\Delta) \tag{5.97}$$

$$\propto \left[1+\frac{(f-\bar{d})^2}{(1+n)\Delta^2}\right]^{-n/2} \tag{5.98}$$

The pre-factor will be the same for both models and drop out of the odds ratio so we don't need to calculate it.

For the model M_1 we have another parameter to represent the extra average flux in this pixel. We will call this parameter s. The likelihood and the prior for θ and σ can be combined in the same way

$$\mathcal{L}_1(f|\theta,\sigma,s)Pr(\theta,\sigma|\bar{d},\Delta) \propto \frac{1}{\sigma^{n+1}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\mathcal{G}\left(f\left|\bar{d}+s,\sigma\sqrt{1+\frac{1}{n}}\right.\right)\mathcal{G}\left(\theta\left|\frac{n(f-s)+\bar{d}}{n+1},\sigma\sqrt{1+\frac{1}{n}}\right.\right) \tag{5.99}$$

Integrating over θ and σ in the same way gives the evidence

$$\mathcal{E}_1(\bar{d},\Delta,f) = \int ds \int d\theta \int d\sigma^2\ \mathcal{L}_1(f|\theta,\sigma,s)Pr(\theta,\sigma|\bar{d},\Delta)Pr(s) \tag{5.100}$$

$$\propto \int ds \left[1+\frac{(f-\bar{d}-s)^2}{(1+n)\Delta^2}\right]^{-n/2} Pr(s) \tag{5.101}$$

Bayes' ratio will be

$$B_{10}(\bar{d},\Delta,f) = \frac{\mathcal{E}_1(\bar{d},\Delta,f)}{\mathcal{E}_0(\bar{d},\Delta,f)} = \frac{\int ds \left[1+\frac{(f-\bar{d}-s)^2}{(1+n)\Delta^2}\right]^{-n/2} Pr(s)}{\left[1+\frac{(f-\bar{d})^2}{(1+n)\Delta^2}\right]^{-n/2}} \tag{5.102}$$

Now we have to choose a prior for s. This is where, in my mind, the ambiguity comes into Bayesian model selection. Lets use a uniform prior from 0 to $s_{\max}$. Unlike in the parameter estimation case, here the normalization of the prior (in this case $1/s_{\max}$) does not drop out.

$$B_{10}(\bar{d},\Delta,f) = \frac{\int_0^{s_{\max}} ds\ \left[1+\frac{(f-\bar{d}-s)^2}{(1+n)\Delta^2}\right]^{-n/2}}{s_{\max}\left[1+\frac{(f-\bar{d})^2}{(1+n)\Delta^2}\right]^{-n/2}} \tag{5.103}$$

The integral can be done, but the result is kind of ugly. Luckily there is an approximation that is valid in many cases. Lets take the number of background pixels to be large. Then we can use the formula we have used before:

$$\lim_{N\to\infty}\left[1+\frac{t^2}{2N\Delta^2}\right]^{-N} = e^{-\frac{t^2}{2\Delta^2}} \tag{5.104}$$

and Bayes ratio becomes

$$B_{10}(\bar{d},\Delta,f) \simeq \frac{e^{\frac{(f-\bar{d})^2}{2\Delta^2}}}{s_{\max}}\int_0^{s_{\max}} ds\ e^{-\frac{(f-\bar{d}-s)^2}{2\Delta^2}} \tag{5.105}$$

$$= \frac{\sqrt{\pi}\Delta}{s_{\max}} e^{\frac{(f-\bar{d})^2}{2\Delta^2}}\left[\mathrm{erf}\left(\frac{f-\bar{d}}{\sqrt{2}\Delta}\right)-\mathrm{erf}\left(\frac{f-\bar{d}-s_{\max}}{\sqrt{2}\Delta}\right)\right] \tag{5.106}$$

You can see that B_{10} gets exponentially larger as $(f-\bar{d})/\Delta$ gets large. This makes sense. If the measured value in this pixel is much larger than the noise, $\sigma \sim \Delta$, then you have strong evidence that there needs to be more flux than the background. But what should we put in for the prior maximum $s_{\max}$? It could be taken from the expected luminosity function of sources. We could have used the expected luminosity function as the prior in the first place. Without such a well defined

and well justified prior the numerical value of the odds ratio doesn't have much meaning, in my opinion.

Sometimes in astronomy the number of photons is so small that the noise is dominated by Poisson counting noise. This is particularly true in the X-rays or higher energy. The same exercise can be done with Poisson noise to assess a detection of a source. I won't go through all of it because it gets a bit ugly with incomplete gamma functions, but you can easily see how this calculation should go. The likelihood for model M_0 where there is only a background is

$$\mathcal{L}_0(n_1 \ldots |\theta) = \prod_{i=1}^{N_p} \frac{\theta^{n_i}}{n_i!} e^{-\theta} = \frac{\theta^{\sum_i n_i}}{\prod_i n_i!} e^{-N_p \theta} \tag{5.107}$$

where N_p is the total number of pixels including and n_i is the observed photon number counts in pixel i. If there are many background pixels the prior on the background and the variance will be very close to the Gaussian prior we used before despite the Poisson errors.

For model M_1 the likelihood (keeping Poisson errors for the background) is

$$\mathcal{L}_1(n_s, n_1 \ldots |\theta, s) = \frac{(\theta + s)^{n_s}}{n_s!} e^{-(\theta+s)} \prod_{i \neq s} \frac{\theta^{n_i}}{n_i!} e^{-\theta} \tag{5.108}$$

$$= \frac{(\theta + s)^{n_s} \theta^{\sum_i^{N_{\rm bk}} n_i}}{n_s! \prod_i n_i} e^{-(1+N_{\rm bk})\theta - s} \tag{5.109}$$

The rest of the calculation goes as before except the integrals are different.

5.8 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 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. However, to get the evidence the likelihood and prior need to be properly normalized.

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.

5.9 Example: luminosity function

Lets consider the problem of measuring the luminosity function from a data set of star (or galaxy or AGN or supernovae, etc.) luminosities. This is the same problem of 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.

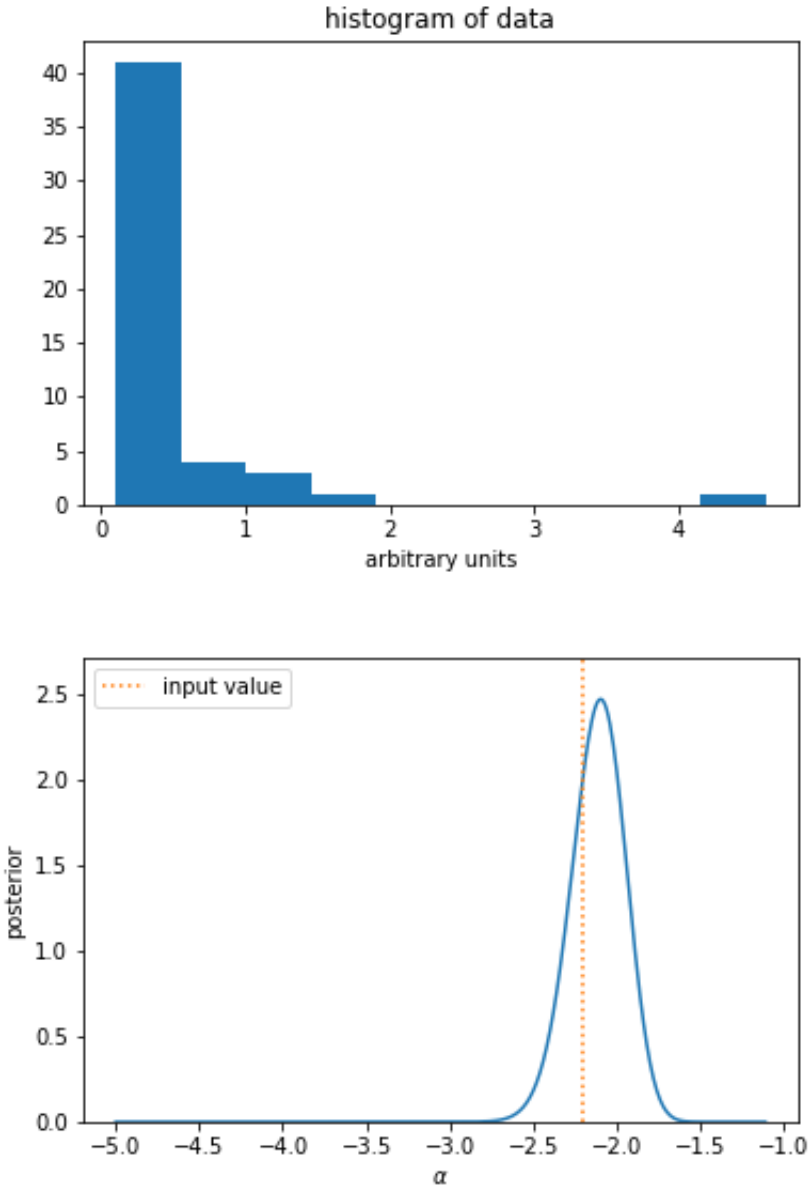

Figure 10: Bayesian constraint on the slope of the luminiosity function or SED. A histogram of the luminosity function is given above. Below is the posterior for the slope α. There are 50 data points that were generated from a power-law with $\alpha = -2.2$.

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

$$f(l) = \frac{(1+\alpha)}{\left[l_{max}^{\alpha+1} - l_{min}^{\alpha+1}\right]}\ l^{\alpha} \tag{5.110}$$

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.

5.9.1 no noise

Lets 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

$$\mathcal{L}(\{m\}|\alpha) = \prod_i f(m_i) = \frac{(1+\alpha)^N}{\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right]^N}\ \prod_i^N m_i^{\alpha} = \frac{(1+\alpha)^N}{\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right]^N}\ \left(\prod_i^N m_i\right)^{\alpha} \tag{5.111}$$

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

$$\ln P(\alpha|\{m_i\}) = N\ln(1+\alpha) - N\ln\left[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}\right] + \alpha\ln\left(\prod_i^N m_i\right) - \ln\mathcal{E}(\{m_i\}) \tag{5.112}$$

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 10.

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 no 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 a incorrect constraint.

5.9.2 with noise

So far we have taken the measurements of the luminosities to be perfect. Lets 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,

$$p(m, m_o) = p(m)p(m_o|m) \tag{5.113}$$

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

$$p(m_o|m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m_o-m)^2}{2\sigma^2}} \Theta(m_{min} < m_o < m_{max}). \tag{5.114}$$

Lets 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 $f(x)$

$$p(m_o, m) = f(m)e^{-\frac{(m_o-m)^2}{2\sigma^2}} \tag{5.115}$$

$$= \exp\left[\ln(f(m)) - \frac{(m_o-m)^2}{2\sigma^2}\right] \tag{5.116}$$

$$= \exp\left[\ln(f(m_o)) + \frac{\partial \ln f}{\partial m}(m-m_o) + \frac{\partial^2 \ln f}{\partial m^2}(m-m_o)^2 - \frac{(m_o-m)^2}{2\sigma^2} + \dots\right] \tag{5.117}$$

$$= f(m_o)\exp\left[\frac{\partial \ln f}{\partial m}(m-m_o) + \frac{\partial^2 \ln f}{\partial m^2}(m-m_o)^2 - \frac{(m_o-m)^2}{2\sigma^2} + \dots\right] \tag{5.118}$$

$$= f(m_o)\exp\left[\alpha\frac{(m-m_o)}{m_o} - \beta\frac{(m-m_o)^2}{m_o^2} - \frac{(m_o-m)^2}{2\sigma^2} + \dots\right] \tag{5.119}$$

where

$$\alpha(m_o) \equiv \left.\frac{\partial \ln f}{\partial \ln m}\right|_{m=m_o} \qquad \beta(m_o) \equiv \left(\frac{\partial \ln f}{\partial \ln m} - \frac{\partial^2 \ln f}{\partial \ln m^2}\right)_{m=m_o} \tag{5.120}$$

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

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

$$\frac{\partial}{\partial m}\ln p(m|m_o) = \frac{\partial}{\partial m}\ln p(m, m_o) \tag{5.121}$$

$$\simeq \frac{\alpha}{m_o} - \frac{2\beta}{m_o^2}(m-m_o) - \frac{1}{\sigma^2}(m-m_o) \tag{5.122}$$

So the maximum posterior is

$$m \simeq m_o + \frac{\alpha\sigma^2 m_o}{(m_o^2 + 2\beta\sigma^2)} \tag{5.123}$$

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, strictly speaking, bias is not a Bayesian concept.

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

$$p(m_o) = \int_{-\infty}^{\infty} dm\ p(m, m_o) \tag{5.124}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} dm\ f(m)e^{-\frac{(m_o-m)^2}{2\sigma^2}} \tag{5.125}$$

$$= \frac{f(m_o)}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} dm\ \exp\left[\alpha\frac{(m-m_o)}{m_o} - \beta\frac{(m-m_o)^2}{m_o^2} - \frac{(m_o-m)^2}{2\sigma^2}\right] \tag{5.126}$$

$$= \frac{f(m_o)}{\sqrt{2\pi}\sigma}\exp\left[\frac{\alpha^2\sigma^2}{2(m_o^2+2\beta\sigma^2)}\right]\int_{-\infty}^{\infty} dm\ \exp\left[-\frac{m_o^2+2\beta\sigma^2}{2\sigma^2 m_o^2}\left(m - m_o - \frac{\alpha m_o\sigma^2}{m_o^2+2\beta\sigma^2}\right)^2\right] \tag{5.127}$$

$$= \frac{f(m_o)}{\sqrt{1+\frac{2\beta\sigma^2}{m_o^2}}}\exp\left[\frac{\alpha^2\sigma^2}{2(m_o^2+2\beta\sigma^2)}\right] \tag{5.128}$$

which is not the true luminosity function. This luminosity function can be used in place of the power-law luminosity function that was used in the previous section when noise in the measurements is significant. $f(m_o)$ may have some internal parameters. 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 form 0 to ∞ rather than $-\infty$ to ∞ and the result will have some erf() functions in it, but will be essentially the same.

5.10 Example: Object detection and measurement

Lets look at the problem of detecting and measuring the brightness of a faint object in a noisy image. We will use the fake data shown in figure 11. We will consider the psf (Point Spread Function) to be known. A point source will have the shape of the psf which will be represented as $m_i(x, y)$ for pixel i if the source is centered at (x, y). It will be normalized so that it is one for the pixel centered on (x, y). In this case we have generated the mock data so we know the "real" or input parameter values.

The noise is Gaussian and uncorrelated between pixels so the likelihood is

$$\mathcal{L}\left(\boldsymbol{d}|f, b, \sigma, \boldsymbol{m}(x, y)\right) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\sum_i^N (d_i - f m_i(x, y) - b)^2\right] \tag{5.129}$$

where f is the strength of the source at it peak, b is a background. We will assume that the variance of the noise σ is known.

5.10.1 detection

Detection is a kind of model selection where the models are one with a source and one without a source. To simplify things a little I will assume the background is zero for this. If there are enough "off source" pixels the background can be measured very accurately and subtracted. We will assume that this has been done. So the likelihood will be

$$\mathcal{L}\left(\boldsymbol{d}|f, b, \sigma, \boldsymbol{m}(x, y)\right) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\sum_i^N (d_i - f m_i(x, y))^2\right] \tag{5.130}$$

We now need to calculate the evidence for this model marginalizing over f. Lets look at the quantity in the exponent:

$$-\frac{1}{2\sigma^2}\sum_i (d_i - f m_i)^2 = -\frac{1}{2\sigma^2}\sum_i \left[d_i^2 - 2 m_i d_i f + m_i^2 f^2\right] = -\frac{1}{2\sigma^2}\left[|\boldsymbol{d}|^2 - 2\boldsymbol{m}\cdot\boldsymbol{d} f + |\boldsymbol{m}|^2 f^2\right] \tag{5.131}$$

This makes the evidence

$$\mathcal{E}_s(\boldsymbol{d}) = \frac{1}{(2\pi)^{N/2}\sigma^N}\int_{-\infty}^{\infty} df\ \exp\left[-\frac{1}{2\sigma^2}\left(|\boldsymbol{d}|^2 - 2\boldsymbol{m}\cdot\boldsymbol{d} f + |\boldsymbol{m}|^2 f^2\right)\right] Pr(f) \tag{5.132}$$

$$= \frac{1}{(2\pi)^{N/2}\sigma^N}\exp\left[-\frac{1}{2\sigma^2}\left(|\boldsymbol{d}|^2 - \left[\frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right]^2\right)\right]\int_{-\infty}^{\infty} df\ \exp\left[-\frac{|\boldsymbol{m}|^2}{2\sigma^2}\left(f - \frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right)^2\right] Pr(f) \tag{5.133}$$

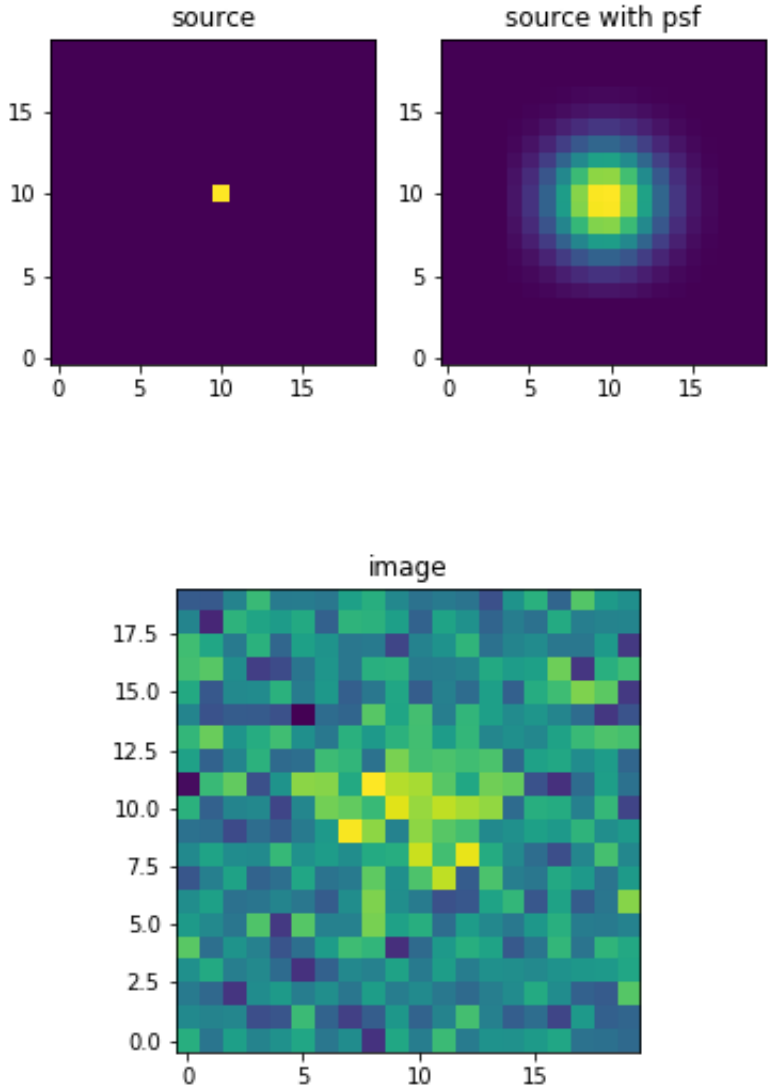

Figure 11: Simulated point source and point source convolved with psf. The source added to Gaussian uncorrelated noise.

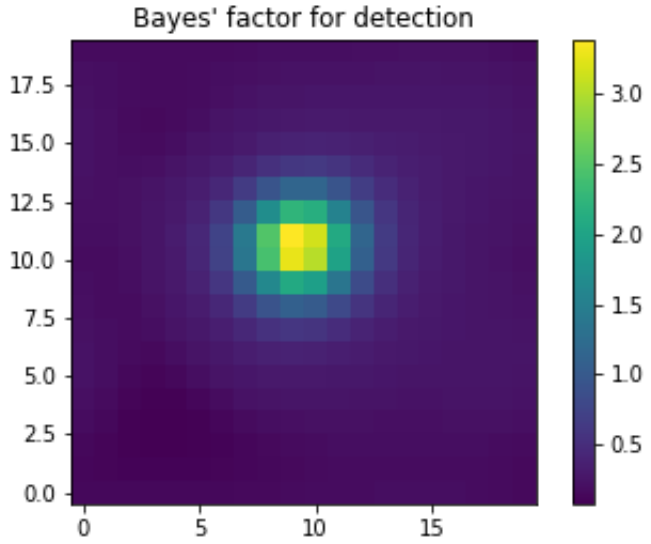

Figure 12: Bayes' factor for a source for each pixel of the image.

Now it is time to choose a prior of f, $Pr(f)$. If we were doing this for a star or galaxy survey we might use the observed luminosity function of the sources. This would be a well justified prior. The result would be the evidence that there is a star in a pixel given that the probability that a star is given by the luminosity function. In this case we would probably need to do the integral above numerically. Instead lets use a uniform prior from zero to f_{max}.

$$\mathcal{E}_s(\boldsymbol{d}) = \frac{1}{(2\pi)^{N/2}\sigma^N f_{max}} \exp\left[-\frac{1}{2\sigma^2}\left(|\boldsymbol{d}|^2 - \left[\frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right]^2\right)\right] \int_0^{f_{max}} df \;\; \exp\left[-\frac{|\boldsymbol{m}|^2}{2\sigma^2}\left(f - \frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right)^2\right] \tag{5.134}$$

$$= \frac{\left(2\pi\sigma^2\right)^{(1-N)/2}}{|\boldsymbol{m}| f_{max}} \exp\left[-\frac{1}{2\sigma^2}\left(|\boldsymbol{d}|^2 - \left[\frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right]^2\right)\right] \left[\mathrm{erf}\left(\frac{\boldsymbol{m}\cdot\boldsymbol{d}}{\sqrt{2}\sigma|\boldsymbol{m}|}\right) + \mathrm{erf}\left(\frac{|\boldsymbol{m}|}{\sqrt{2}\sigma}\left[f_{max} - \frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right]\right)\right] \tag{5.135}$$

What should we set f_{max} to? Well lets say that if the peak of the source were $f = 4\sigma$ the source would be obvious so we could say that we already know it is less than this so we set $f_{max} = 4\sigma$. I don't like this argument much myself, but lets go with it.

The source free model has no free parameters to marginalize. It is just the likelihood (5.130) with $f = 0$. Bayes' ratio is

$$B(x,y) = \frac{\sqrt{2\pi}\sigma}{|\boldsymbol{m}| f_{max}} \exp\left[\frac{1}{2\sigma^2}\left[\frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right]^2\right] \left[\mathrm{erf}\left(\frac{\boldsymbol{m}\cdot\boldsymbol{d}}{\sqrt{2}\sigma|\boldsymbol{m}|}\right) + \mathrm{erf}\left(\frac{|\boldsymbol{m}|}{\sqrt{2}\sigma}\left[f_{max} - \frac{\boldsymbol{m}\cdot\boldsymbol{d}}{|\boldsymbol{m}|^2}\right]\right)\right] \tag{5.136}$$

Using this we can scan through (x, y) in $m_i(x, y)$ to see if there is evidence that a source exists at each pixel compared to no source. Figure 12 shows this. As you can see there is clear evidence for

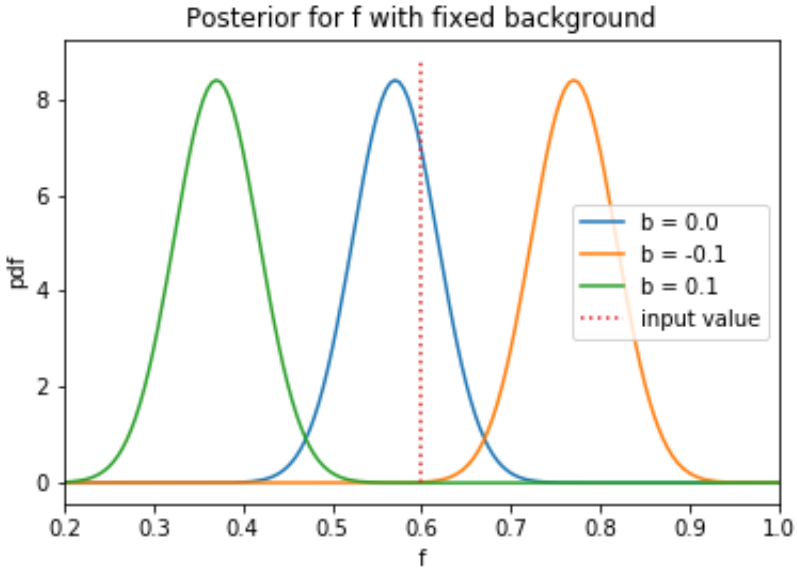

Figure 13: The posterior for f with b set to several values. $b = 0$ was the input value.

a source. The actual numerical value of the peak is dependent on our choice of f_{max} so I don't give it to much weight, but clearly some pixels are much more likely to be the position of a star then others. The pixel with the most evidence will be used as the center of our model source for the rest of this example.

This is an example of a more general technique called **match filtering**. The signal is known to have a certain shape or profile, in this case the psf. A *template* or family of templates are used to "match" the signal. The signal one is looking for is then more specific and spread out over more or area so that the sensitivity is increased and chance of a false detection is reduced over using just the value of one pixel at a time as we did in section 5.7.2.

5.10.2 measuring the background and brightness

Lets put the background and source parameters back in. If we take the priors on these parameters to be uniform the posterior is proportional to

$$P\left(f, b|\boldsymbol{d}, \sigma, \boldsymbol{m}(x,y)\right) \propto \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\sum_i^N (d_i - f m_i(x,y) - b)^2\right]. \tag{5.137}$$

The normalization can be found by summing over pixels in f-b space. Lets first fix the background and find the posterior for f for a given guess for b. Figure 13 shows this. You can see something that is generally true about the Bayesian parameter estimation. (As said before.) *You will always get a result, and it will often look quite nice, even if your model is completely wrong.* We will see in a moment that some of these values for the background are ruled out, but you wouldn't know it from figure 13. If some of the parameters are fixed the posterior for the remaining parameters can be wildly wrong. Remember that the Bayesian method always compares models. If all the models

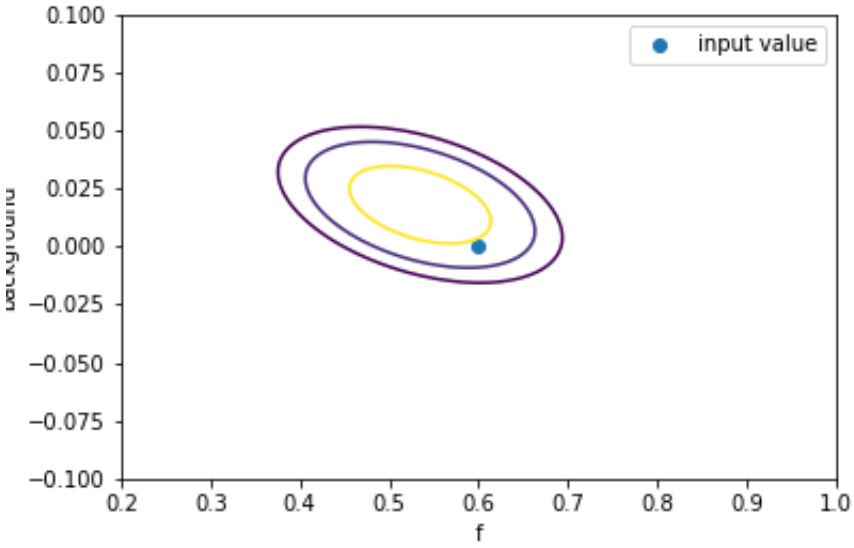

Figure 14: Posterior for the combination of parameters f and b. The contours contain 68%, 95% and 99% of the probability.

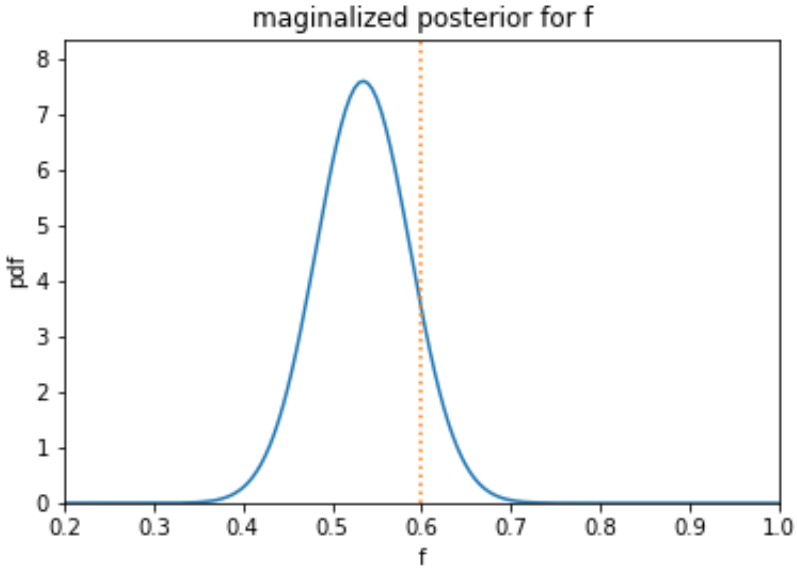

Figure 15: Marginalized posterior for parameter f. The dotted curve is the value used to generate the data.

that are used are bad models the Bayesian method will pick out the model that is the best among the bad choices, but it wont tell you that none of them are any good.

Now we can plot a contour plot of the posterior (5.137) as a function of the two parameters f and b. This is in figure 14. The normalization is found by summing over all pixels. The range of the plot must be expanded until it is clear that the posterior is very close to zero on all the borders, otherwise the normalization will not be correct. This might take a little experimentation. Also the resolution needs to be good enough that the integral is well approximated by the sum. If evaluations of the likelihood are expensive, and the resolution limited, it might be helpful to use a more sophisticated integration algorithm. In this case we can easily make the resolution very high. You can see in this plot that some of the values used for b in the previous plot are strongly inconsistent with the data.

At what level should the contours be drawn? It is traditional to draw the contours so that they *enclose* 68%, 95% and 99% of the total probability. This is in conformity with the one, two and three σ enclosed probabilities for a Gaussian distribution as discussed before. These do not correspond to the same levels of posterior probability in different each case, so the levels need to be found each time by iteratively summing the all pixels above a value and comparing it to the target confidence level then adjusting it until the proper level is found. Here is a Python function that does this. It takes a map of the posterior of any shape and finds the level by bisection.

```
import numpy as np

########################################
#  This function finds the level for a contour that contains
#  a fixed fraction of the total sum of pixels (or voxels)
########################################
def find_level(posterior,fraction) :
    tot = np.sum(posterior)
    max = np.max(posterior)
    min = np.min(posterior)

    ## initialize level to halfway between max and min
    level = 0.5*(max + min)
    ## initialize fraction for this level
    frac = np.sum( posterior[ posterior >= level ]  )/tot
    ## initialize resolution = +/- smallest pixel as fraction of total
    res = np.min( posterior[ posterior >= level ]  )/tot

    ## iterate until frac is within res of the input fraction
    while( abs(frac - fraction) > res  ) :

        ## update max or min
        if( frac > fraction) :
            min = level
        else :
            max = level

        ## update level by bisecting
        level = 0.5*(max + min)
```

```
        ## update frac and res
        frac = np.sum( posterior[ posterior >= level ] )/tot
        res = np.min( posterior[ posterior >= level ]  )/tot

    ## output the level and its actual fraction
    return level,frac
```

Now we are not much interested in the background parameter b so we would like to marginalize over it. In this case we could calculate the marginal posterior for f analytically, but we can also do it easily numbericaly. If we just sum the posterior map along the dimensions that we want to marginalize we will have it. (Again, make sure this is an accurate enough approximation of the integral in any particular case.) The marginal posterior for f is shown in figure 15 along with the input value. You can see that see that $f = 0$ is strongly disfavored by the data.

Problem 9. *Repeat the calculations of section 5.9.2 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 10. *How would you include in the posterior found in section 5.9.1 a constraint on the normalization of the luminosity function?*

6 Linear Models, Curve Fitting, least-squares and Regression

Here we will took 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 quit general solutions can be derived.

A linear model for the data point, d_i, is of the form

$$d_i = \sum_i M_{i\alpha}\theta_\alpha + n_i \qquad \text{or} \qquad \boldsymbol{d} = \boldsymbol{M\theta} + \boldsymbol{n} \tag{6.1}$$

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 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 incite 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

$$y = \sum_{\alpha=0}^{M} \theta_\alpha f_\alpha(\boldsymbol{x}) \tag{6.2}$$

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 measure value y_i. The prediction of the model can be written

$$y_i = M_{i\alpha}\theta_\alpha \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{M\theta} \tag{6.3}$$

where the matrix $\boldsymbol{M}$ contains the values of the functions $f_\alpha(\boldsymbol{x})$ at each point $\boldsymbol{x}_i$

$$\boldsymbol{M} = \begin{pmatrix} f_1(\boldsymbol{x}_1) & f_2(\boldsymbol{x}_1) & \cdots \\ f_1(\boldsymbol{x}_2) & f_2(\boldsymbol{x}_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{6.4}$$

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

$$\ln \mathcal{L} = -\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\theta})^T \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\theta}) + \ln\left((2\pi)^N |\boldsymbol{C}|\right)\right] \tag{6.5}$$

If we take uniform priors on the parameters then the posterior will be proportional to the likelihood.

Lets 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

$$\begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial \theta_\alpha} &= -\frac{1}{2}\frac{\partial}{\partial \theta_\alpha}\left[(y_i - M_{i\beta}\theta_\beta)C^{-1}_{ij}(y_j - M_{j\gamma}\theta_\gamma) + \ln\left((2\pi)^N|\boldsymbol{C}|\right)\right] && (6.6)\\ &= \frac{1}{2}\left[M_{i\alpha}C^{-1}_{ij}(y_j - M_{j\gamma}\theta_\gamma) + (y_i - M_{i\beta}\theta_\beta)C^{-1}_{ij}M_{j\alpha}\right] && (6.7)\\ &= M_{i\alpha}C^{-1}_{ij}y_j - M_{i\alpha}C^{-1}_{ij}M_{j\gamma}\theta_\gamma \qquad \boldsymbol{C} \text{ is symmetric} && (6.8)\\ &= \boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} - \boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\boldsymbol{\theta} && (6.9) \end{aligned}$$

Setting this to zero gives the MLE

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{6.10}$$

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.

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. As we said the posterior is proportional to the likelihood. If we look at (6.5) 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. So **the posterior for the linear model parameters is Gaussian**. We can find the inverse of the covariance by taking derivatives of the log-likelihood

$$-\frac{1}{2}A_{\alpha\beta} = \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \tag{6.11}$$

This is easily done with equation (6.8). The posterior is then

$$p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) = \frac{\sqrt{|\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}|}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right] \tag{6.12}$$

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

$$y = \sum_\alpha \theta_\alpha f_\alpha(x,z,w,\dots) \tag{6.13}$$

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 all linear then this is fitting a hyperplane.

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

$$y = \sum_\alpha \theta_\alpha f_\alpha(\boldsymbol{x}) \quad \text{and} \quad z = \sum_\alpha \theta_\alpha g_\alpha(\boldsymbol{x}) \tag{6.14}$$

we can rearrange this into a matrix form

$$\begin{pmatrix} y_1 \\ z_1 \\ y_2 \\ z_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{x}_1) & f_2(\boldsymbol{x}_1) & \cdots \\ g_1(\boldsymbol{x}_1) & g_2(\boldsymbol{x}_1) & \cdots \\ f_1(\boldsymbol{x}_2) & f_2(\boldsymbol{x}_2) & \cdots \\ g_1(\boldsymbol{x}_2) & g_2(\boldsymbol{x}_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \end{pmatrix} \tag{6.15}$$

which is the same form as before ($\boldsymbol{y}' = \boldsymbol{M}'\boldsymbol{\theta}$) so it can be solved in the same way.

6.2 fitting a line

Lets 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

$$y = \theta_o + \theta_1 x \tag{6.16}$$

Translating this into the matrix form gives

$$\boldsymbol{M} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{pmatrix} \tag{6.17}$$

The inverse of the covariance matrix for uncorrelated errors with equal variances is

$$\boldsymbol{C}^{-1} = \frac{\boldsymbol{I}}{\sigma^2} \tag{6.18}$$

where $\boldsymbol{I}$ is the identity matrix.

The inverse covariance for the parameters is

$$\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M} = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 1 & 1 & \dots \\ x_1 & x_2 & x_3 & \dots \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{pmatrix} \tag{6.19}$$

$$= \frac{1}{\sigma^2}\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \tag{6.20}$$

$$= \frac{N}{\sigma^2}\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \tag{6.21}$$

We can easily invert this matrix to find the parameter covariance

$$\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1} = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \tag{6.22}$$

and the MLE (6.10) is

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{6.23}$$

$$= \frac{1}{N(\overline{x^2} - \overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 & \dots \\ x_1 & x_2 & x_3 & \dots \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{pmatrix} \tag{6.24}$$

$$= \frac{1}{N(\overline{x^2} - \overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} \tag{6.25}$$

$$= \frac{1}{(\overline{x^2} - \overline{x}^2)}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} \tag{6.26}$$

$$= \frac{1}{(\overline{x^2} - \overline{x}^2)}\begin{pmatrix} \overline{x^2}\overline{y} - \overline{x}\ \overline{xy} \\ \overline{xy} - \overline{x}\ \overline{y} \end{pmatrix} \tag{6.27}$$

And (6.12) 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

$$\mathcal{L}(\boldsymbol{x}^o, \boldsymbol{y}^o|\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{(2\pi\sigma_x^2\sigma_y^2)^N} \exp\left[-\frac{1}{2}\sum_i \frac{(y_i^o - \theta_0 - \theta_1 x_i)^2}{\sigma_y^2}\right] \exp\left[-\frac{1}{2}\sum_i \frac{(x_i^o - x_i)^2}{\sigma_x^2}\right] \tag{6.28}$$

$$= \prod_i \mathcal{G}\left(y_i^o \left| \theta_0 + \theta_1 x_i, \sigma_y^2\right.\right) \mathcal{G}\left(x_i^o \left| x_i, \sigma_x^2\right.\right) \tag{6.29}$$

$$= \prod_i \mathcal{G}\left(\theta_1 x_i \left| y_i^o - \theta_0, \sigma_y^2\right.\right) \mathcal{G}\left(x_i \left| x_i^o, \sigma_x^2\right.\right) \tag{6.30}$$

$$= \prod_i \frac{1}{\theta_1} \mathcal{G}\left(x_i \left| \frac{y_i^o - \theta_0}{\theta_1}, \frac{\sigma_y^2}{\theta_1^2}\right.\right) \mathcal{G}\left(x_i \left| x_i^o, \sigma_x^2\right.\right) \tag{6.31}$$

We can use the rule for combining multivariant Gaussians that was introduced in section 3.14.3 to rearrange this

$$\mathcal{L}(\boldsymbol{x}^o, \boldsymbol{y}^o|\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{\theta_1^N} \prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left| x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \mathcal{G}\left(x_i \left| \mu_c, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \tag{6.32}$$

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

$$P(\boldsymbol{\theta}|\boldsymbol{x}^o, \boldsymbol{y}^o) = \int d^n x \; P(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{x}^o, \boldsymbol{y}^o) \tag{6.33}$$

$$= \mathcal{C} \int d^n x \; \mathcal{L}(\boldsymbol{x}^o, \boldsymbol{y}^o|\boldsymbol{\theta}, \boldsymbol{x}) \tag{6.34}$$

$$= \frac{\mathcal{C}}{\theta_1^N} \prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left| x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1}\right.\right) \int dx_i \mathcal{G}\left(x_i \left| \mu_c, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \tag{6.35}$$

$$= \frac{\mathcal{C}}{\theta_1^N} \prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left| x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \qquad \boldsymbol{G} \text{ is normalized} \tag{6.36}$$

$$= \frac{\mathcal{C}}{\left(2\pi(\sigma_y^2 + \theta_1^2\sigma_x^2)\right)^{N/2}} \exp\left[-\frac{1}{2}\sum_i \frac{(y_i^o - \theta_0 - \theta_1 x_i^o)^2}{(\sigma_y^2 + \theta_1^2\sigma_x^2)}\right] \tag{6.37}$$

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

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_0} = \sum_i \frac{(y_i^o - \hat{\theta}_0 - \hat{\theta}_1 x_i^o)}{(\sigma_y^2 + \hat{\theta}_1^2\sigma_x^2)} = 0 \qquad \Rightarrow \qquad \overline{y} - \hat{\theta}_0 - \hat{\theta}_1\overline{x} = 0 \tag{6.38}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_1} = -\frac{N\hat{\theta}_1\sigma_x^2}{(\sigma_y^2 + \hat{\theta}_1^2\sigma_x^2)} + \sum_i \frac{(y_i^o - \hat{\theta}_0 - \hat{\theta}_1 x_i^o)x_i^o}{(\sigma_y^2 + \hat{\theta}_1^2\sigma_x^2)} = 0 \qquad \Rightarrow \quad \overline{yx} - \hat{\theta}_0\overline{y} - \hat{\theta}_1\overline{x^2} - \hat{\theta}_1\sigma_x^2 = 0 \tag{6.39}$$

Solving these equations gives

$$\hat{\theta}_0 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{(\sigma_x^2 + \overline{x^2} - \overline{x}^2)} \tag{6.40}$$

$$\hat{\theta}_1 = \frac{\overline{y}\ \overline{x^2} + \overline{y}\ \sigma_x^2 - \overline{xy}\ \overline{x}}{(\sigma_x^2 + \overline{x^2} - \overline{x}^2)} \tag{6.41}$$

You can see that if $\sigma_x^2 = 0$ the former solution (6.27) is recovered.

6.4 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. 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 works minimize

$$M_{SE} = |\boldsymbol{y} - \boldsymbol{M\theta}|\,|_2^2 \equiv \sum_i \left(y_i - \sum_\alpha M_{i\alpha}\theta_\alpha \right)^2 \tag{6.42}$$

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 out 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

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{y} \tag{6.43}$$

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 often called the **pseudoinverse** or **Moore-Penrose inverse** of the matrix $\boldsymbol{M}$ (replacing the transposes with the Hermitian transpose for complex matrices).

6.4.1 calculating the pseudoinverse

The pseudoinverse is usually found by single-valued decomposition or SVD (see appendix). The SVD decomposition of $\boldsymbol{M}$ is $\boldsymbol{M} = \boldsymbol{SVD}^T$, where $\boldsymbol{V}$ is a diagonal matrix, but it is not square. The number of columns 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

$$\boldsymbol{M}^+ = \boldsymbol{DV}^+\boldsymbol{S}^T \tag{6.44}$$

where $\boldsymbol{V}^+$ is found by taking the reciprocal of the nonzero entries and then taking the transpose, i.e. $V_{ij}^+ = 1/V_{ij}$ for $V_{ij} \neq 0$. The SVD decomposition can be calculated quickly and accurately.

6.5 supervised learning and linear models

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 θ_α's. The problem of finding these 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 sets 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 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 M_{SE} (6.42) 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 fit the data. In supervised learning this is known as the **bias–variance tradeoff**.

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 $\hat{f}_{\{x\}}(x)$ is the model trained or fit to the data set $\{x\}$ that predicts y given x. The average MSE for a point in the data set is

$$\begin{aligned}
\left\langle (y - \hat{f}_{\{x\}}(x))^2 \right\rangle &= \left\langle \left(f(x) + n_y - \hat{f}_{\{x\}}(x) \right)^2 \right\rangle && (6.45)\\
&= \left\langle f(x)^2 + n_y^2 + \hat{f}_{\{x\}}(x)^2 - 2f(x)\hat{f}_{\{x\}}(x) \right\rangle \qquad \langle n_y \rangle = 0 && (6.46)\\
&= f(x)^2 + \sigma_y^2 + \left\langle \hat{f}_{\{x\}}(x)^2 \right\rangle - 2f(x)\left\langle \hat{f}_{\{x\}}(x) \right\rangle && (6.47)\\
&= f(x)^2 + \sigma_y^2 + Var[\hat{f}_{\{x\}}(x)] + \left\langle \hat{f}_{\{x\}}(x) \right\rangle^2 - 2f(x)\left\langle \hat{f}_{\{x\}}(x) \right\rangle && (6.48)\\
&= \sigma_y^2 + Var[\hat{f}_{\{x\}}(x)] + \left(f(x) - \left\langle \hat{f}_{\{x\}}(x) \right\rangle \right)^2 && (6.49)\\
&= \sigma_y^2 + Var[\hat{f}_{\{x\}}(x)] + \text{Bias}\left[\hat{f}_{\{x\}}(x) \right]^2 && (6.50)
\end{aligned}$$

When the model is simple, like a line, the variance in the model prediction $Var[\hat{f}_{\{x\}}(x)]$ will be small. (The variance is over all possible training sets.) The bias will be large for the simple model 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 have a minimum somewhere between too simple and too complex. This decomposition and general behavior is valid linear and nonlinear models. σ_y^2 is an unvoidable lower bound on the MSE because of the random noise.

A common, practical way to address this is to split the data in two and then use one part of the data to fit, or *train*, the model by finding its LSQ solution (or other model) and then use the other part to *validate* the model by calculating M_{SE} on this part. The number of parameters can be increased until the MSE reaches a minimum and starts to increasing due to over fitting. A variation on this is k-fold **cross-validation** where the data is divided into k subsets. One of the subsets is used as a validation set and the remainder as a training set. The MSE is calculated using the validation

set. Then the training is repeated with a different subset until all k subsets are used as validation sets. Finally the model parameters for each training and the MSEs are averaged. This is gives a better estimate of the expected error given a model and uses the training data more efficiently. The model is then made more complicated until as minimum in the MSE is found.

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 xs. 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. The linear model is 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

$$||\boldsymbol{y} - \boldsymbol{M\theta}|\,|_1 =\equiv \sum_i |y_i - \sum_\alpha M_{i\alpha}\theta_\alpha| \tag{6.51}$$

which is less sensitive to outliers.

6.6 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 are 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. 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.

This is related to the overfitting problem. 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 many fields the prior is called the **loss function** or **cost function**.

With a Gaussian likelihood and prior the posterior will take the form

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{M\theta})^T \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{M\theta}) + \ln\left((2\pi)^N |\boldsymbol{C}|\right) - \lambda \boldsymbol{\theta}^T\boldsymbol{\theta}\right] \tag{6.52}$$

where λ is a free parameter that regulates the strength of the prior. And the maximum posterior solution will be

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M} + \lambda \boldsymbol{I}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{6.53}$$

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 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 Laplacian or exponential prior using ℓ_1 prior

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{M\theta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{M\theta}) + \ln\left((2\pi)^N|\boldsymbol{C}|\right) - \lambda\,||\boldsymbol{\theta}||_1\right] \tag{6.54}$$

There is no close form solution for the maximum, 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 the Gaussian. In this way it does a kind of automatic model selection by identifying which parameters can be discarded.

6.7 resampling

Although resampling techniques are not specific to linear models, 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 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.

6.7.1 bootstrap

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 pdf

$$f^{bs}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i) \tag{6.55}$$

This can be considered the maximum likelihood estimate for the pdf of the data itself. It is an estimate for the real 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 it will asymptotically converge to the real distribution. This pdf will only ever return values that were observed.

Lets consider any statistic that is a function of these data point $t(x_1, x_2, \dots x_n)$. Assuming that each data point is statistically independent, the expectation value of this statistic will be

$$E[t(x_1,x_2,\dots x_n)] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n\; p(x_1)\dots p(x_n)t(x_1,x_2,\dots x_n). \tag{6.56}$$

Using the bootstrap estimation of the pdf (6.55) gives

$$E^{bs}[t(x_1,x_2,\dots x_n)] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n\; f^{bs}(x_1)\dots f^{bs}(x_n)t(x_1,x_2,\dots x_n) \tag{6.57}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n} t\left(x_{i_1},x_{i_2},\dots x_{i_n}\right) \tag{6.58}$$

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. Lets consider the arithmetic mean, $\overline{x} = \frac{1}{n}\sum_i x_i$

$$E^{bs}[\overline{x}] = \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n}\left(x_{i_1} + x_{i_2} + \ldots x_{i_n}\right) \tag{6.59}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n} x_{i_1} \qquad \text{all terms are the same} \tag{6.60}$$

$$= \frac{n^{n-1}}{n^n}\sum_{i_1=1}^{n} x_{i_1} \tag{6.61}$$

$$= \overline{x} \tag{6.62}$$

So the bootstrap mean for the sample mean is the same as the sample mean. Okay, now lets look at the bootstrap variance for the error.

$$Var^{bs}[\overline{x}] = E^{bs}[\overline{x}^2] - \overline{x}^2 \tag{6.63}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n^2}\left(x_{i_1} + x_{i_2} + \ldots x_{i_n}\right)^2 - \overline{x}^2 \tag{6.64}$$

$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n^2}\left(x_{i_1}^2 + x_{i_2}^2 + \cdots + x_{i_1}x_{i_2} + x_{i_1}x_{i_3}\ldots\right) - \overline{x}^2 \tag{6.65}$$

$$= \frac{1}{n^n}\left[\frac{1}{n^2}\left(n^{n-1}\sum_{i_1=1}^{n}x_{i_1}^2 + n^{n-1}\sum_{i_2=1}^{n}x_{i_2}^2 + \cdots + n^{n-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}x_{i_1}x_{i_2} + n^{n-2}\sum_{i_1=1}^{n}\sum_{i_3=1}^{n}x_{i_1}x_{i_3}\ldots\right)\right] - \overline{x}^2 \tag{6.66}$$

$$= \left(\frac{1}{n^2}\sum_{i_1=1}^{n}x_{i_1}^2 + \frac{n(n-1)}{n^2}\overline{x}^2\right) - \overline{x}^2 \tag{6.67}$$

$$= \frac{1}{n}\left(\frac{1}{n}\sum_{i_1=1}^{n}x_{i_1}^2 - \overline{x}^2\right) \tag{6.68}$$

This is a asymptotically unbiased estimator of the variance. The ensemble average is

$$E\left[Var^{bs}[\overline{x}]\right] = \left(\frac{n-1}{n}\right)\frac{\sigma^2}{n} \tag{6.69}$$

while the variance is of course σ^2/n.

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 LSE for a parameter is a statistic (Its a function of the data.) and we could estimate its expectation and variance in this way. In practice the sums in (6.58) are difficult to do explicitly because it as many term. There are

$$\binom{2n-1}{n} \tag{6.70}$$

distinct bootstrap samples which is 92378 for $n = 10$ and $\simeq 4.53 \times 10^{58}$ for $n = 100$. 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.

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.

Problem 11. *Show that*

$$\binom{2n-1}{n} \tag{6.71}$$

is the number of distinct bootstrap samples. (Hint: see problem 5.)

6.7.2 jackknife

Another related technique is called the jackknife resampling. Again consider a a statistic $t(x)$ that is being calculated from n data points. Lets 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

$$t_n = t_\infty + \frac{t_b}{n} + \mathcal{O}(n^{-2}) \tag{6.72}$$

Applying this to the $n-1$ case gives

$$t_{n-1} = t_\infty + \frac{t_b}{n-1} + \mathcal{O}(n^{-2}) \tag{6.73}$$

Combining these we can eliminate the lowest order bias and solve for the asymptotic limit

$$t_\infty = nt_n + (1-n)t_{n-1} + \mathcal{O}(n^{-2}) \tag{6.74}$$

$$= t_n + (n-1)\left(t_n - t_{n-1}\right) + \mathcal{O}(n^{-2}) \tag{6.75}$$

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

$$\bar{t}_{n-1}^{J} \equiv \frac{1}{n}\sum_{i=1}^{n} t_{n-1}^{(i)} \tag{6.76}$$

Using this we get the jackknife estimate of the statistic t

$$t_n^J = nt_n + (1-n)\bar{t}_{n-1}^{J} \tag{6.77}$$

We can also calculate the jackknife estimate of the variance variance for our statistic

$$Var^J[t_n] = \frac{n-1}{n}\sum_{i=1}^{n}\left(t_{n-1}^{(i)} - \bar{t}_{n-1}^{J}\right)^2 \tag{6.78}$$

This derivation is a bit dodgy actually because the expansion (6.72) is not unique. It can be shown in general that

$$\left\langle \sum_{i=1}^{n} \left(t_{n-1}^{(i)} - \bar{t}_{n-1}^{J} \right)^2 \right\rangle \leq Var[t_{n-1}] \tag{6.79}$$

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 12. *If* $t_n = \frac{1}{n}\sum_i x_i = \overline{x}$ *show that* $t_n^J = \overline{x}$ *and*

$$Var^J[t_n] = \frac{s^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_i (x_i - \overline{x}^2) \tag{6.80}$$

i.e. an unbiased estimate of the variance.

Problem 13. *If* $t_n = \frac{1}{n}\sum_i (x_i - \overline{x})^2$*, which is biased, show that* $t_n^J = \frac{1}{(n-1)}\sum_i (x_i - \overline{x})^2$ *which is not.*

7 Hypothesis testing & frequentist parameter fitting

Hypothesis testing is the frequentist version of Bayesian model selection. In some cases it is easier to apply hypothesis testing and there are many specific hypothesis tests that are commonly used in practice so are essential to understand.

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 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 any hypothesis test are as follows:

1. State the hypothesis as a well posed true or false question. The goal is the falsify this question.
2. Choose or invent a statistic that should be affected by the truth of the hypothesis.
3. Determine by analytic or numerical methods the probability distribution of the statistic.
4. Calculate the statistic with the data and determine if the measured value is improbable if the hypothesis is true.
5. If the statistic is in a region that 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 there are 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.

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 hypothesis is often called a **null hypothesis** and denoted H_o. The "null" being interpreted as referring to the rejection process.

Errors in hypothesis testing are often 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 (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 is called the **p-value** also known as the **significance** of the test. The smaller it is the more evidence you have against the null hypothesis.

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* 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. The statistic is tailored to test one aspect of the data's distribution and there as other tests to address other aspects.

7.1 frequentist test for constancy of a signal

A simple case will make this clearer. We (again) have n independent data points, x_i.

Lets start with a the null hypothesis "The signal is constant and it's mean is equal to μ." Where μ is some fixed value that is not derived from the data, maybe zero. We might add to that the implicitly assumption that the noise is Gaussian also. The measurement error will be known. The likelihood with this hypothesis is

$$p(x_i|\mu) = \frac{1}{(2\pi))^{n/2}\sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \tag{7.1}$$

Lets use the statistic

$$X^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} \tag{7.2}$$

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 ν is ruled out at the $1 - F_{\chi^2_n}(X^2)$ confidence level.

Lets relax this hypothesis and just test "The signal is constant." We might use a statistic that is very similar

$$X^2 = \sum_{i=1}^{n} \frac{(x_i-\overline{x})^2}{\sigma^2} \tag{7.3}$$

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

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \overline{x} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_1 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix} \tag{7.4}$$

The vector on the left is Gaussian distributed with n degrees of freedom. The first vector on the right is to projection of the data vector onto the $(1, 1, \dots)$ 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 $\overline{x}$). The remaining vector will be in an $n-1$ dimensional space. The magnitude of this vector, X^2 in (7.3), will thus be χ^2_{n-1} distributed. Note that these two vectors are orthogonal so there will be no cross-terms or cross-correlation between the components.

Using this to we can apply a χ^2 **test** to see if the hypothesis that the data is constant by calculating X^2 and seeing if its χ^2_{n-1} p-value is small. In this case a one-sided test is advisable because if there is some variation in the data we would expect X^2 to be large. If X^2 is exceptionally small according to the χ^2_{n-1}-distribution, we have probably overestimated the error bars.

7.2 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

$$Z = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7.5}$$

The null hypothesis is that the means of the populations are equal $\mu_1 = \mu_2$. In this case Z is normally distributed. 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 order, 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

$$T = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{7.6}$$

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

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \tag{7.7}$$

degrees of freedom. This is a **student's t test** for the difference of two means.

7.3 the variance of two populations are the same

We might also wonder if two populations have the same variance. We can do this with the statistic

$$f = \frac{S_1^2}{S_2^2} \tag{7.8}$$

This is of the form

$$\frac{X_\alpha^2/\alpha}{X_\beta^2/\beta} \tag{7.9}$$

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

$$p_F(f) = \frac{\alpha^{\alpha/2}\beta^{\beta/2} f^{\alpha/2-1}}{B(\alpha/2,\beta/2)(\alpha f+\beta)^{(\alpha+\beta)/2}} \tag{7.10}$$

where $B(\alpha,\beta)$ is the beta function. Thus this is called the **F-test** for the difference of two variances.

7.4 hypothesis testing with linear models

We found in section 6.1 that the likelihood for a linear model with Gaussian errors is

$$\ln \mathcal{L} = -\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{M\theta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{M\theta}) + \ln\left((2\pi)^N|\boldsymbol{C}|\right)\right] \tag{7.11}$$

$$= -\frac{1}{2}\left[X^2(\boldsymbol{y},\boldsymbol{\theta}) + \ln\left((2\pi)^N|\boldsymbol{C}|\right)\right] \tag{7.12}$$

where this will define $X^2(\boldsymbol{y},\boldsymbol{\theta})$ and we found that the MLE is

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y}. \tag{7.13}$$

Unlike Bayesian parameter estimation, frequentist null hypothesis testing gives us a way of assessing how well the model fits the data. The null hypothesis will be "The MLE model is the correct model. Any disagreement between the data and the model is caused by the Gaussian noise." In section 7.1 we discussed a test for constancy of the signal which is clearly a special case of the current topic with the simplest model ($y_i = \theta_0$). Just like in that case we construct a statistic by plugging the MLE model into $X^2(\boldsymbol{y},\boldsymbol{\theta})$,

$$X^2(\boldsymbol{y},\hat{\boldsymbol{\theta}}) = \left(\boldsymbol{y}-\boldsymbol{M}\hat{\boldsymbol{\theta}}\right)^T\boldsymbol{C}^{-1}\left(\boldsymbol{y}-\boldsymbol{M}\hat{\boldsymbol{\theta}}\right) \tag{7.14}$$

$$= \left(\boldsymbol{y}-\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y}\right)^T\boldsymbol{C}^{-1}\left(\boldsymbol{y}-\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{y}\right) \tag{7.15}$$

$$= \boldsymbol{y}^T\left(\boldsymbol{I}-\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\right)^T\boldsymbol{C}^{-1}\left(\boldsymbol{I}-\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\right)\boldsymbol{y} \tag{7.16}$$

$$= \boldsymbol{y}^T\left(\boldsymbol{C}^{-1}-\boldsymbol{C}^{-1}\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}\right)\boldsymbol{y} \tag{7.17}$$

$$= \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} \tag{7.18}$$

Problem 14. *Prove line (7.17) from (7.16).*

This is the equivalent of (7.3) when we where testing constancy, but now the data points might be correlated through $\boldsymbol{C}^{-1}$ and the model can have more than one parameter. Just like in that case the the problem can be seen as a projection of the data onto subspaces. If the model has k parameters, it can be shown that $\boldsymbol{A}$ will have k eigenvalues that are zero (problem 15). The matrix can be decomposed as $\boldsymbol{A} = \boldsymbol{R\Lambda R}^T$. The columns of $\boldsymbol{R}$ will be the eigenvalues of $\boldsymbol{A}$. The null space of $\boldsymbol{A}$ will be spanned by the eigenvectors that have zero eigenvalues. Any component of the data that is in the null space will not be in $X^2(\boldsymbol{y},\hat{\boldsymbol{\theta}})$.

Problem 15. *Prove that the matrix $\boldsymbol{A}$ has k eigenvalues that are zero in the following steps:*

1. *Show that the inverse of the covariance matrix can be decomposed as* $\boldsymbol{C}^{-1} = \boldsymbol{R}\boldsymbol{R}^T$ *where* $\boldsymbol{R}^{-1} = \boldsymbol{R}^T$. *Then show that the* **standardized variables**

$$\boldsymbol{x} = \boldsymbol{R}^T\boldsymbol{y} \tag{7.19}$$

will be uncorrelated and have variance 1 for each component, i.e. $\langle x_i x_j \rangle = \delta_{ij}$.

2. *Show that* $\boldsymbol{B} = \boldsymbol{R}\boldsymbol{A}\boldsymbol{R}^T$ *have the property* $B^2 = B$ *and that this implies that all its eigenvalues are either 1 or zero.*

3. *Show that* $tr[\boldsymbol{B}] = n - k$ *and why this implies that there are* k *eigenvalues that are zero.*

So despite appearances $X^2(\boldsymbol{y}, \hat{\boldsymbol{\theta}})$, is a function of $n-k$ uncorrelated standard normally distributed variables. So it is χ^2_{n-k} distributed. Using this distribution we can test if the data is consistent with the best fit model.

To summerize the geometric interpritation of what is going on here

- 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 parameters space.
- The matrix $\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}$ projects the data onto the subspace that has 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 to data space.
- The matrix $\boldsymbol{I} - \boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{C}^{-1}$ 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.

7.5 the F-test

Now lets 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 X^2 to be large. If we add parameters to our model we would expect X^2 to fall rapidly if it is fitting the data better. If we already had a model that fit the data as well as the noise will allow then adding a parameter to the model will make X^2 drop less. How much of a change would we expect if the more complex model is not justified.

Lets say we have a linear model with k parameters and another model with $m = k-r$ parameters. We make the X^2 statistics for each model and relate them with

$$X^2_M = X^2_K + \left(X^2_M - X^2_K\right) \tag{7.20}$$

$$= X^2_K + \Delta X^2 \tag{7.21}$$

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 $\Delta X^2 \sim \chi^2_r$. 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_K 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^2_K. So ΔX^2 and X^2_K are in orthogonal spaces and ΔX^2 is in a r dimensional subspace.

Since ΔX^2 and X^2_K are uncorrelated χ^2 distributed variables, their ratio

$$f = \frac{\Delta X^2}{X^2_k}\left(\frac{n-k}{r}\right) \tag{7.22}$$

is a $F_{r,n-k}$ distributed variable.

In particular if we add one parameter to the model we expect that

$$f = \frac{\Delta X^2}{X_k^2}(n-k) \tag{7.23}$$

will be $F_{1,n-k}$. It terns out that if $f \sim F_{1,n-k}$ and $f = x^2$ then x is t-distributed with $n-k$ degrees of freedom. If the measured value of f has a small chance of occurring according to this distribution we conclude that it is justified to add this parameter.

Problem 16. *Show that X_m^2 and ΔX^2 are statistically independent.*

7.6 frequentist confidence intervals

Once it has been determined that the best fit linear model has an acceptable χ^2 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 6.12). It follows that

$$X^2(\boldsymbol{x},\boldsymbol{\theta}) = X^2(\boldsymbol{x},\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \tag{7.24}$$

where the likelihood is

$$\mathcal{L} = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{C}|}} e^{-\frac{1}{2}X^2(\boldsymbol{x},\boldsymbol{\theta})} \tag{7.25}$$

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 frequentist hypothesis testing for the model was based on. The two parts of $X^2(\boldsymbol{x},\boldsymbol{\theta})$ can be shown to be statistically independent. This should not be a surprise because as we have seen the first term contains only the components of the data that do not affect the parameters and the second term contains only the parts that do affect the k parameters. They are in orthogonal subspaces. The first term in (7.24) 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.

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{M}^T\boldsymbol{C}^{-1}\boldsymbol{M}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) = \text{constant} \tag{7.26}$$

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

$$X^2 = \sum_i^n \frac{(x_i-\theta_o)^2}{\sigma^2} \tag{7.27}$$

$$= \sum_i^n \frac{(x_i-\overline{x})^2}{\sigma^2} + \frac{(\theta_o-\overline{x})^2}{\sigma^2/n} \tag{7.28}$$

The first part we saw before for testing if the signal is constant in section 7.1. The second part will be χ^2_1 distributed. $F_{\chi^2_1}(4) = 0.954$ so 95.4% confidence interval for θ_o is $\overline{x} - \frac{2\sigma}{\sqrt{n}}$ to $\overline{x} + \frac{2\sigma}{\sqrt{n}}$ or usually written $\theta_o = \overline{x} \pm \frac{2\sigma}{\sqrt{n}}$(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 if getting a sample mean that is further away than was measured is less than 5%. (Kind of a convoluted statement really).

7.6.1 Frequentest and Bayesian confidence/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 $\theta_1 = x$.

If the true parameter The frequentist confidence region is the range

7.7 Likelihood ratio test

What statistic should you use for a particular hypothesis? There are a large number of statistical tests and custom made statistics in the literature for particular purposes and we will cover a few more later. You can easily make one up yourself. The hard part is determining what the distribution of the statistic is with your hypothesis. In any but the simplest cases these are not analytically calculable. The likelihood ratio statistic has some appeal, some generality and its distribution can often be found by Monte Carlo if it is not analytic.

Consider the null hypothesis "Parameter(s) $\boldsymbol{\theta}$ has (have) the value $\boldsymbol{\theta}_H$" and the alternative is that it has some other value. The statistic is

$$L_H = \frac{\mathcal{L}\left(D|\boldsymbol{\theta}_H, \hat{\boldsymbol{\alpha}}_H\right)}{\mathcal{L}\left(D|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}\right)} \tag{7.29}$$

$\mathcal{L}\left(D|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}\right)$ is the value of the likelihood at its global maximum. $\hat{\boldsymbol{\alpha}}_H$ are the values for other parameters that maximize the likelihood when $\boldsymbol{\theta} = \boldsymbol{\theta}_H$. In general $\hat{\boldsymbol{\alpha}}_H \neq, \hat{\boldsymbol{\alpha}}$. The value of L_H is between 0 and 1, smaller being a worse fit. If one can find the statistical distribution of L_H you can reject $\boldsymbol{\theta}_H$ if L_H is too low.

This has a Bayesian feel about it because it is the ratio of probabilities, but the significance of the statistic comes from the averaging possible data sets. Note that as in Bayesian parameter fitting, this statistic will not tell you if the model doesn't this the data in a global sense. It could be the that $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}$ is a terrible fit.

7.8 Binned data χ^2 test

Lets return to the problem of determining how some measurements, such as star luminosities or photon energies, are distributed. We tackled this problem with Bayesian parameters estimation already. Now lets see some frequentist techniques.

One option is to bin the data, divide range of the data into intervals and count the number of data points in each bin. Lets 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 3.13):

$$P(\{n_i\}|N,\{p_i\}) = \frac{N!}{\prod_{i=1}^{k} n_i!}\prod_{i=1}^{k} p_i^{n_i} \tag{7.30}$$

Under the assumption that the number of counts in each bin is large we can use Sterling's approximation to simplify the probability

$$\ln P(\{n_i\}|N,\{p_i\}) = \ln(N!) + \sum_i^k \left[n_i \ln(p_i) - \ln(n_i!)\right] \tag{7.31}$$

$$\simeq N\ln(N) - N + \sum_i^k \left[n_i \ln(p_i) - n_i \ln(n_i) + n_i\right] \tag{7.32}$$

$$= N\ln(N) + \sum_i^k \left[n_i \ln(p_i) - n_i \ln(n_i)\right] \qquad \sum_i n_i = N \tag{7.33}$$

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

$$E[n_i] = Np_i \qquad Var[n_i] = Np_i(1-p_i) \tag{7.34}$$

We can expand the log probability around the average number counts using

$$\ln P(n_i = Np_i) = N\ln(N) + \sum_i^k \left[Np_i \ln(p_i) - Np_i \ln(Np_i)\right] \tag{7.35}$$

$$= 0 \qquad \sum_i p_i = 1 \tag{7.36}$$

$$\left[\frac{\partial}{\partial n_i}\ln P(\{n_i\})\right]_{n_i=Np_i} = \left[\ln(p_i) - \ln(n_i)\right]_{n_i=Np_i} = -\ln N \tag{7.37}$$

$$\left[\frac{\partial^2}{\partial n_i^2}\ln P(\{n_i\})\right]_{n_i=Np_i} = \left[-\frac{1}{n_i}\right]_{n_i=Np_i} = -\frac{1}{Np_i} \tag{7.38}$$

So the expansion is

$$\ln P(\{n_i\}) \simeq -\ln N \sum_i (n_i - Np_i) - \frac{1}{2Np_i}(n_i - Np_i)^2 + \mathcal{O}\left[(n_i - Np_i)^3\right] \tag{7.39}$$

$$= -\frac{1}{2Np_i}(n_i - Np_i)^2 + \mathcal{O}\left[(n_i - Np_i)^3\right] \tag{7.40}$$

So we can approximate the distribution of each n_i as being Gaussian so

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \tag{7.41}$$

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. If the approximation might not be valid you could also use this statistic, but find its distribution using Monte Carlo as will be discussed in the next chapter.

7.9 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is another frequentist test that is used to test if the data came form a particular distribution or whether two data sets came from the same distribution. It does not require binning the data which is is advantage over the test just discussed

Consider the sample cumulative distribution

$$F_n(x) = \frac{k}{n} \tag{7.42}$$

where k is the number of measured values below x. This will be a step function. In the limit of $n \to \infty$ we would expect $F_n(x)$ be the true cumulative distribution, $F(x)$. The KS statistic is

$$D_n = \max|F_n(x) - F(x)| \tag{7.43}$$

i.e. the largest vertical distance between the sample and cumulative distributions. Kolmogorov and Smirnov found that the distribution of this statistic is

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq t) = H(t) = 1 - 2\sum i = 1^{\infty}(-1)^{i-1}e^{-2i^2t^2} \tag{7.44}$$

for large n. This is independent of the distribution that is being tested $F(x)$.

7.9.1 two sample KS test

The hypothesis here is that both data sets come from the same data set. Lets say the sample cumulative distributions for these two samples are $F_n(x)$ and $G_m(x)$. The statistic is the maximum vertical distance between the two sample cumulative distributions

$$D_{mn} = \max|F_n(x) - G_m(x)| \tag{7.45}$$

This statistic is distributed like

$$\lim_{n\to\infty} P\left(\sqrt{\frac{mn}{m+n}}D_{mn} \leq t\right) = H(t) \tag{7.46}$$

for large n.

7.10 rank statistics

In most of the statistics we have talked about so far in this chapter, except the KS test, we have had to assume the data was Gaussian distributed or hope that this is a good approximation. Rank statistics avoid this requirement 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. 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. 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.

7.10.1 Spearman's correlation statistic

Here we revisit the problem of determining whether two sets of data points, x_i and y_i 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. We already met **Pearson's correlation coefficient**

$$\rho_{xy} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{7.47}$$

Spearman's correlation coefficient is the same thing, but using the ranks, X_i instead of the values x_i,

$$r_s = \frac{\sum_i^n (X_i - \overline{X})(X_i - \overline{X})}{\sqrt{\sum_i^n (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}} \tag{7.48}$$

This can be simplified by taking into account that the mean and variance of the ranks is always that same. Using these well know sums

$$\overline{X} = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} i = \frac{1}{2} n(n+1) \tag{7.49}$$

and

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} i^2 = \frac{1}{6} n(n+1)(2n+1) \tag{7.50}$$

The variance of both X_i and Y_i are

$$V_X = X_Y = \frac{1}{n} \sum_i (X_i - \overline{X})^2 = \frac{(n^2-1)}{12} \tag{7.51}$$

After some algebra

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_i (X_i - Y_i)^2 \tag{7.52}$$

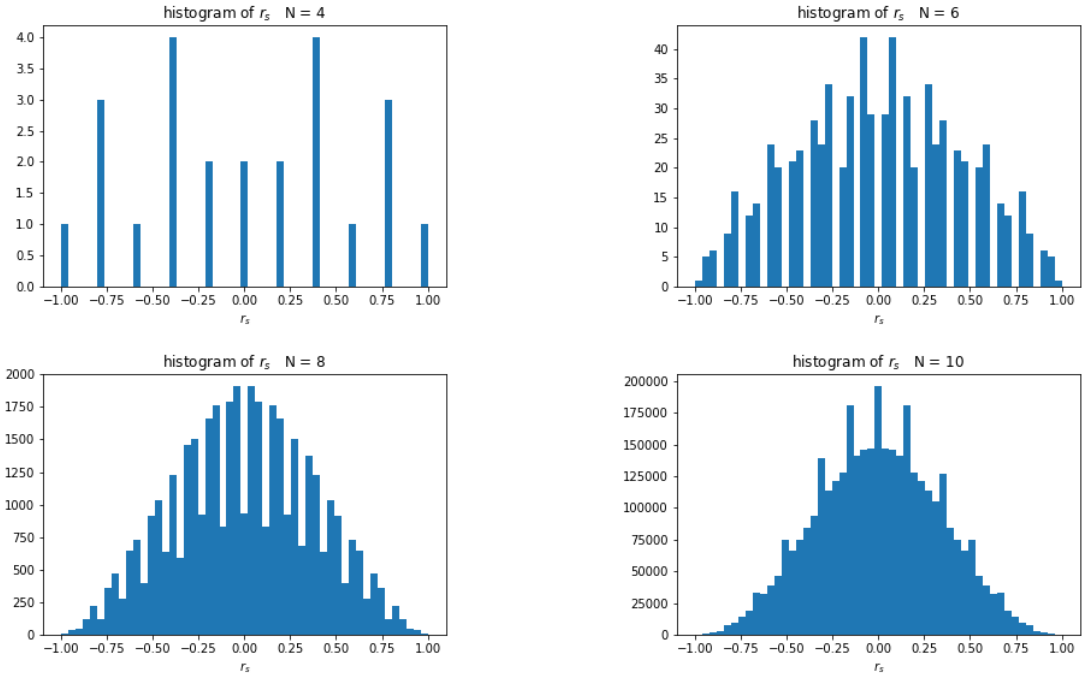

Figure 16: The distribution of Spearman's correlation coefficient, r_s, calculated under the hypothesis that there are no correlations by using all permutations of the ranks for one of the variables.

Problem 17. *Show that (7.52) 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 (7.47). 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 16 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 6.7.1 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 approximates solutions for large n.

It can be shown that

$$z = \sqrt{\frac{n-3}{1.06}}\text{arctanh}(r_s) \tag{7.53}$$

is approximately $\mathcal{N}(0,1)$ distributed. This is know as the Fisher's z-transformation. It is also true that

$$t = r_s\sqrt{\frac{n-2}{1-r^2}} \tag{7.54}$$

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.

7.10.2 Kendall's correlation coefficient

Another rank statistic that is used for detecting correlations is Kendall's τ. Lets 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 $Y_i = \{n, n-1, \ldots, 1\}$. Lets define Q as the number of pairs of Y_i's that are out of order, the number of inversions. In other words if

$$h_{ij} = \begin{cases} 1 & Y_i > Y_j \\ 0 & \text{otherwise} \end{cases} \tag{7.55}$$

then

$$Q = \sum i < j h_{ij} \tag{7.56}$$

So for $Y_i = \{1, 9, 6, 7, 5\}$ $Q = 5$. Kendall's correlation coefficient is

$$t = 1 - \frac{4Q}{n(n-1)} \tag{7.57}$$

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 17. 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

$$\sigma_t^2 = \frac{2(2n+5)}{9n(n-1)} \tag{7.58}$$

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

$$\langle r_s \rangle = \rho_s + \frac{3}{N+1}(\tau - \rho_s) \tag{7.59}$$

A disadvantage is that a particular value for τ might be hard to interpret. For a multivariant Gaussian it can be shown that

$$\tau = \frac{2}{\pi}\arcsin(\rho_{xy}) \tag{7.60}$$

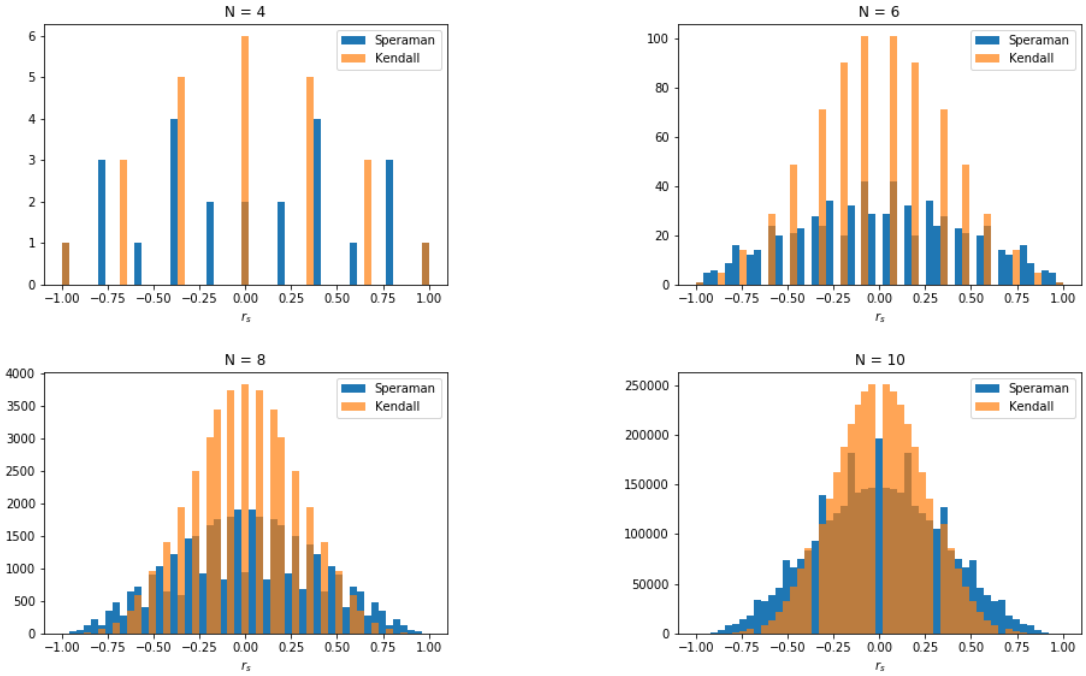

Figure 17: The distribution of Kendall and Spearman's correlation coefficients calculated under the hypothesis that there are no correlations by using all permutations of the ranks for one of the variables.

7.10.3 Wilcoxon's U test

Wilcoxon's U test (also called the **Mann-Whitney test** or **rank-sum test**))is a test for the equality of the means of two samples. In section7.2 we saw a test for the difference of the means that relied on the unterlying 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

$$W = \sum X_i \tag{7.61}$$

where where X_i is the rank for the combined sample X and Y. It is also common to use

$$U = \sum X_i - \frac{1}{2} n_x(n_x + 1) \tag{7.62}$$

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 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

$$\langle U \rangle = \frac{n_x n_y}{2} \tag{7.63}$$

$$\sigma_U^2 = \frac{1}{12} n_x n_y (n_x + n_y + 1) \tag{7.64}$$

Problem 18. *Show that U is in the range $[0, n_x n_y]$.*

7.11 sufficient 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

$$P(\boldsymbol{d}|\theta) = f(\boldsymbol{d}) g\left(\theta, t(\boldsymbol{d})\right) \tag{7.65}$$

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.

7.12 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 contain 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, lets 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 inginuity 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 $\boldsymbol{y}$ obseravbles. But the data has noise in it so if f is not linear and the noise, $\boldsymbol{n}$ is additive

$$\left\langle \tilde{\boldsymbol{\theta}} \right\rangle = \left\langle f^{-1}(\boldsymbol{d}) \right\rangle = \left\langle f^{-1}(\boldsymbol{y} + \boldsymbol{n}) \right\rangle \neq \left\langle f^{-1}(\boldsymbol{y}) \right\rangle \tag{7.66}$$

even if $\langle \boldsymbol{n} \rangle = 0$. The estimator $\tilde{\boldsymbol{\theta}}$ is biased.

A simple example, lets say we want to measure the kth power of y. The estimator $\tilde{\theta}_k = d^k$ would have an average of

$$\left\langle \tilde{\theta}_k \right\rangle = \left\langle (y+n)^k \right\rangle = \sum_{i=0}^{n} \binom{n}{i} y^k \langle n \rangle^{k-i} \tag{7.67}$$

For $k > 2$ this would be a rather bad estimator.

8 Maximum Likelihood, Fisher Information, Error Forecasting and Experimental Design

8.1 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

$$\frac{\partial \mathcal{L}}{\partial \theta_i}(\hat{\boldsymbol{\theta}}) = 0 \tag{8.1}$$

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 on 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

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{\theta}_o) + \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}_o)}{\partial \theta_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)_i + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)_i \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta}_o)}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)_j + \dots \tag{8.2}$$

$$= \ln \mathcal{L}(\boldsymbol{\theta}_o) + (\boldsymbol{\theta} - \boldsymbol{\theta}_o) \cdot \boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}_o) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)^T \boldsymbol{F}(\boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o) + \dots \tag{8.3}$$

where the curvature or Hessian matrix is

$$F_{ij}(\boldsymbol{\theta}) = \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \tag{8.4}$$

If we take the gradient of the Taylor expansion with respect to $\boldsymbol{\theta}$ we get

$$\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}_o) + \boldsymbol{F}(\boldsymbol{\theta}_o)(\boldsymbol{\theta} - \boldsymbol{\theta}_o). \tag{8.5}$$

We want to find the maximum where $\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}) = 0$. This expansion will not be perfect in general, but assuming that it is and iterating. 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

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \boldsymbol{F}^{-1}(\boldsymbol{\theta}_n) \boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}_n) \tag{8.6}$$

This is repeated until $\boldsymbol{\nabla} \ln \mathcal{L}(\boldsymbol{\theta}) = 0$ to within some tolerance. This finds the maximum quickly in most 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 be difficult in some cases. It is also possible that there are multiple maxima and this will converge on a local maximum and not the global one.

8.2 Fisher information and the minimum variance limit

Lets derive an important limitation on all estimators. The normalization of the likelihood is of course

$$\int d^n x \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = 1 \tag{8.7}$$

Taking the derivative of this with respect to a parameter θ_i gives

$$\int d^n x \frac{\partial}{\partial\theta_i}\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \int d^n x \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\frac{\partial \ln\mathcal{L}}{\partial\theta_i} \tag{8.8}$$

$$= \left\langle \frac{\partial \ln\mathcal{L}}{\partial\theta_i} \right\rangle = 0 \tag{8.9}$$

since $\langle \dots \rangle = \int d^n x \mathcal{L}(\boldsymbol{x})(\dots)$. Differentiating this again gives

$$\int d^n x \left(\frac{\partial \ln\mathcal{L}}{\partial\theta_i}\frac{\partial\mathcal{L}}{\partial\theta_j} + \frac{\partial^2 \ln\mathcal{L}}{\partial\theta_i\partial\theta_j}\mathcal{L} \right) = \int d^n x \left(\frac{\partial \ln\mathcal{L}}{\partial\theta_i}\frac{\partial \ln\mathcal{L}}{\partial\theta_j}\mathcal{L} + \frac{\partial^2 \ln\mathcal{L}}{\partial\theta_i\partial\theta_j}\mathcal{L} \right) = 0 \tag{8.10}$$

In other words

$$\mathcal{F}_{ij} \equiv \left\langle \frac{\partial \ln\mathcal{L}}{\partial\theta_i}\frac{\partial \ln\mathcal{L}}{\partial\theta_j} \right\rangle = -\left\langle \frac{\partial^2 \ln\mathcal{L}}{\partial\theta_i\partial\theta_j} \right\rangle \tag{8.11}$$

where $\mathcal{F}_{ij}$ is known as the **Fisher information matrix** (not to be confused the the Hessian $\boldsymbol{F}$ above which is not averaged) .

Say we have an estimator for the parameter θ_i which we will call $\tilde{\theta}_i(\boldsymbol{x})$

$$\int d^n x \tilde{\theta}_i(\boldsymbol{x})\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \theta_i + b(\boldsymbol{\theta}) \tag{8.12}$$

where $b(\boldsymbol{\theta})$ is the bias which could be zero or not. Taking the differential of this with respect to θ_i gives

$$\int d^n x \tilde{\theta}_i(\boldsymbol{x})\frac{\partial\mathcal{L}}{\partial\theta_i} = 1 + \frac{\partial b}{\partial\theta_i} \tag{8.13}$$

$$\int d^n x \tilde{\theta}_i(\boldsymbol{x})\mathcal{L}\frac{\partial \ln\mathcal{L}}{\partial\theta_i} = 1 + b' \tag{8.14}$$

$$\left\langle \tilde{\theta}_i(\boldsymbol{x})\frac{\partial \ln\mathcal{L}}{\partial\theta_i} \right\rangle = 1 + b' \tag{8.15}$$

It follows from (8.9) that

$$\left\langle (\tilde{\theta}_i(\boldsymbol{x}) - \theta_i)\frac{\partial \ln\mathcal{L}}{\partial\theta_i} \right\rangle = 1 + b' \tag{8.16}$$

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

$$\left[\left\langle (\tilde{\theta}_i(\boldsymbol{x}) - \theta_i)\frac{\partial \ln\mathcal{L}}{\partial\theta_i} \right\rangle \right]^2 \le Var[\tilde{\theta}_i] Var\left[\frac{\partial \ln\mathcal{L}}{\partial\theta_i} \right] = Var[\tilde{\theta}_i]\mathcal{F}_{ii} \tag{8.17}$$

or

$$Var[\tilde{\theta}_i] \ge \frac{(1+b')^2}{\mathcal{F}_{ii}} \tag{8.18}$$

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 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 the $\mathcal{F}_{ii}$ is average of the curvature or Hessian matrix of the log-likelihood. If it is evaluated at the maximum of the 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.

Generally, for any finite amount of data the MLE is not necessarily unbiased or an EE. However it can be shown that in the limit of a very large amount of data the MLE becomes an unbiased EE. I will not prove this because it is rather lengthy. This gives the MLE a special status, although in practical situations it is not always the best choice. It could be highly biased and/or there could exist an unbiased estimator with a smaller variance.

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 $\boldsymbol{\theta'}$,

$$\mathcal{F}'_{ab} = \frac{\partial \theta_i}{\partial \theta'_a} \mathcal{F}_{ij} \frac{\partial \theta_j}{\partial \theta'_b} \tag{8.19}$$

Problem 19. *Prove that the sample mean is an efficient estimator of the mean of N uncorrelated Gaussian variables.*

Problem 20. *What is the efficiency of the median as an estimate of the mean in the uncorrelated Gaussian case?*

Problem 21. *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 $\left\langle (A-\mu)^2 \right\rangle$? What is the efficiency of this estimator? Does this violate the minimum variance limit?*

8.3 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

$$Var[\theta] = \sigma_\theta^2 \simeq \frac{1}{\mathcal{F}_{\theta\theta}} \tag{8.20}$$

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 MVE 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 8.4 we will see that there are approximations that try to take this into account.

8.3.1 Example: Simple Cosmological Supernovae

As a simple example lets 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. 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 lets 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

$$D_L(z, H_o, \Omega_m, \Omega_\Lambda) = \frac{(1+z)c}{H_o} \int_0^z dz' \, \frac{1}{\sqrt{\Omega_m (1+z')^3 + \Omega_m - 1}} = \frac{c}{H_o} d_L(z, \Omega_m) \tag{8.21}$$

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 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

$$\begin{aligned} m &= M_o + 5 \log_{10}(D_L(z, H_o, \Omega_m)) && (8.22) \\ &= M_o + 5 \log 10 (H_o/c) + 5 \log_{10}(d_L(z, \Omega_m)) && (8.23) \end{aligned}$$

where M_o is an undetermined constant which includes the intrinsic peak luminosity. With these assumptions the likelihood will is

$$\ln \mathcal{L}(\boldsymbol{m}, \boldsymbol{z} | \Omega_m, \Omega_\Lambda) = -\frac{1}{2} \sum_i \frac{1}{\sigma_i^2} \left(m_i - M_o - 5 \log 10 (H_o/c) - 5 \log_{10} d_L(z_i, \Omega_m) \right)^2 - \frac{1}{2} \sum_i \ln(2\pi\sigma_i^2) \tag{8.24}$$

$$= -\frac{1}{2} \sum_i \frac{1}{\sigma_i^2} \left(m_i - \tilde{M}_o - \mu(z_i, \Omega_m) \right)^2 - \frac{1}{2} \sum_i \ln(2\pi\sigma_i^2) \tag{8.25}$$

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. 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 lets find the Fisher matrix

$$\frac{\partial}{\partial \Omega_m} \ln \mathcal{L}(\boldsymbol{m}, \boldsymbol{z} | \Omega_m) = -\sum_i \frac{1}{\sigma_i^2} \left(m_i - \tilde{M}_o + \mu(z_i, \Omega_m) \right) \frac{\partial \mu(z_i)}{\partial \Omega_m} \tag{8.26}$$

To find the Fisher matrix you have the choice of taking another derivative or squaring this. I'll choose to take another derivative

$$\frac{\partial^2}{\partial\Omega_m^2}\ln\mathcal{L}(\boldsymbol{m},\boldsymbol{z}|\Omega_m) = -\sum_i \frac{1}{\sigma_i^2}\left[\left(\frac{\partial\mu(z_i)}{\partial\Omega_m}\right)^2 + \Big(m_i - \tilde{M}_o - \mu(z_i,\Omega_m)\Big)\frac{\partial^2\mu(z_i)}{\partial\Omega_m^2}\right] \tag{8.27}$$

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

$$\mathcal{F}_{\Omega_m\Omega_m} = -\left\langle\frac{\partial^2}{\partial\Omega_m^2}\ln\mathcal{L}(\boldsymbol{m},\boldsymbol{z}|\Omega_m)\right\rangle \tag{8.28}$$

$$= \sum_i \frac{1}{\sigma_i^2}\left(\frac{\partial\mu(z_i)}{\partial\Omega_m}\right)^2 \tag{8.29}$$

The other components of the Fisher matrix are

$$\mathcal{F}_{M_oM_o} = \sum_i \frac{1}{\sigma_i^2} \tag{8.30}$$

$$\mathcal{F}_{M_o\Omega_m} = \sum_i \frac{1}{\sigma_i^2}\frac{\partial\mu(z_i)}{\partial\Omega_m} \tag{8.31}$$

where

$$\frac{\partial\mu}{\partial\Omega_m} = 5\log_{10}(e)\frac{\partial}{\partial\Omega_m}\ln d_L(z) = -2.17147\frac{(1+z)}{2d_L(z)}\int_0^z dz'\ \frac{(1+z')^3-1}{(\Omega_m(1+z')^3+(1-\Omega_m))^{3/2}} \tag{8.32}$$

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 some thing like $f(z)\propto x^\alpha e^{-z/z_o}$. Using this we can convert the sums into integrals

$$\sum_i \to n\int dz\ f(z) \tag{8.33}$$

So that for example

$$\mathcal{F}_{\Omega_m\Omega_m} = \frac{n}{\sigma^2}\int dz\ f(z)\left(\frac{\partial\mu(z)}{\partial\Omega_m}\right)^2 \tag{8.34}$$

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 $\mathcal{F}_{\Omega_m\Omega_m} = 1.67$, $\mathcal{F}_{M_oMo} = 11.1$, $\mathcal{F}_{\Omega_m M_o} = 3.18$ 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.

8.4 The Asymptotic Normal Approximations

Lets expand the likelihood around the MLE (or MPE for a uniform prior)

$$\ln\mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) \simeq \ln\mathcal{L}(\boldsymbol{d}|\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\cdot\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\left.\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \mathcal{O}\left(|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|^3\right) \tag{8.35}$$

$$= \ln\mathcal{L}(\boldsymbol{d}|\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\left.\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \mathcal{O}\left(|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|^3\right) \tag{8.36}$$

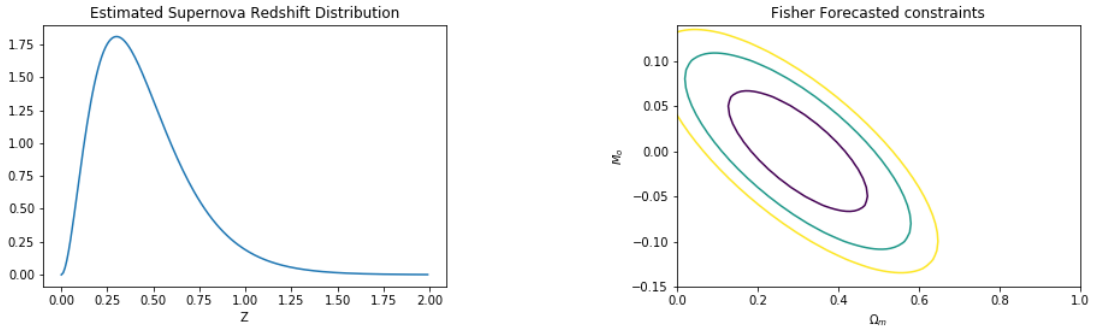

Figure 18: The estimated supernova redshift distribution on the left and the forecasted constraints on Ω_m and and the peak luminosity normalization. Here 100 supernovae are assumed and $\sigma_m = 0.3$ mag. The redshift distribution parameters are $\alpha = 2$ and $z_0 = 0.15$.

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 mount of data gets larger and the constraints on the parameters get stronger.

Ignoring the higher order terms, the average log-likelihood will be

$$\langle \ln \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) \rangle \simeq \left\langle \ln \mathcal{L}\left(\boldsymbol{d}|\hat{\boldsymbol{\theta}}\right) \right\rangle - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{8.37}$$

This leads us to approximate the posterior of a future experiment as

$$p(\boldsymbol{\theta}) \simeq \frac{|\boldsymbol{\mathcal{F}}|}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right] \tag{8.38}$$

at least near its peak. With this you see that the Cramér-Rao limit (8.20) is the *conditional* variance for one parameter given this posterior.

Following the rules for manipulating multivariant Gaussian distributions discussed in section 3.14 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

$$\sigma_\theta^2 \simeq \left[\boldsymbol{\mathcal{F}}^{-1}\right]_{\theta\theta} \tag{8.39}$$

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.

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

$$\boldsymbol{\mathcal{F}}^{tot} = \boldsymbol{\mathcal{F}} + \boldsymbol{C}_{\text{prior}}^{-1} \tag{8.40}$$

$\boldsymbol{C}_{\text{prior}}^{-1}$ could be the inverse covariance 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 the CMB observation?" 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.

Problem 22. *Show that if the likelihood depends only on a single combination of two parameters, that is*

$$\ln \mathcal{L}(\boldsymbol{x}|\theta_1, \theta_2) = \ln \mathcal{L}\left(\boldsymbol{x}|f(\theta_1, \theta_2)\right) \tag{8.41}$$

then the Fisher matrix will have a determinant of zero. 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 8.38 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 18 shows such a plot for our simplified cosmological supernova example.

Problem 23. *Show that the area or volume of an ellipsoid in n dimensional parameter space with*

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) < \chi^2 \tag{8.42}$$

is

$$V = \frac{1}{\sqrt{|\boldsymbol{\mathcal{F}}|}} \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} \frac{\chi^n}{(n-1)} \tag{8.43}$$

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 degenercies 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 strategies. 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.

8.5 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

$$\boldsymbol{\mathcal{F}}_{\alpha\beta} = \boldsymbol{\mu}_{,\alpha}^{T} \boldsymbol{C}^{-1} \boldsymbol{\mu}_{,\beta} + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\right] \tag{8.44}$$

Problem 24. *Show that equation (8.44 is correct. Use the identities*

$$\boldsymbol{C}^{-1}{}_{,\beta} = -\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1} \quad \textit{and} \quad \frac{d}{d\beta}\ln|\boldsymbol{C}| = tr\left[\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\right] \tag{8.45}$$

The solution is in the appendix.

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 (8.6) when finding the maximum likelihood because $\boldsymbol{F}$ can be computationally expensive and $\boldsymbol{\mathcal{F}}$ often works just as well.

9 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 a sufficiently large sample drawn from a distribution one can use the **law of large numbers** to estimate any expectation value

$$E[g(x)] = \int_{-\infty}^{\infty} d^n x \; p(\boldsymbol{x}) g(\boldsymbol{x}) \simeq \frac{1}{n} \sum_i g(\boldsymbol{x}_i) \tag{9.1}$$

where the $\boldsymbol{x}_i$'s are drawn from the distribution $p(\boldsymbol{x})$.

In one dimension this is often possible efficiantly sample from a standard distribution function and any good statisitic saftware package will have functions to do this. There are several methods used to find these deviates such as rejection and transformation methods.

9.0.1 transformation

We already know how to transform variable. 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 $p(x)\frac{dx}{df} = 1$ or $df = p(x)dx$. In other $\int_0^{f(x)} df = F(x)$ is the cumulative distribution function. So if you can invert the cumulative distribution function to get $x = F^{-1}(f)$ then you can draw f from a uniform distribution between 0 and 1 and the corresponding x will be distributed according to the pdf $p(x) = \frac{d}{dx}F(x)$.

For example, say you want a random point within a sphere of radius R. The pdf is

$$p(r, \theta, \phi) dr d\theta d\phi \propto r^2 d\cos(\theta) d\phi \tag{9.2}$$

The cumulative distribution for the radius is

$$F(r) = \left(\frac{r}{R}\right)^3 \tag{9.3}$$

So inverting this gives

$$r = RF^{1/3} \tag{9.4}$$

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 of a random particle within a profile with a particular

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 direction of the vector $\boldsymbol{x}$ is uniformly distributed on the 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

$$\boldsymbol{y} = \frac{r}{||\boldsymbol{x}||}\boldsymbol{x} \tag{9.5}$$

See Press et al. (2007) for more information on generating random deviates.

9.1 Monte Carlo and Confidence Intervals

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 of the sun, cross-section of scatter 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?

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 outcomes that have varying numbers of globular clusters. You observe the number of globular clusters in the galaxies around us. You derive a statistic from them – 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 for the outcomes of 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 probability or confidence level given a sample?

For frequentist this topic falls into the subject of quintile 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 quintile 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

$$P(k|p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{9.6}$$

Assuming a uniform prior and using the integral form of the beta function (appendix E) , we can

renormalize this to get the posterior for p

$$P(p|n) = \frac{\Gamma(n+2)}{\Gamma(n-k+1)\Gamma(k+1)} p^k (1-p)^{n-k} \tag{9.7}$$

$$= \frac{(n+1)!}{(n-k)!k!} p^k (1-p)^{n-k} \tag{9.8}$$

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 euql to zero)

$$p_{ML} = \frac{k}{n} \tag{9.9}$$

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

$$\langle p \rangle = \frac{(n+1)!}{(n-k)!k!} \int_0^1 dp\ p^{k+1} (1-p)^{n-k} \tag{9.10}$$

$$= \frac{(n+1)!(k+1)!}{k!(n+2)!} \tag{9.11}$$

$$= \frac{(k+1)}{(n+2)} \tag{9.12}$$

and we can find its variance

$$\left\langle p^2 \right\rangle = \frac{(n+1)!}{(n-k)!k!} \int_0^1 dp\ p^{k+2} (1-p)^{n-k} \tag{9.13}$$

$$= \frac{(k+2)(k+1)}{(n+3)(n+2)} \tag{9.14}$$

$$\sigma_p^2 = \left\langle p^2 \right\rangle - \langle p \rangle^2 \tag{9.15}$$

$$= \frac{(k+1)}{(n+2)} \left[\frac{(k+2)}{(n+3)} - \frac{(k+1)}{(n+2)} \right] \tag{9.16}$$

$$\simeq \frac{3\langle p \rangle \left(1 - \langle p \rangle\right)}{n} + \mathcal{O}(1/n^2) \tag{9.17}$$

The distribution is plotted in figure 19. 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.

9.2 Monte Carlo Integration

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 (9.1) to estimate a multidimensional integral that

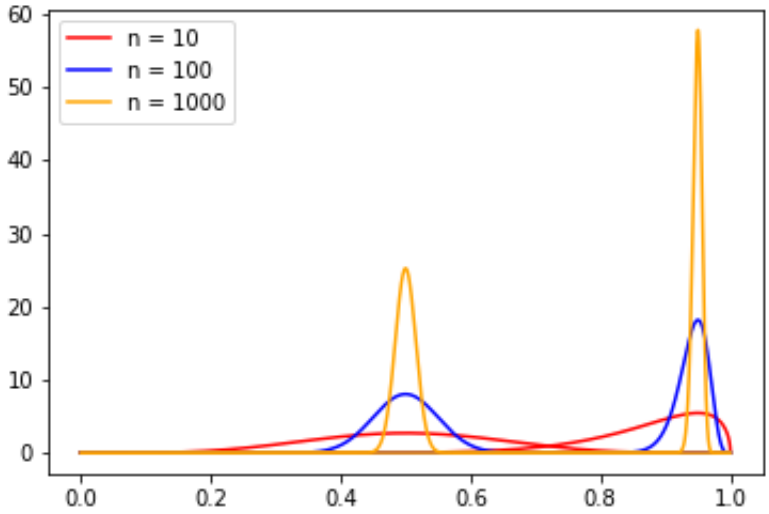

Figure 19: The posterior for the cumulative probability up to a boundary given that the fraction of samples above the boundary is 50% (center) and 95% (right). The the total number of samples is as in the legend.

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.

$$\int_{\partial V} d^n x\ g(\boldsymbol{x}) = \int_{\partial V} d^n x\ p(\boldsymbol{x}) \left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x})}\right) = \int_{-\infty}^{\infty} d^n x\ p(\boldsymbol{x}) \left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \Theta(\boldsymbol{x} \in V)) = \left\langle \left(\frac{g(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \Theta(\boldsymbol{x} \in V) \right\rangle \tag{9.18}$$

$$\simeq \frac{1}{n} \sum_{x_i \in V} \left(\frac{g(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right) \tag{9.19}$$

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

$$\pm \frac{1}{n} \sqrt{\sum_{x_i \in V} \left(\frac{g(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right)^2 - \frac{1}{n} \left(\sum_{x_i \in V} \frac{g(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\right)^2} \tag{9.20}$$

Clearly the special case of a uniform random sampling is $p(\boldsymbol{x}) = 1/V$. 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.

9.3 Markov Chains

In statistics a **chain** is an ordered series of random numbers, $x_1 \dots x_n \dots$ where the conditional probability of each element given the other elements is specified is specified – $p(x_n|x_1 \dots)$. 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 are 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 $\boldsymbol{x}_n$ depends only on the previous element 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 many 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

$$\int_{-\infty}^{\infty} dx_n \; f(\boldsymbol{x}_n)p(\boldsymbol{x}_{1+n}|\boldsymbol{x}_n) = f(\boldsymbol{x}_{1+n}) \tag{9.21}$$

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

$$E[g(\boldsymbol{x})] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} g(\boldsymbol{x}_n) \tag{9.22}$$

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 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 25. *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 $p(\boldsymbol{x}_{1+n}|\boldsymbol{x}_{n-1})$?*

9.3.1 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**:

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) \tag{9.23}$$

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 (9.21).

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
$$\alpha(\boldsymbol{x}_t, \boldsymbol{x}_n) = \min\left\{1, \frac{q(\boldsymbol{x}_n|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_n)}\frac{f(\boldsymbol{x}_t)}{f(\boldsymbol{x}_n)}\right\} \tag{9.24}$$
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

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) = q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)\alpha(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n)f(\boldsymbol{x}_n) \tag{9.25}$$

$$= \begin{cases} q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) & , \quad q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) < q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) \\ q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) & , \quad q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) > q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)f(\boldsymbol{x}_n) \end{cases} \tag{9.26}$$

and

$$p(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) = q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})\alpha(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})f(\boldsymbol{x}_{n+1}) \tag{9.27}$$

which will be the same as above in the same cases so (9.23) is satisfied.

9.3.2 choosing a proposal distribution

Although the MCMC is guaranteed to converge under the conditions mentioned above, it might take a *very log 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 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. The rejection rate is too low the chain will move, but not fast enough to get

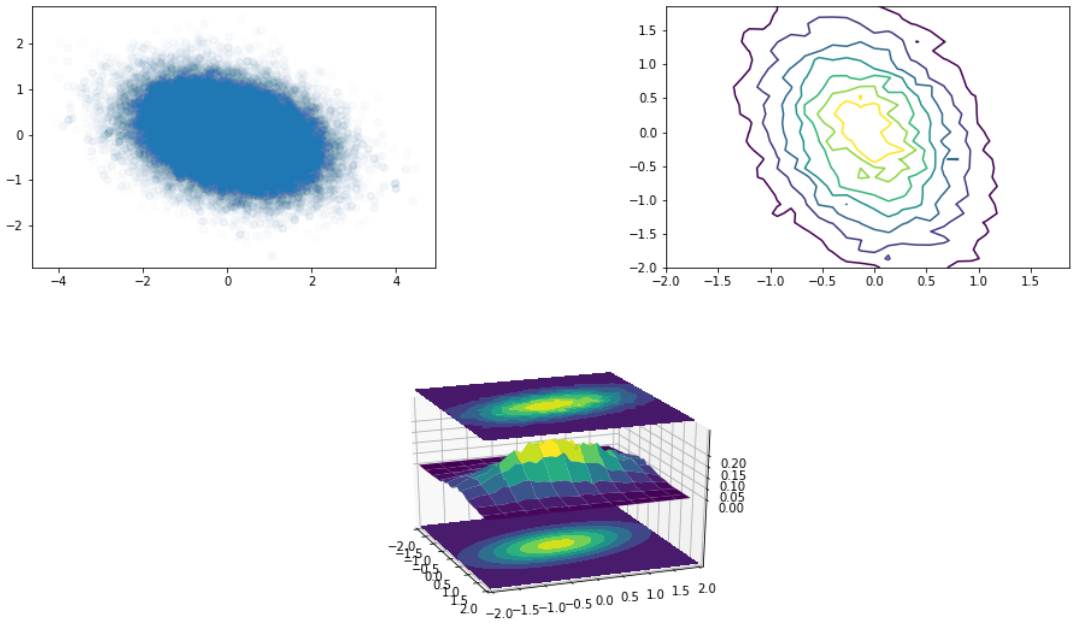

Figure 20: On the top left are the points from a Metropolis-Hastings Markov Chain for a simple 2 dimensional Gaussian. On the top right is a contour plot of the 2d histogram of those points. Below is the target distribution and two representations of the histogram.

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 $q(\boldsymbol{x}_t|\boldsymbol{x}_n)$.

The 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, 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 the settles in in some location like in figure 24). Other times the maximum of the distribution can be found by some minimization technique such as was discussed in section **??** and MCMC is being 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 to get an acceptable rejection rate during the rest of the chain.

The biggest difficulties with MCMC arise when the parameters are deg

- 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 $f(\boldsymbol{x})$ that has 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 in one point in space will not work well in 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 ti deal with. If there are multiple peaks in the distribution that are separated by regions of low probability

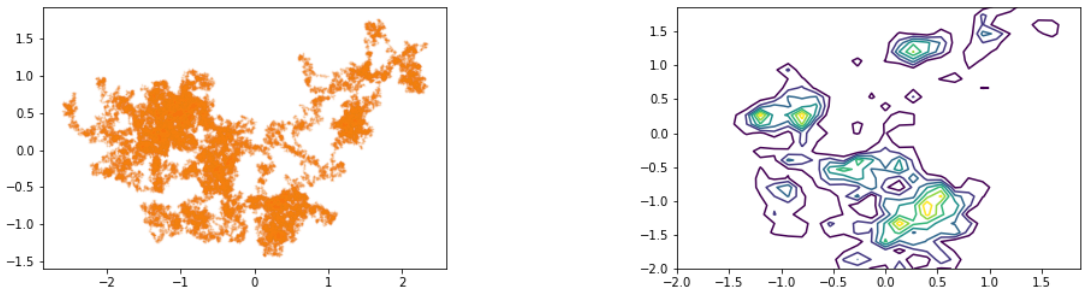

Figure 21: This is the same simple 2D Gaussian as shown in figure 20, but here the width of the proposal distribution was chosen to be too small ($\sigma = 0.01$ here and 0.5 there).

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 22). 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.

curse of large dimensions

Problem 26. *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*

$$q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n) = f(x^{(i)}_{n+1}|\boldsymbol{x}^{(i-)}_n) \tag{9.28}$$

where $\boldsymbol{x}^{(i-)}$ are all the other parameters that are not being updated this step.

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 $f(x^{(i)}_{n+1}|\boldsymbol{x}^{(i-)}_n)$. 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.

9.3.3 example

Figures 23 through 26 show an example. The (fake) data is shown in 23. The likelihood function is Gaussian with varying but known σ's. The model here is $y = 2\left((\frac{p_1}{2} + p_2)x\right)^2 + p_1$. Figure 24 shows the burn in period. You can see that the chain seems to have been attracted to some steady state solution. Figure 25 shows a chain 100,000 points long. The acceptance rate for this chain was 21.6%. An isotropic Gaussian proposal distribution was used with $\sigma = 0.3$.

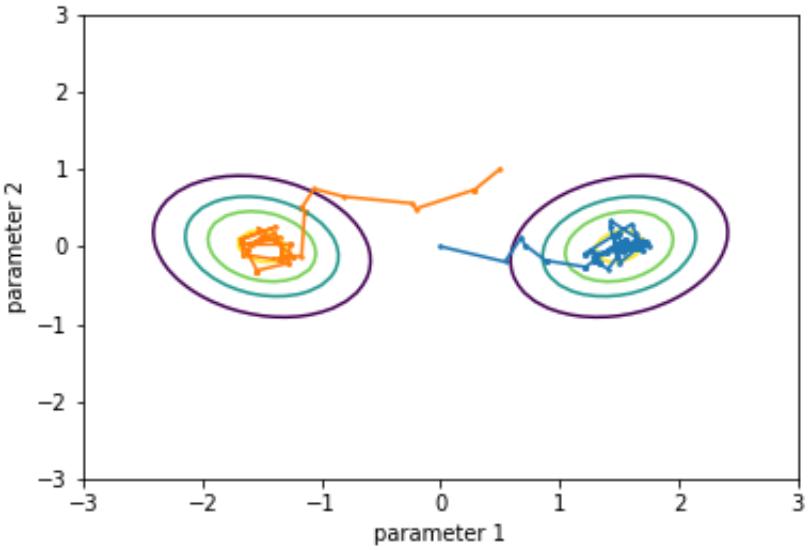

Figure 22: Two chains started at different initial points in a mulit model distribution. In this case the chance of jumping between modes within a single chain is small.

9.3.4 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*, the separation in the chain. It can be defines as

$$C_{\alpha,\beta}(m) = \frac{\sum_{i=1}^{N-m}(\alpha_i - \overline{\alpha})(\beta_{i+m} - \overline{\beta})}{\sqrt{\left(\sum_{i=1}^{N-m}(\alpha_i - \overline{\alpha})^2\right)\left(\sum_{i=m}^{N}(\beta_i - \overline{\beta})^2\right)}} \tag{9.29}$$

where α and β are parameter values. In the case of the autocorrelation $\alpha = \beta$. $C_{\alpha,\beta}(0) = 1$. 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**. Lets call this N_{corr}. Points separated be less than the correlation length will not be independent. You can define an effective number of independent samples in the chain as

$$N_{eff} = \frac{N_{chain}}{N_{corr}} \tag{9.30}$$

We want this number to be large. We also want any statistic we are interested to depend on a number of points that is much larger than N_{corr}. For example the difference between the 95% and 99% contour levels depend on only 4% of the particles. This might be smaller than N_{corr}.

Figure 27 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.

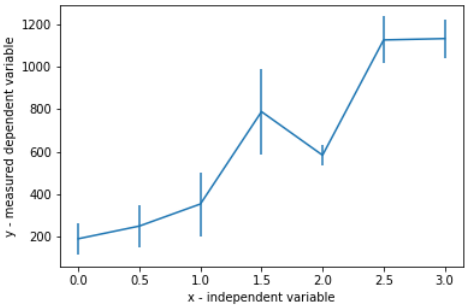

Figure 23: Simulated data.

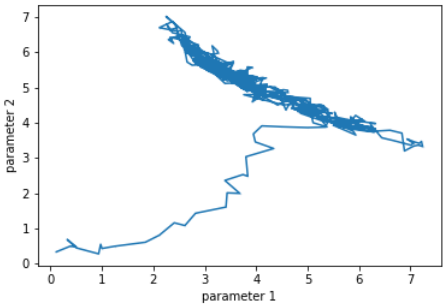

Figure 24: The first 1,000 steps of the MCMC starting at an initial guess $(0, 0)$.

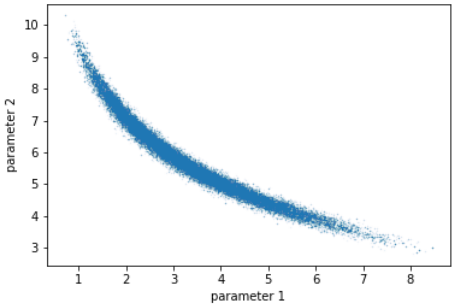

Figure 25: The 100,000 steps after discarding the first 1,000.

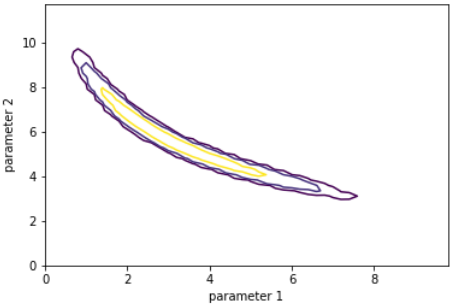

Figure 26: Contours surrounding 68%, 95% and 99% of the points.

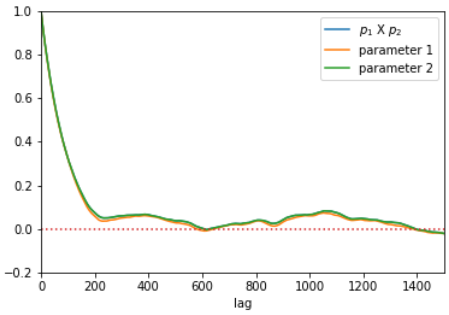

Figure 27: The correlation coefficient as a function of lag in the MCMC chain. Shown are the autocorrelation for the two parameters and the crosscorrelation between them.

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 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:

$$\overline{\theta}^\alpha = \frac{1}{n}\sum_i^n \theta_i^\alpha \qquad \overline{\overline{\theta}} = \frac{1}{m}\sum_\alpha^m \overline{\theta}^\alpha \tag{9.31}$$

$$s_\alpha^2 = \frac{1}{n-1}\sum_i^n (\theta_i^\alpha - \overline{\theta}^\alpha)^2 \qquad B = \frac{n}{m-1}\sum_\alpha^m \left(\overline{\theta}^\alpha - \overline{\overline{\theta}}\right)^2 \tag{9.32}$$

$$W = \frac{1}{m}\sum_\alpha^m s_\alpha^2 \qquad V = \frac{n-1}{n}W + \frac{M+1}{nm}B \tag{9.33}$$

$$\hat{R} = \sqrt{\frac{V}{W}} \tag{9.34}$$

$\hat{R}$ 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.

9.3.5 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 26) 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 a 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.

9.4 nested sampling & calculation of evidence

Another numerical technique that has become wide spread for solving the Bayesian infirence problem and in astrophysics in particular is **nested sampling**. The application of this to the inference problem is due to J. (2004) (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

$$\mathcal{E} = \int d^n\theta \; \mathcal{L}(\boldsymbol{d}|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}) \tag{9.35}$$

For the moment lets take the prior $Pr(\boldsymbol{\theta})$ to be uniform, but restricted to a finite volume in parameter space. Lets 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 $\mathcal{L}(\boldsymbol{\theta}_1) < \mathcal{L}(\boldsymbol{\theta}_2) < \cdots < \mathcal{L}(\boldsymbol{\theta}_2)$. We know from our study of the extreme values (section 4.5) and Monte Carlo confidence levels (section 9.1) 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

$$V_{pr}(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1)) \simeq \left(1 - \frac{1}{N_a}\right) \tag{9.36}$$

Now lets 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 point might now be $\boldsymbol{\theta}_1$ or it might not. 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 $V_{pr}^2(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1)) \simeq \left(1 - \frac{1}{N_a}\right) V_{pr}$. If we continue to do this we will in the nth cycle get an estimate for the volume of

$$V_{pr}(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1^n)) = V_{pr}^n \simeq \left(1 - \frac{1}{N_a}\right)^n \tag{9.37}$$

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

$$v_n = \frac{1}{2}\left(V_{pr}^{n-1} - V_{pr}^{n+1}\right) \tag{9.38}$$

Using this we can estimate the evidence as

$$\mathcal{E} \simeq \sum_{i=n}^{M} \mathcal{L}_n v_n \tag{9.39}$$

where M is the total number of cycles used. Typically the calculation is continued until new cycles changed $\mathcal{E}$ by the desired accuracy.

Any expectation value for any function of the parameters can then be estimated with

$$E\left[f(\boldsymbol{\theta})\right] \simeq \frac{\sum_{i=n}^{M} f(\boldsymbol{\theta}_1^n)\mathcal{L}_n v_n}{\mathcal{E}} \tag{9.40}$$

The approximation is often made that

$$V_{pr}^n = \exp\left[\ln(V_{pr}^n)\right] \tag{9.41}$$

$$= \exp\left[n \ln\left(1 - \frac{1}{N_a}\right)\right] \tag{9.42}$$

$$\simeq \exp\left[-\frac{n}{N_a}\right] \tag{9.43}$$

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 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

$$\sigma^2_{\ln \mathcal{E}} \simeq \frac{`H}{N_a} = \frac{1}{N_a} \sum_{n=1}^{M} \frac{\mathcal{L}_n v_n}{\mathcal{E}} \ln\left(\frac{\mathcal{L}_i}{\mathcal{E}}\right) \tag{9.44}$$

where H is an estimate of the *relative entropy* (more on this later).

9.4.1 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 be comes 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 $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_n$. A popular software package that does this is *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 $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_n$ while 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 point. 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 9.0.1 and then stretching it with the axis ratios of the ellipsoid.

A Selected Problem Solutions

[17] **Problem 17.**
First

$$\sum_i (X_i - Y_i)^2 = \sum_i (X_i - \overline{X} - Y_i + \overline{Y})^2 \qquad \overline{X} = \overline{Y} \tag{A.1}$$

$$= \sum_i \left[(X_i - \overline{X})^2 + (Y_i - \overline{Y})^2 - 2(X_i - \overline{X})(Y_i - \overline{Y}) \right] \tag{A.2}$$

$$= 2nV_X - 2\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) \tag{A.3}$$

Then using

$$r_s = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{nV_X} \tag{A.4}$$

if follows that

$$r_s = 1 - \frac{1}{2nV_X} \sum_i (X_i - Y_i)^2 \tag{A.5}$$

$$= 1 - \frac{6}{n(n^2-1)} \sum_i (X_i - Y_i)^2 \tag{A.6}$$

using (7.51).

[24] **Problem 24.**
The log of the likelihood is

$$\ln \mathcal{L}(\boldsymbol{d}) = -\frac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) - \frac{1}{2} \ln |\boldsymbol{C}| - \frac{n}{2} \ln [2\pi] \tag{A.7}$$

Taking the first derivative with respect to a parameter, α, gives

$$\frac{\partial}{\partial \alpha} \ln \mathcal{L}(\boldsymbol{d}) = \frac{1}{2}(\boldsymbol{\mu}_{,\alpha})^T \boldsymbol{C}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{\mu}_{,\alpha}) - \frac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}{}_{,\alpha} (\boldsymbol{d} - \boldsymbol{\mu}) \tag{A.8}$$

$$-\frac{1}{2} \frac{d}{d\alpha} \ln |\boldsymbol{C}| \tag{A.9}$$

where the subscript with commas are derivatives. Using these formulas

$$\frac{d}{d\beta} \ln |\boldsymbol{C}| = \mathrm{tr} \left[\boldsymbol{C}^{-1} \boldsymbol{C}_{,\beta} \right] \tag{A.10}$$

we can change the last term

$$\frac{\partial}{\partial \alpha} \ln \mathcal{L}(\boldsymbol{d}) = \frac{1}{2}(\boldsymbol{\mu}_{,\alpha})^T \boldsymbol{C}^{-1} (\boldsymbol{d} - \boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{\mu}_{,\alpha}) - \frac{1}{2}(\boldsymbol{d} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}{}_{,\alpha} (\boldsymbol{d} - \boldsymbol{\mu}) \tag{A.11}$$

$$-\frac{1}{2} \mathrm{tr} \left[\boldsymbol{C}^{-1} \boldsymbol{C}_{,\alpha} \right] \tag{A.12}$$

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,

$$\left\langle \frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathcal{L}(\boldsymbol{d})\right\rangle = -\frac{1}{2}(\boldsymbol{\mu}_{,\alpha})^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{,\beta}) - \frac{1}{2}(\boldsymbol{\mu}_{,\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{,\alpha}) - \frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}{}_{,\alpha\beta}(\boldsymbol{d}-\boldsymbol{\mu}) \quad \text{(A.13)}$$

$$-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}{}_{,\alpha}\boldsymbol{C}_{,\beta} + \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta\alpha}\right] \quad \text{(A.14)}$$

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

$$\mathcal{F}_{\alpha\beta} = -\left\langle \frac{\partial^2}{\partial\alpha\partial\beta}\ln\mathcal{L}(\boldsymbol{d})\right\rangle = \boldsymbol{\mu}_{,\alpha}^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_{,\beta} + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}{}_{,\alpha\beta}\boldsymbol{C}\right] + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^{-1}{}_{,\alpha}\boldsymbol{C}_{,\beta} + \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta\alpha}\right] \quad \text{(A.15)}$$

Using

$$\boldsymbol{C}^{-1}{}_{,\beta} = -\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1} \quad \text{(A.16)}$$

and the chain rule

$$\boldsymbol{C}^{-1}{}_{,\alpha\beta} = \boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1} + \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta}\boldsymbol{C}^{-1}\boldsymbol{C}_{,\alpha}\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{C}_{,\beta\alpha}\boldsymbol{C}^{-1} \quad \text{(A.17)}$$

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 (8.44).

[25] **Problem 25.**

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

$$p(\boldsymbol{x}_{n+1},\boldsymbol{x}_n,\boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n,\boldsymbol{x}_{n-1}) \quad \text{(A.18)}$$

$$= p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})p(\boldsymbol{x}_{n-1}) \quad \text{(A.19)}$$

$$= p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})p(\boldsymbol{x}_{n-1}) \quad \text{(A.20)}$$

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 $\boldsymbol{x}_n$,

$$p(\boldsymbol{x}_{n+1},\boldsymbol{x}_{n-1}) = \int_{-\infty}^{\infty} d\boldsymbol{x}_n\ p(\boldsymbol{x}_{n+1},\boldsymbol{x}_n,\boldsymbol{x}_{n-1}) \quad \text{(A.21)}$$

$$= p(\boldsymbol{x}_{n-1})\int_{-\infty}^{\infty} d\boldsymbol{x}_n\ p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \quad \text{(A.22)}$$

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

$$p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_{n-1}) = \int_{-\infty}^{\infty} d\boldsymbol{x}_n\ p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \quad \text{(A.23)}$$

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.

[26] **Problem 26.**

The acceptance probability for a Gibbs step will be

$$\alpha(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n) = \frac{q(\boldsymbol{x}_n|\boldsymbol{x}_{n+1})}{q(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)} \frac{f(\boldsymbol{x}_{n+1})}{f(\boldsymbol{x}_n)} \tag{A.24}$$

$$= \frac{f(x_n^{(i)}|\boldsymbol{x}_{n+1}^{(i-)})}{f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})} \frac{f(\boldsymbol{x}_{n+1}^{(i)}, \boldsymbol{x}_{n+1}^{(i-)})}{f(\boldsymbol{x}_n^{(i)}, \boldsymbol{x}_n^{(i-)})} \tag{A.25}$$

$$= \frac{f(x_n^{(i)}|\boldsymbol{x}_n^{(i-)})}{f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})} \frac{f(\boldsymbol{x}_{n+1}^{(i)}, \boldsymbol{x}_n^{(i-)})}{f(\boldsymbol{x}_n^{(i)}, \boldsymbol{x}_n^{(i-)})} \qquad \boldsymbol{x}_{n+1}^{(i-)} = \boldsymbol{x}_n^{(i-)} \tag{A.26}$$

$$= \frac{f(x_n^{(i)}|\boldsymbol{x}_n^{(i-)})}{f(x_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})} \frac{f(\boldsymbol{x}_n^{(i-)}) f(\boldsymbol{x}_{n+1}^{(i)}|\boldsymbol{x}_n^{(i-)})}{f(\boldsymbol{x}_n^{(i-)}) f(\boldsymbol{x}_n^{(i)}|\boldsymbol{x}_n^{(i-)})} \tag{A.27}$$

$$= 1 \tag{A.28}$$

B Matrix basics

$$(\boldsymbol{ABC}\dots)^T = \dots\boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T \tag{B.1}$$
$$(\boldsymbol{ABC}\dots)^{-1} = \dots\boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \tag{B.2}$$
$$(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T \tag{B.3}$$
$$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T \tag{B.4}$$

Some properties of the determinant

$$|\boldsymbol{A}| = \prod_i \lambda_i \quad \text{where } \lambda_i \text{ are the eigenvalues} \tag{B.5}$$
$$|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}| \tag{B.6}$$
$$|\boldsymbol{BA}| = |\boldsymbol{B}||\boldsymbol{A}| \tag{B.7}$$
$$|c\boldsymbol{A}| = c^n|\boldsymbol{A}| \tag{B.8}$$
$$|\boldsymbol{A}^T| = |\boldsymbol{A}| \tag{B.9}$$

Some properties of the trace

$$\mathrm{tr}(\boldsymbol{A}) = \sum_i A_{ii} \tag{B.10}$$
$$\mathrm{tr}(\boldsymbol{A}) = \sum_i \lambda_i \quad \text{where } \lambda_i \text{ are the eigenvalues} \tag{B.11}$$
$$\mathrm{tr}(\boldsymbol{A}^T) = \mathrm{tr}(\boldsymbol{A}) \tag{B.12}$$
$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}) \tag{B.13}$$
$$\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B}) \tag{B.14}$$

derivatives of matricies

$$\frac{d}{d\beta}\boldsymbol{C}^{-1} = -\boldsymbol{C}^{-1}\left[\frac{\partial \boldsymbol{C}}{\partial \beta}\right]\boldsymbol{C}^{-1} \tag{B.15}$$
$$\tag{B.16}$$
$$\frac{d}{d\beta}\ln|\boldsymbol{C}| = \frac{d}{d\beta}\ln\left(\prod_i \lambda_i\right) = \frac{d}{d\beta}\sum_i \ln\lambda_i = \sum_i \frac{1}{\lambda_i}\frac{d\lambda_i}{d\beta} = \mathrm{tr}\left[\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \beta}\right] \tag{B.17}$$

$\boldsymbol{A}$ is an **orthogonal matrix** if and only if

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} \tag{B.18}$$

An orthogonal matrix has the following properties

$$\boldsymbol{A}^T = \boldsymbol{A}^{-1} \tag{B.19}$$
$$|\boldsymbol{A}| = \pm 1 \tag{B.20}$$

The $|\lambda_i| = 1$ for all eigenvalues and the magnitude of all eigenvactors are 1.

$\boldsymbol{C}$ is a **positive definite matrix** if

$$\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} > 0 \quad \forall \boldsymbol{x}. \tag{B.21}$$

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.

C Matrix decompositions

Eigenvalue decomposition
If $\boldsymbol{A}$ is a $N \times N$ matrix with linear independent columns the it can be decomposed as

$$\boldsymbol{A} = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{M}^{-1} \tag{C.1}$$

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

$$\boldsymbol{A} = \boldsymbol{S}\boldsymbol{V}\boldsymbol{D}^{\dagger} \tag{C.2}$$

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 $\boldsymbol{A}^{\dagger}\boldsymbol{A}$.

D Notation

Notation may vary but in general I follow the guide in table 2

E Some useful integrals and mathematical definitions

E.1 Gaussian integrals

$$\int_{-\infty}^{\infty} dx\ e^{-\frac{x^2}{2}} = \sqrt{2\pi} \tag{E.1}$$

$$\begin{aligned}\int_{-\infty}^{\infty} dx\ e^{-(ax^2+bx+c)} &= e^{-c}\int_{-\infty}^{\infty} dx\ e^{-\left(\sqrt{a}x+\frac{b}{2\sqrt{a}}\right)^2+\frac{b^2}{4a}} = e^{-c+\frac{b^2}{4a}}\int_{-\infty}^{\infty}\frac{dy}{\sqrt{a}}\ e^{y^2} \\ &= \sqrt{\frac{\pi}{a}}e^{-c+\frac{b^2}{4a}}\end{aligned} \tag{E.2}$$

"A and B"	A, B
"A or B"	$A \cup B$
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	$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu},\boldsymbol{C}\right)$
$\boldsymbol{x}$ 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 2: notation

$$\int_0^\infty dx\; x^n e^{-\frac{1}{2}Ax^2} = 2^{\frac{n-1}{2}} A^{-\frac{n+1}{2}} \Gamma\left(\frac{n+1}{2}\right) \qquad n > -1 \tag{E.3}$$

E.2 Stirling's approximation

$$\ln N! \simeq N \ln N - N \text{ for } N \gg 1 \tag{E.4}$$

or more accurately

$$N! \simeq \sqrt{2\pi N}\left(\frac{N}{e}\right)^N \text{ for } N \gg 1 \tag{E.5}$$

E.3 The Gamma function

$$\begin{array}{lll} \int_0^\infty dx\; x^n e^{-x^2} & = \frac{1}{2}\Gamma\left(\frac{n+1}{2}\right) & \\ \int_0^\infty dx\; x^{z-1} e^{-x} & = \Gamma(z) & \\ \Gamma(n) & = (n-1)! & n = 1, 2, \ldots \\ \Gamma\left(\frac{1}{2}+n\right) & = \frac{(2n)!}{4^n n!}\sqrt{\pi} & n = 0, 1, 2, \ldots \end{array} \tag{E.6}$$

E.4 Error function

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2} du = \frac{1}{\sqrt{\pi}}\int_{-z}^z e^{-u^2} du \tag{E.7}$$

$$\frac{1}{\sqrt{2\pi}\sigma}\int_b^a dx\ e^{-\frac{x^2}{2\sigma}} = \frac{1}{2}\left[\mathrm{erf}\left(\frac{a}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right] \tag{E.8}$$

$$\mathrm{erf}(\infty) = 1 \tag{E.9}$$

$$\mathrm{erf}(-x) = -\mathrm{erf}(x) \tag{E.10}$$

E.5 Beta function

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 dx\ x^{p-1}(1-x)^{q-1} \tag{E.11}$$

There are also many integral forms of this beta function.

E.6 Miscellaneous approximations

$$\lim_{N\to\infty}\left[1 + \frac{t^2}{2N}\right]^N = e^{\frac{t^2}{2}} \tag{E.12}$$

Index

References

Feroz F., Hobson M. P., 2008, MNRAS, 384, 449

Gregory P., 2006, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press

J. S., 2004, in AIP Conf. Proc. Vol. 735 Bayesian inference and maximum entropy methods in science and engineering. Am. Inst. Phys., Melville, NY, p. 395

Jaynes E., 2003, Probability Theory - The Logic of Science

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

Silvia D., Skilling J., 2006, Data Analysis a Bayesian Tutorial, 2nd edn. Oxford University Press