An introduction to statistics for astronomers

Robert Crittenden

May 24, 2013

Abstract

Lecture notes from short course in statistics for the graduate students of the ICG. Original notes, 2004. Revised notes: March 2009, May 2011, May 2013.

1 Introduction

Physicists and astronomers, even very theoretical ones, always want to know if their theories have any relation to the real universe. At some point they must compare the predictions of their theory to the observations, and this is where an understanding of statistics becomes essential.

Given the limited time, I will not be giving rigorous derivations, but will instead focus on giving an intuitive understanding of the topics. The goal is to give a general treatment that will provide working knowledge of the subject and how it is applied.

1.1 Bayesians vs Frequentists

There are two basic approaches to statistics. The Bayesian approach described below dates back to Bayes and Laplace, but has only really become popular recently. For most of the last two centuries, statistics has been discussed in terms of 'frequentist,' or orthodox, statistics, which is distilled in the work by Fisher, Neyman and Pearson.

The key difference between these approaches is the definition of probability. For a Bayesian, it simply indicates the degree of belief, or plausibility of something being true. The frequentist approach instead defines the probability of something as the fraction of times, or long term frequency, that thing will occur given infinitely many repeated trials. In this approach, one invents *estimators* of the quantities of interest; there might be many for a given quantity which you must choose between given their sampling properties. (Estimator theory will be the topic of a later chapter.)

Most of the debate between the camps centres on the notion of priors. The frequentist approach does not have any explicit dependence on priors, and frequentists see the need for these as a weakness of the Bayesian approach. The Bayesian camp argues that in choosing one particular estimator, you are implicitly making a choice of priors, and that at least in their approach the priors are made explicit.

In effect, Bayesians and frequentists are asking different questions. A Bayesian asks, 'How likely is a given parameter value given the data?', while a frequentist

asks, 'How probable is the data, given certain parameters?' The good news is that if the data are good, it doesn't really matter, both sides should come to the same conclusions. Unfortunately in cosmology, we are usually at the limits of what the data are telling us, and in such cases it is important to clearly state what prior assumptions you are making, or your results will not be reproducible.

Here I will generally take the Bayesian viewpoint, if and when it matters. I see it as a cleaner, more transparent formalism which mirrors the scientific method as a whole.

2 Elements of probability theory

Most of statistics reduces down to describing probability distributions, or, given some theoretical assumptions or hypothesis, using some probability distribution to infer another. These probability distributions can be on a set of either discrete or continuous variables, and can be used to describe a range of things, for example:

- Something which is directly observable, e.g. the CMB temperature in a certain direction
- A statistic which summarises some observables given certain elementary assumptions, e.g. a power spectrum
- A parameter which is defined in the context of a particular model, e.g. the value of the cosmological constant
- The relative probability of a model within a larger class of models.

2.1 Probability and the probability density function

For discrete variables, we assume we have some variable X which can take any one of a set of discrete possible values, x_i. The probability that X takes a particular value x_i is given by the expression $\mathcal{P}_X(x_i)$. This function is assumed to be always positive take a value between 0 and 1.

For continuous variables, one begins with a variable X which can take a range of values, which we will assume stretches from negative infinity to positive infinity. The probability that the variable lies in the range $[x, x + dx]$ is given by

$$\mathcal{P}(x \leq X \leq x + dx) = f_X(x)dx, \tag{1}$$

where $f_X(x)$ is known as the *probability distribution function* (pdf), or probability density. Virtually any function can serve as a probability density, as long as it is positive definite and has a finite integral over the range in question. (Sometimes called the probability mass function for discrete variables.) A closely related function is the *cumulative distribution function*, defined as the probability X is less than some value x:

$$\mathcal{P}(X \leq x) = F_X(x) \equiv \int_{-\infty}^{x} f_X(x')dx'. \tag{2}$$

Equivalently, $dF_X(x) = f_X(x)dx$.

2.2 Basic rules of probability

Its essential we establish a few basic rules for dealing with probabilities and probability distributions. Luckily, we need rather few rules and the ones we need are fairly intuitive. The first, sometimes called the sum rule, states that if we add up the probabilities for all the possible outcomes of an observation, it should be unity. For a discrete example, it means

$$\mathcal{P}_X(A) + \mathcal{P}_X(A^c) = 1, \tag{3}$$

where I have denoted by A a set of possible values of X, and A^c is the complement of that set, i.e. all the other possibilities. Equivalently, one could have said

$$\sum_i \mathcal{P}_X(x_i) = 1, \tag{4}$$

where the sum goes over all the possible values of X. For the continuous case, the sum rule can be written as

$$\int_{-\infty}^{\infty} f_X(x)dx = 1, \tag{5}$$

or equivalently that $F_X(\infty) = 1$.

The second thing we need to do is to formalise the notion of *conditional probability*. Suppose that X and Y are possible events, we define the conditional probability of X given Y, $\mathcal{P}(X|Y)$, is the probability of observing X given that Y is known to be true. This should be the ratio of the probability both X and Y are true ($\mathcal{P}(X, Y)$) to the probability that Y is true:

$$\mathcal{P}(X|Y) = \mathcal{P}(X, Y)/\mathcal{P}(Y). \tag{6}$$

This relation, known as the product rule, we will take as defining conditional probability.

Two events are said to be *independent* if the likelihood of one being true does not depend on whether the other is true, $\mathcal{P}(X|Y) = \mathcal{P}(X)$. From the product rule, this implies $\mathcal{P}(X, Y) = \mathcal{P}(X)\mathcal{P}(Y)$.

2.3 Marginalisation and Bayes' theorem

Marginalisation is a generalisation of the sum rule to the case where more than one event is being considered. It basically says

$$\mathcal{P}(X) = \sum_i \mathcal{P}(X, Y_i) = \int dY f(X, Y), \tag{7}$$

i.e., if you add up all the joint probabilities of X and Y, including everything which might happen with Y, you should get the probability of X.

Another essential element of probability theory follows directly from the product rule. We can use it to relate two different conditional probabilities:

$$\mathcal{P}(X, Y) = \mathcal{P}(X|Y)\mathcal{P}(Y) = \mathcal{P}(Y|X)\mathcal{P}(X). \tag{8}$$

From this quickly follows *Bayes' theorem*,

$$\mathcal{P}(X|Y) = \mathcal{P}(Y|X)\mathcal{P}(X)/\mathcal{P}(Y), \tag{9}$$

which is the basis of much of what we will talk about below.

2.4 Bayesian reasoning

Bayes' theorem gives a basis for reasoning, and evaluating the probability of a theory given some data. More precisely, it gives us a system for updating our confidence in some theory given data, so it implicitly contains the whole of the scientific method. Take the event X above to represent the truth of some hypothetical model, H, and let event Y represent some observed data D. Then Bayes' theorem is transcribed as

$$\mathcal{P}(H|D) = \mathcal{P}(D|H)\mathcal{P}(H)/\mathcal{P}(D). \tag{10}$$

Its worth describing each of these terms a little. $\mathcal{P}(H)$ is called the *prior distribution*, the probability we described to a theory based on what we knew previous to seeing this data. It may be based on earlier data to some extent, or it may be purely based on some expectations of naturalness in the context of our theories. Bayes' theorem relates the prior to the *posterior distribution*, $\mathcal{P}(H|D)$, which is the probability we should assign to the hypothesis given what we thought before and the data we see now.

The prior and posterior distributions are related by two factors. $\mathcal{P}(D|H)$ is the *likelihood* of the hypothesis, or the probability of the data assuming the hypothesis is correct. $\mathcal{P}(D)$ is called the *Bayesian evidence*, and often it is ignored as simply a normalisation factor; however, sometimes it plays a key role, such as when we want to compare very different hypotheses. We will return to this when we come to the issue of model comparison.

More generally, we can also apply this in the context of parameters of some assumed model. Suppose we assume the dataset, $\vec{D}$ behaves as some model M_A which has the set of parameters, $\vec{\theta}_A$. We find the posterior distribution of the parameters

$$\mathcal{P}(\vec{\theta}_A|\vec{D}, M_A) = \mathcal{P}(\vec{D}|\vec{\theta}_A, M_A)\mathcal{P}(\vec{\theta}_A|M_A)/\mathcal{P}(\vec{D}|M_A). \tag{11}$$

This distribution should be normalised to unity, so that the evidence for this model

$$\mathcal{P}(\vec{D}|M_A) = \int d^n\vec{\theta}_A \mathcal{P}(\vec{D}|\vec{\theta}_A, M_A)\mathcal{P}(\vec{\theta}_A|M_A). \tag{12}$$

2.5 Updating conclusions with new data

One advantage of the Bayesian method is that it provides a simple means of incorporating previous data in the method, via the prior. Imagine we perform two experiments. In the Bayesian method, we can perform the first and calculate the posterior distribution based on it and some initial prior assumption; this posterior then becomes the prior for the second experiment. The final posterior distribution is identical to what you would have gotten analysing the two experiments together (assuming the measurements are independent and uncorrelated so that their likelihoods factorise.) That is,

$$\begin{aligned} \mathcal{P}(\theta|\mathbf{D_1}, \mathbf{D_2}, M) &\propto \mathcal{P}(\mathbf{D_1}, \mathbf{D_2}|\theta, M)\mathcal{P}(\theta|M) = \mathcal{P}(\mathbf{D_2}|\theta, M)\mathcal{P}(\mathbf{D_1}|\theta, M)\mathcal{P}(\theta|M) \\ &\propto \mathcal{P}(\mathbf{D_2}|\theta, M)\mathcal{P}(\theta|\mathbf{D_1}, M). \end{aligned} \tag{13}$$

Thus, it is straight forward to combine various independent observations in the Bayesian approach without having to consider all observations simultaneously.

2.6 Priors

The necessity of priors is usually what puts people off of the Bayesian method. These can be subjective, meaning that conclusions can vary from person to person. One should always be clear on the priors being used to insure that your results are reproducible. As discussed above, with good data the priors become irrelevant (unless they were pathological to begin with!) Unfortunately, in cosmology the data are not always conclusive, so priors cannot always be ignored. But how does one know what prior or even what parameterisation is best to use?

The priors are meant to reflect all previous information, excluding the data at hand. The first place to start is to look for other observational constraints. It is extremely rare that we are totally ignorant of some physical parameter; usually we can find some bounds, however weak they may be. Observations are rarely the first of their kind, so some bounds likely exist on the quantities at hand. In the absence of any data, theory can and must play an important role. If there is a model for how a given quantity has arisen, then it can often be used to help define the prior.

In the absence of either observations and theoretical expectations, one is forced to make some choice of prior based on the symmetries of the problem. If one is interested in what are known as *location parameters*, where the likelihood is expected to be a function of the difference between the observation and the parameter ($\mathcal{P}(D|\theta) = g(D - \theta)$), then most often a flat prior is recommended. For example, trying to determine the mean of a Gaussian, or the position of a galaxy in a field of view. If instead one is interested in a multiplicative parameter (called a *scale parameter*), where instead $\mathcal{P}(D|\sigma) = \sigma^{-1} g(D\sigma)$, it is more common to use a prior uniform in the logarithm of the parameter ($\propto 1/\sigma$, sometimes called a reciprocal prior.) These parameters are typically positive definite, such as the variance or rms of a Gaussian distribution. Remember that for these distributions, one needs to define upper and lower bounds which requires some physical input; one reasonable choice is to use the limitations of the experiment itself to set these.

There is a growing literature on finding Bayesian priors which are "objective" or "non-informative," or perhaps more accurately one should think of them as a standard which is to be used when no other prior information is forthcoming. These often should not be formally thought of as priors, as much as a means of making a reasonable posterior which is minimally influenced by the prior choice. One proposal is to use a prior, called the Jeffrey's prior, which is simply proportional to the square root of the Fisher information. This has the advantage of being invariant under reparameterisations, and though it often leads to improper prior distributions, the resulting posterior is usually proper. This approach seems to work for 1-dimensional distributions, but it is less reliable in higher dimensions. A closely related prior, called a "reference prior", attempts to maximise the impact of the data on the posterior by maximising some measure of the divergence between the prior and posterior (though there is some arbitrariness about what divergence measure to use.) For both this approach and for the Jeffrey's prior, the chosen priors depend on the model of the likelihood, so they are not priors in the usual sense.

Finally, if one really has no clue how to assign the prior, one can also appeal to information theory. Some statisticians argue the best choice is a prior which

maximises the entropy of the distribution, where the entropy is defined as

$$S = -\int dx f(x) \log\left(\frac{f(x)}{m(x)}\right), \tag{14}$$

where $m(x)$ is the measure. A good discussion of such issues can be found in the book by Jaynes.

3 Describing 1-d distribution functions

We'll begin with different ways of describing one-dimensional functions. While we will be primarily concerned with application to probability distributions, these descriptions are often used in many other contexts. I will initially focus on continuous distributions, though most of what I say will apply equally well to discrete distributions, with any integrations replaced by sums.

3.1 Expectations and moments

For any function of X, like $\theta(X)$, an *expectation value*, or mean value, for that function can be defined given a pdf:

$$\langle\theta(X)\rangle \equiv \int_{-\infty}^{\infty} f_X(x)\theta(x)dx. \tag{15}$$

One way of describing a pdf is by giving its *moments*, which are simply the expectation values of X^n:

$$\mu'_n = \langle X^n\rangle = \int_{-\infty}^{\infty} f_X(x)x^n dx \tag{16}$$

The prime denotes a straightforward algebraic moment, rather than a central moment which will be defined below. For example, we know that since $f_X(x)$ is a probability distribution function and since X must take some value in the range, it follows that $\mu'_0 = \int_{-\infty}^{\infty} f_X(x)dx = 1$ always. The mean of the distribution is first moment, $\bar{X} = \mu'_1 = \int_{-\infty}^{\infty} f_X(x)xdx$.

We are more often interested what are called *central moments*, which are the moments of the distribution if the mean of the distribution were subtracted off:

$$\mu_n = \langle(X - \bar{X})^n\rangle = \int_{-\infty}^{\infty} f_X(x)(x - \bar{X})^n dx \tag{17}$$

As above, $\mu_0 = 1$, but in this case $\mu_1 = 0$ by definition. The *variance* of a distribution (usually denoted σ^2) is the second central moment $\sigma^2 = \mu_2 = \langle(X - \bar{X})^2\rangle$ and is a measure of the width of a distribution. σ is sometimes called the *standard deviation* or the R.M.S., short for the root mean square of the distribution.

Other useful quantities are the *skewness*, γ_1, and *kurtosis*, γ_2, of the distribution, which are related to the third and fourth central moments respectively, each normalised by the appropriate power of the variance. Specifically, these are defined as:

$$\gamma_1 \equiv \frac{\mu_3}{\mu_2^{3/2}} \tag{18}$$

$$\gamma_2 \equiv \frac{\mu_4}{\mu_2^2} - 3 \tag{19}$$

These are defined in a way so that they are both zero for a Gaussian distribution, and are actually ratios of the cumulants (see below.)

In general a central moment is simply related to the algebraic moments of the same order and below. For example, the variance of a distribution, $\mu_2 = \langle (X - \bar{X})^2 \rangle$ can be shown to be $\mu_2 = \mu_2' - (\mu_1')^2$. (QUESTION: Express μ_3 in terms of μ_1', μ_2', μ_3'.)

3.2 The characteristic function

One very useful function is called the *characteristic function*, defined as

$$\phi_X(t) \equiv \langle e^{itX} \rangle = \int_{-\infty}^{\infty} f_X(x) e^{itx} dx = \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \langle X^r \rangle. \tag{20}$$

This is effectively the Fourier transform of $f_X(x)$ and can be easily inverted,

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_x(t) e^{-itx} dt. \tag{21}$$

One nice thing about the characteristic function is that it is easy to use it to generate moments of a pdf:

$$\mu_n' = i^{-n} (\frac{d}{dt})^n \phi_X(t)|_{t=0} \tag{22}$$

The central moments can be generated similarly by including a factor which removes the mean of the distribution:

$$\mu_n = i^{-n} (\frac{d}{dt})^n e^{-it\bar{X}} \phi_X(t)|_{t=0}. \tag{23}$$

Often we want to evaluate the distribution function of a sum of two variables, each of which has their own pdf; the result requires a convolution of their pdf's. Another nice feature of the characteristic function is that in Fourier space, convolutions in real space become simple products:

$$\phi_{X+Y}(t) = \phi_X(t) \phi_Y(t) \tag{24}$$

(QUESTION: Use this feature to show that the sum of two Gaussians is also Gaussian and find the resulting variance.)

The characteristic function is closely related to the *moment generating function*, which is defined as the Laplace transform rather than the Fourier transform, of the distribution function. That is,

$$M_X(t) \equiv \langle e^{tX} \rangle = \int_{-\infty}^{\infty} f_X(x) e^{tx} dx. \tag{25}$$

Many of the properties of the characteristic function apply also to the moment generating function.

3.3 Cumulants and connected moments

Related to the characteristic function is the *cumulant generating function*:

$$K_X(t) \equiv \ln \phi_X(t) \equiv \sum_{r=0}^{\infty} \kappa_r \frac{(it)^r}{r!}. \tag{26}$$

Here we have defined the *cumulants*, κ_n, which are also called the connected (or reduced) moments of the distribution. The cumulants are related to the cumulant distribution function in the same way as the moments are related to the characteristic function:

$$\kappa_n = i^{-n}(\frac{d}{dt})^n K_X(t)|_{t=0}. \tag{27}$$

Cumulants are clearly related to the ordinary moments of the distribution. The 'reduced' label relates to the fact that higher order moments are generated by the existence of lower order ones. The cumulants are what are left over from a moment when the effects of the lower moments are subtracted off. For example, consider distribution with zero mean distribution variance σ^2. It will have a fourth order moment, μ_4, part of which arises from the variance which is $3\sigma^4$, so the resulting cumulant is $\kappa_4 = \mu_4 - 3\sigma^4$. (The factor of three arises from the different ways two pairs can be made from four variables.) Note that this is why the kurtosis is defined as it is.

PICTURES A LA BERNARDEAU ET AL.

(QUESTION: Show that for a Gaussian distribution, all cumulants higher than κ_2 vanish.)

This leads directly to Wick's theorem below: since the Gaussian distribution has no higher cumulants, all higher moments arise from the variance.

4 Some useful distributions

4.1 Flat and reciprocal distributions

The simplest distribution to consider is the flat, or uniform distribution. Unfortunately, it suffers from not being normalisable, so some limits must be introduced. For example, if we choose the limits to be a and b, then the uniform distribution is $f_X(x) = 1/(b-a)$ when x is between a and b and is zero otherwise. This distribution has mean $(a+b)/2$ and variance $(b-a)^2/12$. It is often used as a prior for 'location' kinds of variables.

For the uniform prior, the probability that X lies in a range dx is proportional to the size of the range, $\mathcal{P}(x \leq X \leq x + dx) \propto dx$. Another simple distribution, called the *reciprocal distribution*, has the probability proportional to the logarithmic interval; that is, $\mathcal{P}(x < X < x + dx) \propto dx/x$. This type of distribution is used for 'scale' kinds of variables, such as proportionality constants, which are often multiplicative factors. Like the flat prior, the reciprocal prior is not well behaved without some limits. With the same choice of limits as above, the Jeffreys' distribution is $f_X(x) = 1/x \ln(b/a)$ when x is between a and b and is zero otherwise and its mean is $(b-a)/\ln(b/a)$.

4.2 Gaussian distribution

The most commonly encountered distribution, and in many ways the best behaved, is called the Gaussian, or normal, distribution. The Gaussian distribution is

$$f_X(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{-\frac{1}{2}(x-\bar{X})^2/\sigma^2}. \tag{28}$$

It has mean $\bar{X}$ and variance σ^2, and its characteristic function also takes a Gaussian form:

$$\phi_X(t) = e^{-\frac{1}{2}t^2\sigma^2 + it\bar{X}} \tag{29}$$

(It's worth understanding how this is derived, as Gaussian integrals often arise in probability theory.) The Gaussian has no higher cumulants beyond κ_1 and κ_2, which are just given by the mean and the variance respectively.

In some sense the Gaussian is most generic distribution. The *Central Limit theorem* shows that if you consider the sum of a large number of identically distributed independent variables, its distribution will approach a Gaussian as the the number of variables grows larger. This is independent of the distribution function of the variables. (Assuming the distribution isn't too pathological, and has well defined mean and variance.)

4.3 χ^2 distribution

The χ^2 distribution is the distribution expected from the sum of the square of a number of Gaussianly distributed variables. The number of such variables is called the number of degrees of freedom, p. The χ^2 distribution is

$$f_X(x, p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}. \tag{30}$$

It has mean p and variance $2p$. Its characteristic function is given by,

$$\phi_X(t, p) = (1 - 2it)^{-p/2}. \tag{31}$$

Note that as you increase the number of degrees of freedom, the distribution begins to look more and more like a Gaussian, as predicted by the Central Limit theorem.

Often, under the assumption of Gaussian errors, a 'goodness of fit' statistic called the χ^2 statistic is quoted for a theoretical fit to a set of data, defined as

$$\chi^2 = \sum_{i=1}^{N} (y_i - g(x_i))^2/\sigma_i^2, \tag{32}$$

where, the data are $[x_i, y_i]$, with y errors σ_i, and model for the data given by the function $g(x)$. This should be distributed as a χ^2 with the number of degrees of freedom given by $p = N - M$, where N is the number of data points and M is the number of parameters in the model. (This is because a model with the same number of parameters as there are data points can usually fit the data exactly, without any residuals, so $\chi^2 = 0$.)

Sometimes, the 'reduced χ^2' value is given instead, defined as

$$\chi^2_{red} = \chi^2/(N - M). \tag{33}$$

This is defined so that its expected value for a good model of the data is 1. Values of χ^2_{red} much smaller than 1 suggest that the error bars of the data were over estimated. Values of χ^2_{red} much higher than 1 could either indicate an underestimation of the error bars, or that you have a poor model for the data and should be looking for a better one. Sometimes however, you might just be unlucky, and you might have to live with a high χ^2_{red} until you get more data or can think of a better model.

Note that the definition of the χ^2 statistic given above implicitly assumes that there are no correlations between the error estimates of the various measurements. The more general expression is

$$\chi^2 = \sum_{i,j} (y_i - g(x_i)) C_{ij}^{-1} (y_j - g(x_j)), \tag{34}$$

where C_{ij} is the error covariance matrix defined below. When the error estimates are uncorrelated, C_{ij} is diagonal with $C_{ii} = \sigma_i^2$, so the expression trivially reduces to the one above.

4.4 Discrete distributions

Binomial distribution The first discrete distribution is called the binomial distribution. This distribution often occurs if one has a experiment which can take one of two values ('success' or 'failure'), such as flipping a coin (which is not necessarily a fair coin.) The distribution is given by:

$$\mathcal{P}_X(x, p, n) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \tag{35}$$

where x can take integer values between 0 and n, and p is a probability between 0 and 1. If we take the chance of success of a single trial to be p, then the binomial distribution gives the chances of seeing x successes in n trials. The mean of this distribution is np, and its variance is $np(1-p)$. One application of the binomial distribution would be if you are trying to estimate the probability of a galaxy being a particular type given some finite sample of galaxies.

Poisson distribution The Poisson distribution is a very generic discrete distribution which arises in a wide number of physics and astronomy contexts. For example, if you are given the average density of galaxies, what is the number you actually observe on a given patch? If you ignore correlations and assume the galaxies are randomly laid down, then the observed number should obey a Poisson distribution about the expected number (that is, the density*area).

The Poisson distribution is

$$\mathcal{P}_X(x, \lambda) = \frac{1}{x!} e^{-\lambda} \lambda^x, \tag{36}$$

where $x = 0, 1, 2, 3, \ldots$ and λ is the number expected on average. Thus, the mean is λ, but λ is also the variance of the distribution. If one is trying to estimate the density of sources given that you've observed N of them on a patch of sky of area A, the estimator is given as N/A with error $\sqrt{N}/A$.

The Poisson distribution from fairly few assumptions about the nature of the data. The central assumption is the independence of the events, i.e., the

probability of observing an event should be independent of what was seen previously. It also can be derived as the limit of a binomial distribution in the limit as $n \to \infty$ and where $\lambda = pn$ is fixed.

4.5 Other distributions

There are a large number of other distributions commonly occurring in the statistical literature: Cauchy, Gamma, exponential, lognormal, student-t, etc. These are related to those above in various extreme limits and while occasionally helpful, do not come up as often as those above. I just want to mention some of the more useful ones here.

Cauchy distribution One interesting distribution I want to mention is the Cauchy distribution, which is notable for its pathological properties. It comes up rarely, but it illustrates some reasons to be cautious. The Cauchy distribution is

$$f_X(x, \alpha, \sigma) = \frac{\sigma}{\pi(\sigma^2 + (x - \alpha)^2)}. \tag{37}$$

What is notable about the Cauchy distribution is that its mean and variance are undefined. Thus, one must be cautious about applying things like the Central Limit theorem to such distributions. (See the lighthouse problem in the book by Sivia for a simple case where the Cauchy distribution arises.)

Student t distribution The Student t distribution, or t distribution, can arise in a Bayesian context, though its more commonly seen in frequentist statistics.

$$f_T(t, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{38}$$

This distribution looks similar to a Gaussian, but with broader tails. It arises when trying to infer how far the true mean of a sample is from the sample mean (assuming a Gaussian distribution), after marginalising over an unknown variance (assuming a Jeffrey's prior on the variance and a flat prior on mean.)

Exponential distribution The exponential distribution is defined for positive arguments ($x \geq 0$):

$$f_X(x, \lambda) = \lambda e^{-\lambda x} \tag{39}$$

It has mean λ^{-1} and variance λ^{-2} and occurs naturally in the lengths of times between events in Poisson processes (e.g. radioactive decay.)

Lognormal distribution The lognormal distribution is also defined for positive arguments ($x \geq 0$):

$$f_X(x, \mu, \sigma) = \frac{1}{(2\pi)^{\frac{1}{2}} x\sigma} e^{-\frac{1}{2}(\ln x - \bar{X})^2/\sigma^2}. \tag{40}$$

If a variable, y, has a normal distribution, then the distribution of $x = e^y$ is lognormal. Its mean is $e^{\mu+\frac{1}{2}\sigma^2}$, and its variance is $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$. The lognormal distribution is sometimes used to model the one point matter density distribution.

5 Multi-variate distributions

Thus far we have been focusing on one dimensional distributions, but all the concepts carry over directly to higher dimensional distribution functions. In this case, the one dimensional variable X is generalised into a multi-dimensional data vector, e.g., $\mathbf{X} = (X_1, X_2, \cdots X_N)$. It may be the case that the distributions are continuous in some directions and discrete in others. Expectations and marginalizations generalise into multi-dimensional integrations or sums, or a combination of the two.

5.1 Describing an arbitrary distribution

Bayesian reasoning provides a direct route to the posterior information, which contains all we know about the parameters at hand; but describing this distribution is not always easy, particularly for large dimensional spaces.

A starting point is finding the parameters where the posterior distribution peaks. However, it must be remembered that the distribution may easily be bimodal or multimodal, meaning that it has more than one local maximum. If the peaks are of comparable amplitude, you must be very careful.

In one or two dimensions, one can usually plot the full distribution function. Even so, you often want to summarise it with a few numbers. Typical is to give one or two sigma (68% or 95%) confidence ranges. Even here there is some ambiguity, e.g. for a 1-d confidence range, you could either define it as the region where the probability exceeds some threshold, or such that there is equal probability of being above or below the range. Alternatively, if you are trying to give an upper or lower limit, you might give a one sided range.

Beyond two dimensions, one must look at different ways to summarise the data. Most typical is either to look at moments of the distribution, or to find the curvature matrix around the peak. In directions where the curvature is high, the peak is sharply defined and the corresponding error bars are small. For flat directions, the parameters are poorly constrained and often correspond to *parameter degeneracies.*

For example, one can consider finding the eigenvectors of the curvature matrix and ordering them according to their eigenvalues. The eigenvectors correspond to linear combinations of the parameters which are independent of each other, at least around the peak. Some combinations will be well constrained, but those with small eigenvalues will be poorly constrained and indicate degeneracies. (Of course this presumes some metric for the space to compare curvature in different directions.)

One can also consider lower dimensional slices or projections of the data. To obtain correct lower dimensional distributions, one should marginalise, projecting the distributions down all other dimensions. Sometimes however, one considers simply taking slices through the peak of the distribution, assuming

that one somehow fixes the other parameters. These two approaches can give vastly different answers if the errors are correlated.

5.1.1 A cautionary note

Since it can be difficult to keep track of the full distribution functions, so often people resort to keeping a few moments as summary statistics. (This is particularly true when the parameter space gets very large.) For example, sometimes just the mean and the variance of a distribution are kept, and thenceforth the distribution is treated as Gaussian. While this is often a reasonable approximation, it can lead to very unphysical distributions. A common example is the estimation of the variance of a distribution, such as a power spectrum. The variance is a positive definite function, and if the underlying statistics are Gaussian, then the variance distribution will typically be of a positive definite χ^2 form. With many degrees of freedom, a Gaussian is a good approximation to the χ^2 form. However, with only a few degrees of freedom, the χ^2 is very non-Gaussian. A Gaussian with the same mean and variance will typically have a substantial tail into non-physical negative values and lead to non-physical results.

5.2 Moments

The discussion of moments in one dimension carries over to multiple dimensions, except now there are an array of new moments generated by combining different variables. That is, not only does one have moments such as $\langle X_1^n \rangle$ and $\langle X_2^n \rangle$, but you also have possibilities like $\langle X_1^n X_2^m \rangle$. The mean becomes an N dimensional vector with elements $\bar{X}_i = \langle X_i \rangle$, where N is the dimension of the probability space.

In addition, the variance is generalised into what is known as a covariance matrix. The covariance matrix is defined as

$$C_{ij} \equiv \langle (X_i - \bar{X}_i)(X_j - \bar{X}_j) \rangle. \tag{41}$$

(This is also sometimes called σ_{ij}^2.) This matrix is clearly symmetric, but while the diagonal entries must be non-negative, off-diagonal entries may be negative, and if this is the case the two variables are said to be anti-correlated.

The covariance of the two variables cannot be arbitrarily large, but is limited by the fact that two variables cannot be more correlated than one variable is with itself. This result is summarised in what is called the Cauchy-Schwarz inequality

$$|\langle X_i X_j \rangle| \leq \langle X_i^2 \rangle^{1/2} \langle X_j^2 \rangle^{1/2}, \tag{42}$$

which itself is a special case of an arbitrary dimensional generalisation called Hölder's inequality. The inequality becomes an equality only in the special case where the variables are simply linearly related.

Characteristic functions can also be generalised to higher dimensions, except now every variable X_i must have its own counterpart t_i, so it also becomes a vector $\mathbf{t}$. Thus,

$$\phi_{\mathbf{X}}(\mathbf{t}) \equiv \langle e^{i\mathbf{t}\cdot\mathbf{X}} \rangle = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{i\mathbf{t}\cdot\mathbf{x}} d^N\mathbf{x}. \tag{43}$$

The moments are generated as you might expect, for example,

$$\langle X_i X_j \rangle = i^{-2} \frac{d^2}{dt_i dt_j} \phi_{\mathbf{X}}(\mathbf{t})|_{\mathbf{t}=\mathbf{0}}. \tag{44}$$

5.3 Multi-variate Gaussian distribution

The space of multi-variate distributions is very large; even if the variables are independent, one can still consider all the possible products of the one dimensional distributions discussed above. Here, I just want to demonstrate a simple distribution for which the variables are not independent.

The multi-variant Gaussian looks like,

$$f_{\mathbf{X}}(\mathbf{x}) = \left[(2\pi)^N \det|\tilde{C}|\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{X}})^T \tilde{C}^{-1}(\mathbf{x}-\bar{\mathbf{X}})\right], \tag{45}$$

where $\tilde{C}$ is just shorthand for the covariance matrix defined above, and $\bar{\mathbf{X}}$ is the mean of the distribution. This distribution is often used to approximate more complicated distributions. The characteristic function for multi-variate Gaussian is given by the generalisation of the one dimensional result,

$$\phi_X(t) = e^{-\frac{1}{2}\mathbf{t}^T \tilde{C} \mathbf{t} + i\mathbf{t}\cdot\bar{\mathbf{X}}}; \tag{46}$$

this can be useful for marginalising over distributions.

Even if the distribution is non-Gaussian, it can still be useful to make a Gaussian approximation to it. If we know the position of the peak of an arbitrary distribution, we can always expand the log of the distribution in a Taylor series,

$$\begin{aligned} \ln f_{\mathbf{X}}(\mathbf{x_p}+\delta\mathbf{x}) &= \ln f_{\mathbf{X}}(\mathbf{x_p}) + \sum_i \left.\frac{\partial \ln f_{\mathbf{X}}(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x_p}} \delta x_i \\ &\quad + \frac{1}{2}\sum_{ij} \left.\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x})}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\mathbf{x_p}} \delta x_i \delta x_j + \cdots \end{aligned} \tag{47}$$

Cutting off at the third term leads to a Gaussian approximation. The first term shows how high the peak of the function is. If we evaluate about the peak of the likelihood, then the linear terms are zero. $-\frac{\partial^2 \ln f_{\mathbf{X}}(\mathbf{x})}{\partial x_i \partial x_j}$ is sometimes called the *curvature matrix* for the parameters or the *Hessian*. For a Gaussian distribution, the curvature matrix is simply the inverse of the correlation matrix ($\tilde{C}^{-1}$) and the peak corresponds to the mean of the distribution.

5.3.1 Moments of a Gaussian

As mentioned above, all the connected moments of the Gaussian distribution above the second ones are zero. All the information about the distribution is contained in the correlation function (assuming we work in coordinates where the mean is zero.) It follows that any higher order moments of the Gaussian distribution can be written in terms of the correlation function alone.

Precisely how this is done is described in Wick's (or Isserlis') theorem. By symmetry, all odd higher order central moments are zero. All even moments can be generated by adding all possible permutations of pairs; that is,

$$\langle X_i X_k X_l X_k \cdots X_p X_q \rangle = \langle X_i X_j \rangle \langle X_k X_l \rangle \cdots \langle X_p X_q \rangle + \text{permutations}. \tag{48}$$

For example, the three point moment $\langle X_i X_k X_l \rangle = 0$, while the general four point moment is given by

$$\langle X_i X_k X_l X_k \rangle = \langle X_i X_j \rangle \langle X_k X_l \rangle + \langle X_i X_k \rangle \langle X_j X_l \rangle + \langle X_i X_l \rangle \langle X_j X_k \rangle. \quad (49)$$

In general, if one is interested in $2n$-point moments, there are $(2n)!/2^n n!$ permutation terms.

Deriving the moments Wick's theorem follows from taking derivatives of the characteristic function, i.e.

$$\langle X_i X_k \cdots X_p X_q \rangle = i^{-n} \frac{\partial}{\partial t_i} \frac{\partial}{\partial t_j} \cdots \frac{\partial}{\partial t_p} \frac{\partial}{\partial t_q} e^{-\frac{1}{2}\mathbf{t}^T \tilde{C} \mathbf{t}}|_{\mathbf{t}=\mathbf{0}}. \quad (50)$$

Each application of the derivative generates a polynomial in components of $\mathbf{t}$ multiplied times the Gaussian factor; however, only the final constants in the final polynomial survive setting $\mathbf{t} = \mathbf{0}$. For odd moments, the polynomial will be odd, leaving nothing when $\mathbf{t} = \mathbf{0}$. For even moments, the constants in the polynomial arise when half the derivatives act on the prefactor and half act on the exponential. This automatically generates the permutations of pairs. For example,

$$\begin{aligned}
\langle X_1 X_2 X_3 X_4 \rangle &= \frac{\partial}{\partial t_1} \frac{\partial}{\partial t_2} \frac{\partial}{\partial t_3} (-\tilde{C}_{4i} t_i) e^{-\frac{1}{2}\mathbf{t}^T \tilde{C} \mathbf{t}}|_{\mathbf{t}=\mathbf{0}} \\
&= \frac{\partial}{\partial t_1} \frac{\partial}{\partial t_2} (-\tilde{C}_{43} + \tilde{C}_{4i} t_i \tilde{C}_{3j} t_j) e^{-\frac{1}{2}\mathbf{t}^T \tilde{C} \mathbf{t}}|_{\mathbf{t}=\mathbf{0}} \\
&= \frac{\partial}{\partial t_1} (\tilde{C}_{43} \tilde{C}_{2i} t_i + \tilde{C}_{42} \tilde{C}_{3i} t_i + \tilde{C}_{32} \tilde{C}_{4i} t_i - \tilde{C}_{4i} t_i \tilde{C}_{3j} t_j \tilde{C}_{2k} t_k) e^{-\frac{1}{2}\mathbf{t}^T \tilde{C} \mathbf{t}}|_{\mathbf{t}=\mathbf{0}} \\
&= \tilde{C}_{43} \tilde{C}_{21} + \tilde{C}_{42} \tilde{C}_{31} + \tilde{C}_{32} \tilde{C}_{41}. \quad (51)
\end{aligned}$$

5.3.2 Marginalization of Gaussian distributions

Marginalisation over a subset of variables is straight forward for Gaussian distributions using the characteristic function. For characteristic functions, marginalisation over a variable is equivalent to setting its reciprocal t variable to zero. Thus, the covariance matrix elements for the remaining variables is unchanged, and the covariance matrix is replaced by a subset, or partition, of the original matrix.

On the other hand, the inverse covariance matrix is dramatically changed; however, it can be found using the partition method for inverting matrices. That is, suppose we know the full inverse covariance matrix, then we can partition it as

$$\tilde{C}^{-1} = \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix} \quad (52)$$

where $\tilde{P}$ and $\tilde{S}$ are $p \times p$ and $s \times s$ square matrices, while $\tilde{Q}$ and $\tilde{R}$ are $p \times s$ and $s \times p$ matrices respectively. We can find the $p \times p$ inverse matrix after marginalisation over the s variables to be,

$$\tilde{D}^{-1} = \tilde{P} - \tilde{Q} \tilde{S}^{-1} \tilde{R}. \quad (53)$$

Thus, we can get the new inverse correlation matrix only having to invert the submatrix describing the parameters we are marginalising over. This is quite useful when marginalising over a single parameter.

Sometimes, rather than marginalising, we might decide instead to fix some parameters to particular values, effectively looking at a slice through the probability density. In this case it is the inverse covariance elements which are fixed, while the covariance matrix changes. However, the same trick using the partition method can be used, with $\tilde{C}$ replacing $\tilde{C}^{-1}$ above. Note that if the values chosen for the slices differ from their mean values, this can lead to a shift in the mean of the remaining variables if cross correlations exist.

5.4 Principal component analysis

Principal component analysis tries to find combinations of the parameters which are independent and well constrained; this is done by finding the eigenmodes of either the correlation or inverse correlation matrix (or the Fisher matrix, which is the projected inverse correlation matrix.) One begins by solving for a normalised eigenvector basis. (The covariance and inverse covariance matrices share eigenvectors, though their eigen values are inverted.) From the normalised eigenvectors, $\mathbf{w}_i$ we can construct a rotation matrix $W = (\mathbf{w_1}, \mathbf{w_2}, \cdots \mathbf{w_n})$, where $W^T W = 1$. In the eigen mode basis, the covariance matrix is diagonal,

$$D = W^T C W \tag{54}$$

with values given by the eigenvalues of C, and the eigenvectors are typically ranked in order of their eigenvalues. Whether the large or small eigenvalues are most interesting depends on the problem one wants to solve.

If one is trying to represent known data with principal components, the most interesting modes are those with large variance because most of the amplitudes will lie in these directions. One can usually do a very good job of representing the data using just the modes where the variance is seen to be largest. Thus, PCA approaches can be a very efficient means of compressing the data.

On the other hand, if one is interested in understanding what combinations of data are or will be well constrained, then it is the smallest variances which are the most interesting. Large variances will correspond to combinations where the constraints are flat, which correspond to parameter degeneracies. Through PCA approaches, one can learn precisely what a given kind of observation will tell you. (Clearly, if one is working instead with the inverse correlation or Fisher matrix, the eigenvalues reflect the curvature of the probability space, and well constrained modes correspond to large curvature modes.)

This method is perhaps most interesting when the parameters are of the same type, such as in *non-parametric models.* (After other parameters have been marginalised out perhaps.) Non-parametric models generally means parameterising functions by some kind of binning or expansion, rather than tying them to parameters of a specific underlying model. The term nonparametric is somewhat a mis-nomer; it does not imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance.

5.5 Changing variables

Often one has a known distribution in terms of some variables, $\mathbf{x}$, and one wishes to find the distribution of some other variables $\mathbf{u} = \mathbf{g}(\mathbf{x})$. Assuming this transformation of variables is one-to-one, this can be done by finding the inverse transformation, $\mathbf{x} = \mathbf{h}(\mathbf{u})$. It is straight forward to show,

$$f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{X}}(\mathbf{h}(\mathbf{u}))|J|. \tag{55}$$

Here J is the *Jacobian* of the transformation, defined as the determinant of a matrix of partial derivatives, that is,

$$J = \det\left(\frac{\partial h_i}{\partial u_j}\right). \tag{56}$$

We can use this to derive the fact that the distribution function for $X + Y$ is given by the convolution of the distribution functions of X and Y. If X and Y are independent (see below), we can write their combined distribution as the product of their individual distributions:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{57}$$

We now make the change variables from X, Y to $U = X + Y, V = Y$. Inverting this transformation, we have $X = U - V$ and $Y = V$. The Jacobian in this case looks like

$$J = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix}, \tag{58}$$

so the determinant of the Jacobian is unity. Substituting this into the above expression and integrating over the possible values of V gives

$$f_U(u) = \int dv f_{U,V}(u, v) = \int dv f_X(u - v) f_Y(v). \tag{59}$$

This proves the earlier statement that the distribution function for the sum of two independent variables is just the convolution of their individual distributions.

Another example is to derive the χ^2 distribution from the Gaussian distribution. That is, supposing X is Gaussianly distributed with unit variance, what is the distribution for X^2? We can find this by making the transformation from X to $U = X^2$. To ensure transformation is one to one, we will assume that X is positive definite and double the usual probability density to keep the total probability integral unity. Solving for $x = u^{1/2}$, we can find the Jacobian is $J = u^{-1/2}/2$. Thus the χ^2 distribution with one degree of freedom is:

$$f_U(u) = f_X(x = u^{1/2})|J| = \frac{u^{-1/2}}{\sqrt{2\pi}} e^{-u/2}. \tag{60}$$

QUESTION: Derive the χ^2 distribution for 2 and 3 degrees of freedom. That is, if X, Y and Z are Gaussianly distributed with zero mean and unit variance, find the distribution functions for $X^2 + Y^2$ and $X^2 + Y^2 + Z^2$.

Similar arguments can be used to show that if you add two independent Poisson distributed numbers with means θ and λ, you obtain a Poisson distributed number with mean $\theta + \lambda$.

6 Estimator theory

Since orthodox statistics still dominates most of science, it is important to understand how the frequentist approach works in practise. A key concept in this approach is the notion of an *estimator*, which in fact is any function of a sample. That is, any statistic can be an estimator. The goal of this approach is to find statistics of samples which will provide good estimates of the underlying parameters of the distribution. Thus this approach is also referred to as *sampling theory*.

6.1 Kinds of estimators

There are many ways of finding estimators. For example, in *the method of moments,* you measure the lowest k moments of a given sample, where k is the number of parameters of the model. You also can determine the expectation of these moments for a set of parameters, and these relations can be inverted to give the parameters in terms of the expected moments. Estimators for the parameters are found by evaluating these functions substituting the measured moments.

There is nothing really special about the use of moments in this technique. Any set of k statistics could have been used in the same way to find estimators of the parameters. Thus, the family of possible estimators is infinite.

Another popular technique is the *maximum likelihood estimator* (MLE), which attempts to find the parameters for which the likelihood is highest; that is, it finds the parameters for which the data would be most likely to be observed. This is closely related to the Bayesian approach, except it ignores any prior knowledge.

6.2 Evaluating estimators

There are arbitrary many ways of constructing estimators, so how does one choose between them? Some criteria have been suggested which I will quickly go through here.

The first question is whether the estimator is *biased*. Suppose you have a statistic W which is an estimator for the parameter θ. The bias of the statistic is given by $b = \langle W \rangle - \theta$. Usually, we prefer to consider unbiased estimators, though sometimes biased estimators can have advantages.

Another measure of the quality of an estimator is the mean squared error (MSE), defined as $\langle (W-\theta)^2 \rangle$. It is easy to show that the MSE is related to both the variance of the estimator and the bias of the estimator,

$$\langle (W-\theta)^2 \rangle = \langle (W - \langle W \rangle)^2 \rangle + (\langle W \rangle - \theta)^2 = \mathrm{var}(W) + b^2. \tag{61}$$

For an estimator to have small MSE, it must have small variability (precise) and have small bias (good accuracy). The choice of the MSE isn't unique; we could have used any increasing function of the absolute distance $|W-\theta|$ as a measure of the quality of the estimator.

6.3 Example: estimating mean and variance

Suppose you have many samples from a Gaussian distribution and are trying to determine the mean and variance of the distribution. Assume the underlying values are given by μ and σ^2. The standard estimator of the mean is

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{62}$$

which is unbiased since $\langle \bar{X} \rangle = \mu$. The MSE of this estimator is σ^2/N, the same as the variance of the estimator.

There are two estimators to consider for the variance. The standard one is

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2. \tag{63}$$

It is straightforward to show that this is unbiased, and has MSE of $2\sigma^4/(N-1)$. However, the maximum likelihood estimator for the variance is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2, \tag{64}$$

which is biased. However, despite the bias, it has MSE of $(2N-1)\sigma^4/N^2$ which is smaller than that of S^2. Thus, by trading off bias for a smaller variance, the MSE is improved. Thus, a biased estimator isn't always a bad thing.

QUESTION: Show $\hat{\sigma}^2$ is the maximum likelihood estimator, and derive the MSE for it and for S^2.

QUESTION: Measurements of the CMB power spectrum come down to $2\ell+1$ samples from a Gaussian distribution with zero mean. Estimate the percent uncertainty in the power (variance) as a function of ℓ using the MLE defined above. This is known as cosmic variance.

6.4 The Cramer-Rao bound

With the notions of bias and MSE, we can search for the best estimator for a particular parameter. Unfortunately, there is not a prescribed way of finding the best estimator, but we can at least figure out how good an estimator is compared to how good it could possibly be.

If we restrict attention to unbiased estimators, then the MSE is given by the variance of the estimator. We can then use the Cauchy-Schwarz inequality to derive a bound on how small the variance could be. This is known as the Cramer-Rao bound, and is given by,

$$\langle (W - \langle W \rangle)^2 \rangle \geq \left\langle \left(\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta) \right)^2 \right\rangle^{-1}. \tag{65}$$

The right hand side is the inverse of the Fisher information. The bound for biased estimators is modified by a factor of $(\partial\langle W\rangle/\partial\theta)^2$ on the right hand side, and the equality holds only when θ is Gaussian distributed (shown below.)

This generalises for multi-dimensional parameter estimation. For a large family of distributions, it can be shown that

$$F_{ab} \equiv \left\langle \left(\frac{\partial}{\partial \theta_a} \ln f_{\mathbf{X}}(\mathbf{x}|\theta) \right) \left(\frac{\partial}{\partial \theta_b} \ln f(\mathbf{x}|\theta) \right) \right\rangle = - \left\langle \frac{\partial^2}{\partial \theta_a \partial \theta_b} \ln f_{\mathbf{X}}(\mathbf{x}|\theta) \right\rangle . \quad (66)$$

This is defined to be the Fisher information matrix. Formally in estimator theory, $f_{\mathbf{X}}(\mathbf{x}|\theta)$ is the likelihood function, rather than the full posterior distribution. The Fisher matrix is useful for making predictions for what a proposed experiment will tell us about the underlying theory and it is simply the expectation of the curvature matrix, where the expectation denotes an averaging over data consistent with a given model. Thus, the Fisher matrix is defined for a particular theory and is independent of any actual data.

The Fisher matrix is useful because it allows us to place a lower bound on the derived parameter errors resulting from an experiment. In particular, if all the parameters are fixed but one, then its error is bounded by $\sigma_i^2 \geq 1/F_{ii}$. If we marginalise over the other parameters, $\sigma_i^2 \geq (F^{-1})_{ii}$. These limits are the same as what we found above for the Gaussian case, where the bound is saturated. However, it must be realised that the Fisher matrix approach only gives lower bounds. Real experiments must deal with potential non-Gaussianity and possible systematic errors, so the resulting error bounds will rarely achieve the predictions of the Fisher matrix approach.

6.4.1 Deriving the Cramer-Rao bound

Let us briefly go through the derivation of the Cramer-Rao bound in the one dimensional case where we have a single estimator, $W(\mathbf{x})$, for the parameter θ. It is helpful to define a new quantity, $V \equiv \partial \ln f_{\mathbf{X}}(\mathbf{x}|\theta)/\partial\theta$. This quantity has zero expectation:

$$\begin{aligned} \langle V \rangle &= \int d^N\mathbf{x} f_{\mathbf{X}}(\mathbf{x}|\theta) \frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta) = \int d^N\mathbf{x} \frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x}|\theta) \\ &= \frac{\partial}{\partial\theta} \int d^N\mathbf{x} f_{\mathbf{X}}(\mathbf{x}|\theta) = \frac{\partial}{\partial\theta} 1 = 0. \end{aligned} \quad (67)$$

Here we have used the fact that the probability density is normalised to unity.

We next calculate the covariance of W and V (here for simplicity we assume that W is and unbiased estimator for θ, $\langle W \rangle = \theta$); the covariance is

$$\begin{aligned} \langle (W-\theta)V \rangle &= \langle WV \rangle = \int d^N\mathbf{x} f_{\mathbf{X}}(\mathbf{x}|\theta) W(\mathbf{x}) \frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta) \\ &= \int d^N\mathbf{x} W(\mathbf{x}) \frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x}|\theta) = \frac{\partial}{\partial\theta} \int d^N\mathbf{x} W(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta) \\ &= \frac{\partial}{\partial\theta} \langle W \rangle = \frac{\partial}{\partial\theta} \theta = 1. \end{aligned} \quad (68)$$

We can think of the space of functions on $\mathbf{x}$ as a vector space, where the scalar product is defined by the two point expectation value. Since the probability is positive definite, it follows that the two point moment of any non-zero function is positive, and basically can be used to define a scalar magnitude $||g(\mathbf{x})|| \equiv \langle g(\mathbf{x})^2 \rangle^{1/2}$. Given this, it is straight forward to derive the Cauchy-Schwarz inequality,

$$\langle g(\mathbf{x}) h(\mathbf{x}) \rangle^2 \leq \langle g(\mathbf{x})^2 \rangle \langle h(\mathbf{x})^2 \rangle. \quad (69)$$

As in the usual vector case, the bound is saturated only when the functions are linearly proportional to one another. We can apply this to the functions $W(\mathbf{x}) - \theta$ and $V(\mathbf{x})$ to find

$$\langle (W(\mathbf{x}) - \theta)V(\mathbf{x})\rangle^2 \leq \langle (W(\mathbf{x}) - \theta)^2\rangle\langle V(\mathbf{x})^2\rangle. \tag{70}$$

Combining this with the calculation of the covariance, we derive the Cramer-Rao bound

$$\langle (W(\mathbf{x}) - \theta)^2\rangle \geq 1/\langle V(\mathbf{x})^2\rangle \tag{71}$$

where

$$\langle V(\mathbf{x})^2\rangle = \langle (\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta))^2\rangle \tag{72}$$

is the Fisher Information.

We can take a further derivative of $\langle V(\mathbf{x})\rangle$, which we know to be zero. From this we see,

$$\frac{\partial}{\partial\theta}\langle V(\mathbf{x})\rangle = 0 = \frac{\partial}{\partial\theta}\int d^N\mathbf{x} f_{\mathbf{X}}(\mathbf{x}|\theta)\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{x}|\theta) \tag{73}$$

The derivative yields two terms which must be equal and opposite,

$$\langle (\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta))^2\rangle = -\langle \frac{\partial^2}{\partial\theta^2} \ln f_{\mathbf{X}}(\mathbf{x}|\theta)\rangle \tag{74}$$

which provides the alternative form for the Fisher Information.

When is the bound saturated? It is interesting to know when Cramer-Rao bound is saturated; this is generally when the probability distribution of the parameter θ is Gaussian, which can be inferred from the discussion above. The Cauchy-Schwarz inequality is saturated when the functions are linearly related, $V = a(W - \theta)$, implying

$$a(W(\mathbf{x}) - \theta) = \frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{x}|\theta) \tag{75}$$

which can be integrated to find,

$$\ln f_{\mathbf{X}}(\mathbf{x}|\theta) = a(-\theta^2/2 + W(\mathbf{x})\theta) + C. \tag{76}$$

Taking the exponential of both sides, we can see that $f_{\mathbf{X}}(\mathbf{x}|\theta)$ must be Gaussian in the parameter θ (and peaking at $\theta = W(\mathbf{x})$) if the Cramer-Rao bound is saturated.

6.4.2 Fisher matrix for Gaussian likelihoods

The most common application of the Fisher matrix is when the probability of the data is Gaussianly distributed around some mean vector μ and with some covariance matrix $\tilde{C}$. Either μ or $\tilde{C}$ may in principle depend on the parameters θ. Recall for a Gaussian distribution,

$$\ln f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{1}{2}(\mathbf{x} - \mu)^T\tilde{C}^{-1}(\mathbf{x} - \mu) - \frac{1}{2}\ln\det\tilde{C} - \frac{N}{2}\ln(2\pi). \tag{77}$$

Let us first examine the case where the covariance is fixed and only the mean value μ depends on the parameters θ, which is often the case when one's error bars are determined solely by the measurement apparatus. In this case, the derivative with respect to a single parameter is given by

$$\frac{\partial}{\partial\theta_a}\ln f_{\mathbf{X}}(\mathbf{x}|\theta) = -\frac{\partial\mu^T}{\partial\theta_a}\tilde{C}^{-1}(\mu - \mathbf{x}) \tag{78}$$

remembering that $\tilde{C}$ and $\tilde{C}^{-1}$ are symmetric matrices, so that $\mathbf{x}^T\tilde{C}^{-1} = \tilde{C}^{-1}\mathbf{x}$ We can then evaluate the Fisher matrix,

$$F_{ab} = \left\langle \frac{\partial\mu^T}{\partial\theta_a}\tilde{C}^{-1}(\mu-\mathbf{x})\frac{\partial\mu^T}{\partial\theta_b}\tilde{C}^{-1}(\mu-\mathbf{x})\right\rangle \tag{79}$$

This in principle is the product of two scalars, each of which is a linear combination of components of the vector $\mu - \mathbf{x}$. Rearranging and recalling that the expectation of $\langle(\mu-\mathbf{x})_i(\mu-\mathbf{x})_j\rangle = \tilde{C}_{ij}$ we find that

$$F_{ab} = \frac{\partial\mu^T}{\partial\theta_a}\tilde{C}^{-1}\frac{\partial\mu}{\partial\theta_b}. \tag{80}$$

(If you have trouble seeing this, it helps to write all the indices explictly.)

Linear regression example We can use this to estimate errors which would result from a simple linear regression. Suppose we have a set of observations $\mathbf{x}, \mathbf{y}$ where the errors on the independent variable $\mathbf{x}$ are negligible, and the errors on the dependent variable are assumed to be uniform uncorrelated errors, $\tilde{C}_{ij} = \sigma^2\delta_{ij}$. Further, let us assume that there is a model with a linear relationship between the variables, $\mathbf{y} = a\mathbf{x} + b$ and our goal is to understand the expected errors on the parameters a and b. We evaluate the above expression assuming the mean values $\mu = a\mathbf{x}+b$. (Note that $\mathbf{x}$ now refers to the independent variable, and $\mathbf{y}$ takes the place of the data variable from above.) We can calculate the Fisher matrix to be, $F_{aa} = N/\sigma^2$, $F_{bb} = \Sigma_i x_i^2/\sigma^2$ and $F_{bb} = \Sigma_i x_i/\sigma^2$. Inverting the matrix, we find $\sigma_{aa}^2 = \sigma^2\Sigma_i x_i^2/(N\Sigma_i x_i^2 - (\Sigma_i x_i)^2)$ and $\sigma_{bb}^2 = N\sigma^2/(N\Sigma_i x_i^2 - (\Sigma_i x_i)^2)$. These match the expected errors seen for the maximum likelihood estimators seen in standard regression results.

The other alternative, where the mean is fixed and the covariance depends on theory parameters, arises all the time in analysing the statistics of cosmological fields, such as the density field or the temperature of the CMB. In this case,

$$\begin{aligned}\frac{\partial}{\partial\theta_a}\ln f_{\mathbf{X}}(\mathbf{x}|\theta) &= \frac{1}{2}x_i\left(\tilde{C}^{-1}\frac{\partial\tilde{C}}{\partial\theta_a}\tilde{C}^{-1}\right)_{ij}x_j - \frac{1}{2}\tilde{C}^{-1}_{ij}\frac{\partial\tilde{C}}{\partial\theta_a}|_{ji} \\ &= \frac{1}{2}(x_ix_j - \tilde{C}_{ij})\left(\tilde{C}^{-1}\frac{\partial\tilde{C}}{\partial\theta_a}\tilde{C}^{-1}\right)_{ji}\end{aligned} \tag{81}$$

where I used the relation $\frac{\partial\tilde{C}^{-1}}{\partial\theta_a} = -\tilde{C}^{-1}\frac{\partial\tilde{C}}{\partial\theta_a}\tilde{C}^{-1}$ and Jacobi's formula, $\frac{\partial\det\tilde{C}}{\partial\theta_a} = \det\tilde{C}\,\mathrm{Tr}\left(\tilde{C}^{-1}\frac{\partial\tilde{C}}{\partial\theta_a}\right)$. Note that when the covariance is changing, it is important to include the normalisation determinant term.

We now are in a position to calculate the Fisher matrix,

$$F_{ab} = \frac{1}{4}\left\langle (x_i x_j - \tilde{C}_{ij}) \left(\tilde{C}^{-1}\frac{\partial \tilde{C}}{\partial \theta_a}\tilde{C}^{-1}\right)_{ji} (x_k x_l - \tilde{C}_{kl}) \left(\tilde{C}^{-1}\frac{\partial \tilde{C}}{\partial \theta_b}\tilde{C}^{-1}\right)_{lk} \right\rangle . \tag{82}$$

Again, this is a product of two scalar terms, but they become mixed when the expectation values are applied. Using Wick's theorem for quartic moments, we find things simplify dramatically,

$$F_{ab} = \frac{1}{2}\mathrm{Tr}\left[\tilde{C}^{-1}\frac{\partial \tilde{C}}{\partial \theta_a}\tilde{C}^{-1}\frac{\partial \tilde{C}}{\partial \theta_b}\right] . \tag{83}$$

Example Let us consider the problem of estimating the variance of a number of uncorrelated measurements with zero mean. The model parameter is the variance σ^2, and the model for the covariance is simply $\tilde{C}_{ij} = \sigma^2\delta_{ij}$. Thus, $\tilde{C}^{-1} = \sigma^{-2}\delta_{ij}$ and $\frac{\partial \tilde{C}_{ij}}{\partial \sigma^2} = \delta_{ij}$, making the Fisher information $F = \sigma^{-4}/2\mathrm{Tr}\delta = N/2\sigma^4$. This means that the smallest error in an unbiased estimator for the variance is $2\sigma^4/N$, which should be compared to the results seen for the MSE of S^2 seen above. (However in that case the mean was assumed to be unknown; were it assumed to be zero, there would be exact agreement between the approaches.)

For the general case, where both the mean and variance depend on theory parameters, the Fisher matrix is the sum of these two terms:

$$F_{ab} = \frac{\partial \mu^T}{\partial \theta_a}\tilde{C}^{-1}\frac{\partial \mu}{\partial \theta_b} + \frac{1}{2}\mathrm{Tr}\left[\tilde{C}^{-1}\frac{\partial \tilde{C}}{\partial \theta_a}\tilde{C}^{-1}\frac{\partial \tilde{C}}{\partial \theta_b}\right] . \tag{84}$$

(Its hard to think of an example where both dependences are there, but they must exist.)

Note that these expressions do not depend on the basis in which the observables and their correlations are calculated. In practise, we often choose a basis where this is easy to evaluate, for example working in Fourier space or spherical harmonic space where the correlations are diagonal, and the Fisher matrix becomes a simple sum of independent terms. Formally this diagonal property requires measurements over the whole space (e.g. over the full sky); however, often the partial coverage of the volume is put in simply by scaling the information by a simple factor, such as the fraction of the sky over which measurements exist. While this may not be exact, it should usually give a reasonable answer, given the aim is to forecast.

7 Applications of the Bayesian method

It is worth going through a few basic examples to demonstrate the Bayesian approach. In order to show analytical solutions, here we focus on Gaussian distributions; hopefully, these should help provide intuition for more complex cases.

7.1 Template fitting

One common question is pulling a weak signal out of noisy data, where one has an idea for what the signal looks like (a template), but no idea of its amplitude. This is essentially the equivalent to fitting for the slope of a data set, where you are looking for a signal in the data which is proportional to the independent coordinate. Template fitting is a common problem in astronomy, e.g. finding weak gravitational wave signals in noisy data, or searching for non-Gaussianity in noisy measurements of the three point moment.

Suppose we have many correlated measurements of some observable $\mathbf{y}$ which are subject to some noise or other source of covariance. Thus, our data model is $\mathbf{y} = b\mathbf{x} + \mathbf{n}$. What is the best estimate of b given the data and an error covariance matrix $\langle n_i n_j \rangle = C_{ij}$?

The first step is to examine the likelihood function, or the probability of observing the data given a particular value of b. If the errors are Gaussian (a common assumption), then the likelihood is simply:

$$\mathcal{L}(\mathbf{y}|b) = \frac{1}{(2\pi)^{N/2}|C|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - b\mathbf{x})^T C^{-1}(\mathbf{y} - b\mathbf{x})\right]. \tag{85}$$

If the prior is uniform over a wide range of b, the posterior distribution of b is simply proportional to the likelihood, which also happens to be Gaussian in b. The posterior distribution for b peaks where $\ln \mathcal{L}$ peaks, which we can determine by finding where the derivative with respect to b is zero. (This is known as the maximum likelihood value, so we will denote it as b_{ml}.) We can solve for it by,

$$\left.\frac{\partial \ln \mathcal{L}}{\partial b}\right|_{b=b_{ml}} = \mathbf{x}^T C^{-1}(\mathbf{y} - b_{ml}\mathbf{x}) = 0. \tag{86}$$

This has the solution that $b_{ml} = (\mathbf{x}^T C^{-1}\mathbf{y})/(\mathbf{x}^T C^{-1}\mathbf{x})$. Similarly, the curvature of the likelihood can be found by taking another derivative which allows us to find the variance, $\sigma^2_{ml} = (\mathbf{x}^T C^{-1}\mathbf{x})^{-1}$.

The maximum likelihood estimates are good so long as the data are much more informative than the prior on b. When they are comparable, both need to be kept to find the final posterior solution. This generically will not be solvable analytically, but if we consider a Gaussian prior on b, then we can do it in this case. Suppose our prior is Gaussian with mean b_{pr} and variance σ^2_{pr}. It is straightforward to show that with this prior, the posterior is Gaussian with mean

$$b_{post} = \frac{b_{ml}\sigma^2_{pr} + b_{pr}\sigma^2_{ml}}{\sigma^2_{pr} + \sigma^2_{ml}} \tag{87}$$

and variance

$$\sigma^2_{post} = \frac{\sigma^2_{ml}\sigma^2_{pr}}{\sigma^2_{ml} + \sigma^2_{pr}}. \tag{88}$$

These have precisely the behaviour we expect: when the prior data is weak, its variance will be large, and the posterior values reduce to the maximum likelihood values. Similarly, when the priors are very informative compared to the data, the posterior values are little changed from the prior values.

QUESTION - derive the above results for the posterior distribution.

7.2 Wiener filtering

Sometimes one is interested in trying to reconstruct the best estimate of a signal which has been measured in the presence of noise. Given the expected power spectra of the signal and the noise, it is possible to construct an optimal estimate of the signal, by applying what is known as a Wiener filter. This is sometimes derived as the optimal linear estimate of the signal (optimal in a least-square sense, e.g. Press et al.); however it also naturally results from the Bayesian approach, which is the approach I take here.

We model the data as, $\mathbf{d} = \mathbf{s} + \mathbf{n}$ and we assume we know the statistics of the signal and the noise are both Gaussian distributed. Their variances are $\langle s_i s_j \rangle = S_{ij}$ and $\langle n_i n_j \rangle = N_{ij}$ and for simplicity we assume they both have zero mean. We then wish to determine the signal which optimises the posterior probability,

$$\mathcal{P}(\mathbf{s}|\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}|\mathbf{s})\mathcal{P}(\mathbf{s})}{\mathcal{P}(\mathbf{d})} \propto e^{-\frac{1}{2}(\mathbf{d}-\mathbf{s})^T N^{-1}(\mathbf{d}-\mathbf{s})} \times e^{-\frac{1}{2}\mathbf{s}^T S^{-1}\mathbf{s}}. \tag{89}$$

The first term represents the likelihood of the data given the signal, and the second term is the prior probability for the signal. Since the signal and noise expectations are fixed, we simply need to minimise the term in the exponential with respect to the signal. From this we find,

$$-N^{-1}(\mathbf{d} - \mathbf{s}) + S^{-1}\mathbf{s} = 0 \tag{90}$$

so the signal map which maximises the posterior probability has the solution,

$$\mathbf{s} = S(S + N)^{-1}\mathbf{d}. \tag{91}$$

Thus far the analysis is completely general, but if one considers homogeneous fields, one can do the analysis in Fourier space where S and N are diagonal and are effectively the power spectra ($S(k)$ and $N(k)$). There the Wiener filter takes its standard form:

$$\mathbf{s_k} = \frac{S(k)}{S(k) + N(k)}\mathbf{d_k}. \tag{92}$$

Typically, low frequencies are signal dominated while high frequencies are noise dominated; in such a case, the Wiener filter becomes effectively a low pass filter.

7.3 Marginalising over nuisance parameters

Sometimes there are systematic effects in the data which have a known shape (or template) but an unknown amplitude. If one has a Gaussian prior for the possible amplitude of the systematic, one can marginalise over the prior to obtain a description of the data where the systematic no longer appears in the data model, but rather makes its impact in a modified covariance matrix.

Suppose we have a known foreground template, $\mathbf{f}$, which could be contaminating our data with an unknown amplitude, in the presence of Gaussian noise ($\langle n_i n_j \rangle = N_{ij}$.) Our data model thus is,

$$\mathbf{d} = \mathbf{s} + A\mathbf{f} + \mathbf{n} \tag{93}$$

and the corresponding likelihood is

$$\mathcal{P}(\mathbf{d}|\mathbf{s},A,\mathbf{f}) = \frac{1}{(2\pi)^{d/2}\det|N|^{d/2}} e^{-\frac{1}{2}(\mathbf{d}-\mathbf{s}-A\mathbf{f})^T N^{-1}(\mathbf{d}-\mathbf{s}-A\mathbf{f})} \tag{94}$$

Suppose we introduce a Gaussian prior on A, $\mathcal{P}(A) \propto e^{-A^2/2\sigma_A^2}$ and then marginalise over A. The terms involving A in the exponential are

$$\begin{aligned} &-\frac{1}{2}A^2(\sigma_A^{-2} + \mathbf{f}^T N^{-1}\mathbf{f}) - A(\mathbf{d}-\mathbf{s})^T N^{-1}\mathbf{f} \\ &= -\frac{1}{2}(A-\hat{A})^2(\sigma_A^{-2} + \mathbf{f}^T N^{-1}\mathbf{f}) + \frac{1}{2}((\mathbf{d}-\mathbf{s})^T N^{-1}\mathbf{f})^2/(\sigma_A^{-2} + \mathbf{f}^T N^{-1}\mathbf{f}) \end{aligned} \tag{95}$$

where the second line has been determined by completing the square and $\hat{A} = (\mathbf{d}-\mathbf{s})^T N^{-1}\mathbf{f}/(\sigma_A^{-2} + \mathbf{f}^T N^{-1}\mathbf{f})^{1/2}$. We can now marginalise over A to find a new effective likelihood,

$$\mathcal{P}(\mathbf{d}|\mathbf{s},\mathbf{f}) \propto e^{-\frac{1}{2}(\mathbf{d}-\mathbf{s})^T \tilde{N}^{-1}(\mathbf{d}-\mathbf{s})} \tag{96}$$

where

$$\tilde{N}^{-1} = N^{-1} - N^{-1}\mathbf{f}\,\mathbf{f}^T N^{-1}/(\sigma_A^{-2} + \mathbf{f}^T N^{-1}\mathbf{f}). \tag{97}$$

Perhaps remarkably this is not only invertible, but has a very simple form,

$$\tilde{N} = N + \sigma_A^2 \mathbf{f}\;\mathbf{f}^T \tag{98}$$

which follows from the *Sherman-Morrison formula.* If more nuisance parameters are marginalised, similar results can be derived using the Woodbury formula, a generalisation of the Sherman-Morrison formula.

The net effect then of the nuisance parameters is to change the effective covariance matrix, increasing the noise of the data modes which coincide with the nuisance templates. The addition variance of these modes depend on the prior for the noise amplitude. If a flat prior was assumed, this corresponds with $\sigma_A^2 \to \infty$, or infinite noise for these modes. In this case the covariance matrix itself becomes infinite, so its generally useful to put some weak prior on the nuisance amplitude.

8 Comparing models with the Bayesian Evidence

The discussion above centred finding the best parameters given some possible model. However, often we are not sure what the parameters are, or even how many of them there should be. Thus, we need to be able to compare between very different models with completely diffrent parameterisations. The Bayesian approach provides a way of addressing these issues, in which one of the most important principles of reasoning, *Occam's razor*, arises naturally. The basic idea of Occam's razor is that given two hypotheses which explain the observations equally well, we should prefer the simpler explanation, or effectively the one with the fewer free parameters.

Here is where the notion of the Bayesian evidence becomes important. Previously it was just used as a normalisation factor, but the evidence is more than that. It is the probability of observing the data given a particular model, but

integrating over all possible parameters of that model. That is, if model A has set of parameters θ_A for which you have priors $\mathcal{P}(\theta_A|M_A)$, then the evidence is given by

$$\mathcal{P}(\mathbf{D}|M_A) = \int d^n\theta_A \mathcal{P}(\mathbf{D}|\theta_A, M_A)\mathcal{P}(\theta_A|M_A). \tag{99}$$

Suppose you have two models, A and B, each with their own set of parameters (θ_A and θ_B) and you wish to discover which is most consistent with the priors and observations. Consider the ratio

$$R_{AB} = \frac{P(M_A|D,I)}{P(M_B|D,I)} = \frac{P(M_A|I)}{P(M_B|I)} \times \frac{P(D|M_A,I)}{P(D|M_B,I)}. \tag{100}$$

Before seeing the data, you begin with some relative confidences in the two models given all your prior information I, which is given by the first term on the right. The second represents the ratio of the evidences for each of the models, and is sometimes called the *Bayes factor*. R_{AB} represents the relative confidences in the two models, as modified by the observations. When R_{AB} is much larger than one, model A is favoured while model B is preferred if R_{AB} is small. If the ratio is of order 1, then one cannot make a clear preference.

One can get an intuition for the evidence by making some simplifying assumptions. Assume first that the prior is flat in some region, uniform in the region $\theta_{A,i}^{min} < \theta_{A,i} < \theta_{A,i}^{max}$; in this region it is simply given by the inverse of the prior volume, $V_A^{prior} = \prod_i(\theta_{A,i}^{max} - \theta_{A,i}^{min})$. Further, assume not only that the data are Gaussianly distributed, but that they imply that the inferred parameters are also Gaussianly distributed. (This is effectively assuming that the Cramer-Rao bound is saturated.) This implies that the likelihood is effectively,

$$\begin{aligned}\mathcal{P}(\mathbf{x}|\theta) &= \frac{1}{(2\pi)^{N/2}|\det C|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)} \\ &= \mathcal{P}(\mathbf{x}|\hat{\theta}) \times e^{-\frac{1}{2}(\theta-\hat{\theta}(\mathbf{x}))^T F(\theta-\hat{\theta}(\mathbf{x}))}.\end{aligned} \tag{101}$$

Here, $\hat{\theta}_A(\mathbf{x})$ are the parameters which maximise the likelihood, and $\mathcal{P}(\mathbf{x}|\hat{\theta})$ is just the value of the likelihood for those parameters, beyond which the function drops off exponentially with a curvature F described by the Fisher matrix (or inverse theory covariance.)

Now we can evaluate the evidence,

$$\mathcal{P}(\mathbf{x}|M_A) = \int d^n\theta_A \mathcal{P}(\mathbf{x}|\hat{\theta}_\mathbf{A}) \times e^{-\frac{1}{2}(\theta_\mathbf{A}-\hat{\theta}_\mathbf{A}(\mathbf{x}))^T F(\theta_\mathbf{A}-\hat{\theta}_\mathbf{A}(\mathbf{x}))} \times \frac{1}{V_A^{prior}} \tag{102}$$

where the integral is over the range defined by the prior. If we assume the data are infomative relative to the prior, and that the Gaussian is only takes significant values within the prior volume, we can perfom the integration to find,

$$\mathcal{P}(\mathbf{x}|M_A) = \mathcal{P}(\mathbf{x}|\hat{\theta}_A) \times (2\pi)^{n/2}|\det F^{-1}|^{1/2} \times \frac{1}{V_A^{prior}}. \tag{103}$$

The middle factor is effectively the volume of parameter space left after having seen the data (V_A^{post}), leaving the evidence as

$$\mathcal{P}(\mathbf{x}|M_A) = \mathcal{P}(\mathbf{x}|\hat{\theta}_A) \times \frac{V_A^{post}}{V_A^{prior}}, \tag{104}$$

or the probability of the best fit model times the fraction of the original prior volume that gives comparable likelihood to the best fit model. Thus it is not enough that the best fit model is likely; the model should also be relatively predictive in the sense that a large fraction of the parameter space should be consistent with the observations. This second factor is sometimes called *Occam's factor*.

8.1 A simple example

It may be helpful to see this in action on a very simple example. Assume that you measure some observable and find a some value with Gaussianly distributed error, $D = \overline{D} \pm \sigma$. Suppose you wish to compare two models: model A which predicts that $D = 0$, and model B which predicts that $D = \lambda$, where $\lambda_{min} > \lambda > \lambda_{max}$. Model A has no parameters, so the evidence is simply

$$P(D|M_A, I) = (2\pi)^{-\frac{1}{2}}\sigma^{-1}e^{-\overline{D}^2/2\sigma^2}. \tag{105}$$

The evidence for model B requires an integration over possible values of λ :

$$P(D|M_B, I) = \int_{\lambda_{min}}^{\lambda_{max}} d\lambda P(\lambda|M_B, I)(2\pi)^{-\frac{1}{2}}\sigma^{-1}e^{-(\lambda-\overline{D})^2/2\sigma^2}. \tag{106}$$

Let us assume that the prior for λ is flat over the range, so that $P(\lambda|M_B.I) = (\lambda_{max} - \lambda_{min})^{-1}$. Further, if we assume that the observation is well within the allowed range and that $\sigma \ll \lambda_{max} - \lambda_{min}$, then we can do the integration exactly to find $P(D|M_B, I) = P(\lambda = \overline{D}|M_B, I) = (\lambda_{max} - \lambda_{min})^{-1}$.

Thus we find

$$R_{AB} = \frac{P(M_A|I)}{P(M_B|I)} \times \frac{\lambda_{max} - \lambda_{min}}{(2\pi)^{\frac{1}{2}}\sigma} \times e^{-\overline{D}^2/2\sigma^2}. \tag{107}$$

Again, the first factor is the ratio of the priors, while the exponential factor is effectively the ratio of the best fit likelihood for model A to that for the best fit model of model B (i.e. $\lambda = \overline{D}$.) The second factor effectively penalises model B because it has a larger parameter space. Given two models which fit the data equally well, this factor will tell us to prefer the simpler model, the one with a smaller parameter space. Basically, for more complex models, the probability of any particular set of parameters is smaller, and so such models are disfavoured unless the data are inconsistent with simpler models.

In this case, if the initial parameter space for model B is small ($\lambda_{max} - \lambda_{min} \sim 20\sigma$), then model A is preferred only if $\overline{D} < 2\sigma$. However, this cutoff could grow significantly if the B parameter space were increased exponentially. That is, you should prefer the assumption that $D = 0$ if the observations show its within a few sigma of 0. But when its safe to infer $D \neq 0$ depends on how big your observed value is compared to the typical value you thought it might have before the observation.

QUESTION: One application of this is to compare a models with and without a cosmological constant. Particle physicists argue that the typical value is of order 10^{120} times larger than the best fit value for the observations. Taking this seriously, estimate how many sigma we need to conclude that we are really observing a cosmological constant.

8.2 Template searches

An important application of this discussion is trying to find particular signals in noisy data. For example, do we detect a galaxy in a map dominated by Gaussian noise? Or, is there evidence for a particular kind of gravitational wave source in the LIGO data? To make this decision, we need to evaluate the ratio of the evidences for models with and without the signals of interest.

Previously I described the concept of template matching, and showed that if we model the data as $\mathbf{d} = A\mathbf{t} + \mathbf{n}$, where A is the amplitude of the template $\mathbf{t}$ and the noise is assumed to be described by the covariance matrix N, then the maximum likelihood estimate for the template amplitude A is given by

$$\hat{A} = \mathbf{d}^T N^{-1}\mathbf{t}/\mathbf{t}^T N^{-1}\mathbf{t}. \tag{108}$$

The variance on this estimate was shown to be given by $\sigma_A^2 = 1/\mathbf{t}^T N^{-1}\mathbf{t}$. This process will necessarily produce some answer for the best amplitude of the template, but how do we decide whether the template actually exists in the data?

To answer this question, we must decide on what threshold of the amplitude is significant enough to consider whether the template has been detected. The amplitude of this threshold must be set by the theoretical priors and depends on how likely we believe we were to see it. The more rare the expected event, the higher the threshold that must be met to infer a detection. (E.g., when trying to detect point sources in astronomical data, it is useful to understand theoretically what their expected number density and amplitude distribution is.) The prior model should describe both the probability that the template exists (essentially the prior Bayes ratio), and the expected distribution of the amplitude if it does exist. Deciding the question of whether the template has been detected would require evaluating the evidence ratio for the two hypotheses, that the template is in the data and that it is not. This then reduces to the simple example discussed above. The lower the expectation of finding the template, the smaller is its prior Bayes ratio; thus, we require a higher evidence ratio to favor the template hypothesis and so a higher signal-to-noise detection.

Example: searching for point sources An application of this is the problem of point source detection in a noisy field (or one dominated by Gaussian fluctuations, such as the CMB.) In this case, if the point source is unresolved, then it should appear with a known shape given by the beam of the instrument. From theoretical considerations or other observations, we should have some idea of the density of such sources on the sky and thus the probability that it will be seen centered at any given point. We will also have some understanding of the luminosity distribution of the sources, which will play the role of the prior amplitude distribution. These considerations will inform our threshold for the detection of sources. In general, we would be looking for sources over the entire map, so we want to calculate $\hat{A}$ for templates centered at every point, giving us a map of the inferred amplitudes. This map is effectively a convolution of the original map with the beam shape, weighted by the noise and divided by a local sensitivity factor if the noise is inhomogeneous.

9 MCMC techniques

Suppose you have a method of determining the posterior distribution, $P(\mathbf{x})$, of a given set of parameters and you want to understand what the whole probability surface looks like, e.g. to find expectation functions or to perform marginalizations. One way of doing this is to evaluate the posterior on a grid of points in parameter space, and then perform some interpolation to estimate what the function looks like. As long as the grid is sufficiently small to resolve all the interesting features of the distribution, this should work reasonably well.

Unfortunately, evaluating the posteriors takes time, particularly if one has to run a relatively time consuming code like CMBfast for each point. Thus you are limited somewhat on the number of gridpoints you can use. This is not usually a problem if the parameter space is low dimensional, but quickly becomes difficult as the number of dimensions rises. Even for fairly smooth distributions, one needs a minimum ~ 10 points, so the total number rises as 10^n, where n is the dimension of parameter space. This starts to get difficult if the number of parameters exceeds 5 or 6.

In addition, to improve the accuracy of the results, ideally you want to sample better the regions which have higher probability, because they will dominate the integral. Otherwise, you will spend most of the time evaluating the function in regions where there's little chance of ending up, which is not very efficient. Ideally, we need a way of sampling the distribution where its highest, so we can get accurate answers with the minimum number of evaluations of the posterior distribution. This is where MCMC techniques come in.

MCMC stands for Markov Chain Monte Carlo. Monte Carlo indicates in some sense that rather than trying to solve the question exactly, there will be some element of randomness involved. The basic idea is to take some random path through parameter space evaluating the posterior function along the way. We then arrange it so that the paths are concentrated in the parts of parameter space where the distribution is highest.

The paths we choose are called Markov chains. The basic definition of such a chain is that the n^{th} point in the chain depends only where it was for the $n-1$ point, but not on any earlier points. Thus, the path has a very short term memory. The most basic example of a Markov chain is a random walk. For example, in one dimension suppose you flip a coin at every step, moving forward or back one unit depending on the outcome. Your position follows a random walk; where you go next depends on where you are now, but not how you got there.

9.1 Metropolis-Hastings algorithm

There are many MCMC algorithms; here we will focus on one simple example called the *Metropolis-Hastings algorithm*. One starts at a random place in parameter space, $\mathbf{x_0}$. Based on some function, called the *proposal density*, you randomly determine a trial next step, $\mathbf{x_t}$. The proposal density, $Q(\mathbf{x_0}, \mathbf{x_t})$, is can be any fixed function, but often is taken as a Gaussian centred on the current position. (Often we assume the proposal density is symmetric on interchanging $\mathbf{x_0}$ and $\mathbf{x_t}$, which simplifies things a bit.) If the probability of that trial step is higher than that of the present step, then we "accept" it as the next step. If the likelihood is lower than the present likelihood, then you should make that

step only sometimes, with probability equal to the ratio of the probabilities, $P(\mathbf{x_t})/P(\mathbf{x_0})$. If you decide not to take a given trial step, throw it out and repeat the process from the previous step.

The more general Metropolis-Hastings algorithm, with an asymmetric proposal density, is not much harder:

1. Randomly choose a starting point $\mathbf{x_0}$.
2. Use $Q(\mathbf{x}, \mathbf{x_t})$ to randomly choose a trial next point, $\mathbf{x_t}$.
3. Evaluate the ratio:
$$\alpha = \frac{P(\mathbf{x_t})Q(\mathbf{x_t}, \mathbf{x_0})}{P(\mathbf{x_0})Q(\mathbf{x_0}, \mathbf{x_t})} \tag{109}$$
4. If $\alpha > 1$, $\mathbf{x_t}$ becomes the next point in the chain.
5. If $\alpha < 1$, $\mathbf{x_t}$ becomes the next point only with probability α. Otherwise the $\mathbf{x_0}$ is repeated.
6. Begin again with this new starting point.

In the long time limit, this will sample the distribution perfectly. However, there are some ways of speeding up the process. One is to recognise that you might have begun in a poor place, so you might spend a lot of time initially making it to the peak of the distribution. At least initially, your chains will be biased because of this, over representing the region near the starting point. Thus the initial phase of getting near the peak, called the 'burn in', is generally thrown away. (For example, you might choose to ignore all points before the chain reached 10% of the maximum probability.)

Another important issue is deciding you have enough points in the chain. This will depend on the precise statistic you are interested in, and the nature of the probability distribution. For example, for a bimodal distribution with two isolated maxima, it may take quite a while for the chain to discover the second maximum and sample it sufficiently. One approach to decide a stopping point is to monitor how the statistic of interest changes as the chains grow in length; e.g., you can split the chain in two and see how the estimate changes from the first half to the second half. Alternatively, you can also consider multiple chains starting out at different points, and see the variance between the estimators for different chains. Typical chains in cosmology run of order 10^4 steps.

9.1.1 The proposal density

The other essential thing is to have a reasonable choice of proposal function. While the exact shape of the transition function is not that crucial (its often taken as a multi-variate Gaussian), it is important to get the width of the function correct. If the typical step size is too small, it will take forever to reach the likelihood maximum and the samples will be highly correlated. In addition, a small step size means the chain could be stuck for a long time in a local maximum which is not the global maximum, which would strongly bias your results.

On the other hand, if the step size is too large, you will find most of your steps being rejected. If you are near the peak of the distribution, every step you try to take will be to a much lower likelihood, and will be rejected with high

probability. Thus, it will take a long time for the chains to grow significantly. A typical way around these difficulties is to measure the local curvature of the function, or to run an initial chain to determine the curvature, and to use this to determine the initial transition function.

Also, if there are variables which are strongly correlated in the target distribution, the proposal distribution should reflect these correlations. Effectively, this keeps the chain targeting the regions of parameter space where the target distribution is highest. For a well tuned proposal density, typical acceptance rates are around 25%.

MCMC methods can be much faster at than simple integration techniques for large numbers of dimensions. This is because one effectively does a random walk in every direction simultaneously. A lower bound to the number of steps is $(L/\sigma)^2$ where L is the typical distance that the random walk must transverse, and σ is the typical step size in that dimension.

9.2 Formal criteria

What are the formal criteria to ensure that the Markov chain samples with the likelihood density? One criteria is that if you have an ensemble points of which is distributed according to the target probability density, and apply the transition function for each of these points, the resulting ensemble should also be distributed with the same probability. This is called *invariance of the distribution* under the transition function. One also has to be careful that all points in the distribution can be reached from the starting point using the transition function, that is, there are no detached paths, or paths that are periodically repeating themselves. Finally, often a constraint called *detailed balance* is imposed, which implies that the paths are reversible. Detailed balance means that the probability density of being at point $\mathbf{x}_1$ times the transition probability of going from point $\mathbf{x}_1$ to point $\mathbf{x}_2$, is identical to the probability density of being at point $\mathbf{x}_2$ times the transition probability of going from point $\mathbf{x}_2$ to point $\mathbf{x}_1$.

QUESTION - Show that this is true for the Metropolis-Hastings algorithm.

For more information on MCMC methods, I like the book by David MacKay, "Information theory, inference, and learning algorithms," Cambridge University Press (2003), Chapters 29 and 30. Also, see the COSMO-MC paper by Antony Lewis and Sarah Bridle.

9.3 Importance sampling

Given a chain, it is straight forward finding expectation values for the probability distribution. For any function of the parameters, the expectation is simply

$$\langle f(\mathbf{x})\rangle = \int d^N\mathbf{x} P(\mathbf{x}) f(\mathbf{x}) \simeq \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x_i}) \tag{110}$$

One caution on MCMC however, is that while they are usually good at finding expectation values, they are inefficient in finding reliable information about the tails of the distribution. Thus one should be very cautious in using them to evaluate statistics which are sensitive to the tails. Instead one should run

chains which sample that parameter space most relavant to the statistics one is trying to evaluate.

Sometimes one wishes to find expectations for a distribution $P'(\mathbf{x})$ which is similar to $P(\mathbf{x})$ but not the same. As long as it similar enough, you can use the same chain found for $P(\mathbf{x})$, with a relative weighting to account for the difference, i.e.,

$$\langle f(\mathbf{x})\rangle_{P'} = \int d^N\mathbf{x} P'(\mathbf{x}) f(\mathbf{x}) \simeq \frac{1}{N}\sum_{i=1}^{N} \frac{P'(\mathbf{x})}{P(\mathbf{x})} f(\mathbf{x_i}). \tag{111}$$

This is known as *importance sampling* and can be useful in a number of circumstances, for example if one wants to see how the results respond to slight changes in the prior assumptions.

In the CMB, where evaluating a likelihood for a set of parameters exactly can be computationally very expensive, sometimes an approximate code is used to find the chains initially. These chains can be sampled to remove strong correlations between steps, and then the true likelihoods can be found only at these steps. Then importance sampling is used to find the true posterior distribution.

10 Statistics of fields

Thus far the discussion has been as general as possible, but it's worth spending some time discussing statistics which characterise fields. Applications of this arise all the time in astronomy: e.g., in analysing time stream data (1-d), images on the sky (2-d) or density fields in space (3-d).

For fields, the data are indexed by a position variable, $\mathbf{x}$ (or in one dimension, it is sometimes taken as a time variable , t.) These data may be well determined or they may contain noise. One usually assumes one is seeing a random sample, and trying to find characteristics of the underlying distribution.

A few key assumptions make it possible to probe the underlying distribution. The first is *homogeneity,* which requires that any joint probability distributions remain the same when the set of positions, $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \cdots$, is translated (but not rotated). This implies that the probabilities depend on the relative, not absolute, positions. (This is sometimes called *stationarity* for one dimensional fields, particularly in the time domain.) The second property is *isotropy*, which implies that any joint probabilities are unchanged when the set of positions are rotated. Finally, *ergodicity* implies that all the information about an underlying distribution can be learnt from a single (if infinite) realisation of the field. Ergodicity is obviously of great importance in a cosmological context, since we have only one realisation of the Universe!

Ergodicity means that we can imagine an infinite single field which we are sampling at a few finite points. From homogeneity, any statistical properties should be invariant of where we measure, so if we average over a big enough volume, we should find the underlying statistical properties of the field. Thus in effect, we can replace expectations by a volume average.

10.1 Correlation functions

For example, consider a density field $\rho(\mathbf{x})$. The mean of the field can be found by

$$\bar{\rho} \equiv \langle \rho(\mathbf{x}) \rangle = \lim_{V \to \infty} \frac{1}{V} \int_V d^N \mathbf{x_0}\, \rho(\mathbf{x_0}), \tag{112}$$

which is the one point moment, and it must be independent of position by homogeneity. We could also measure the one point distribution function, which tells us what fraction of random samples the density lies in some range.

Let's now consider higher order moments of the field, but for simplicity subtract off the mean of the field; i.e., we will focus on $\delta(\mathbf{x}) = (\rho(\mathbf{x}) - \bar{\rho})/\bar{\rho}$. (In the earlier language, we focus on the central moments.)

The two point correlation is simply the covariance between the field measured at two points,

$$\xi(\mathbf{x}) = \langle \delta(\mathbf{x_0}) \delta(\mathbf{x_0} + \mathbf{x}) \rangle. \tag{113}$$

Again, homogeneity implies it can only be a function of the relative positions of the points. If in addition the field is isotropic, then the correlation is a function of the distance between the points alone, $\xi(|\mathbf{x}|)$. Finally, the variance of the field is the two point correlation function evaluated at the same point, $\xi(0)$.

The three point function is defined similarly,

$$\zeta(\mathbf{x_1}, \mathbf{x_2}) = \langle \delta(\mathbf{x_0}) \delta(\mathbf{x_0} + \mathbf{x_1}) \delta(\mathbf{x_0} + \mathbf{x_2}) \rangle. \tag{114}$$

Again, isotropy here means that the function only depends on the triangle of the relative positions of the points. There are three degrees of freedom, uniquely defined by the lengths of the edges.

We can define arbitrarily N point correlation functions in exactly the same way, and potentially they can tell us something fresh about the distribution. However, if we wish to focus on the new information, it makes sense to define connected correlation functions, completely analogously to the cumulants defined earlier. All the information of Gaussian fields is contained in the two point correlation function; any higher order correlations are simple functions of it, and can be found using Wick's theorem.

10.2 Fourier space statistics

We can also consider an alternate basis for the fields, and consider the statistics in this basis. The most useful basis to consider is Fourier space, defined as

$$\delta_{\mathbf{k}} = \frac{1}{V} \int d^N \mathbf{x} \delta(\mathbf{x}) e^{i \mathbf{k} \cdot \mathbf{x}}. \tag{115}$$

This has the inverse,

$$\delta(\mathbf{x}) = \frac{V}{(2\pi)^N} \int d^N \mathbf{k} \delta_{\mathbf{k}} e^{-i \mathbf{k} \cdot \mathbf{x}}. \tag{116}$$

We can now consider the two, three and higher point moments in Fourier space. The two point moments is

$$\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle = \frac{1}{V^2} \int d^N \mathbf{x} d^N \mathbf{x}' \langle \delta(\mathbf{x}) \delta(\mathbf{x}') \rangle e^{i(\mathbf{k} \cdot \mathbf{x} + \mathbf{k}' \cdot \mathbf{x}')}. \tag{117}$$

We can make the change of variable $\mathbf{x}'' = \mathbf{x}' - \mathbf{x}$ and use the fact that the two point function is only a function of the relative position $\mathbf{x}''$ to show,

$$\begin{aligned}\langle \delta_{\mathbf{k}} \delta_{\mathbf{k}'} \rangle &= \frac{1}{V^2} \int d^N\mathbf{x} d^N\mathbf{x}'' \langle \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{x}'') \rangle e^{i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{x} + i\mathbf{k}'\cdot\mathbf{x}''} \\ &= \delta_D(\mathbf{k}+\mathbf{k}') \frac{(2\pi)^N}{V^2} \int d^N\mathbf{x}'' e^{i\mathbf{k}'\cdot\mathbf{x}''} \xi(\mathbf{x}'') \\ &\equiv \frac{(2\pi)^N}{V} \delta_D(\mathbf{k}+\mathbf{k}') P(\mathbf{k}). \end{aligned} \tag{118}$$

Here, we have integrated over $\mathbf{x}$ to get the Dirac delta function and defined the *power spectrum* $P(\mathbf{k})$ as the Fourier transform of the two point correlation function. (Be warned, there are many slightly different conventions for this definition.) The correlation function is simply the inverse transform of the power spectrum:

$$\xi(\mathbf{x}) = \frac{V}{(2\pi)^N} \int d^N\mathbf{k} P(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{119}$$

If we further assume isotropy, the correlation depends only on the separation distance, and so the power spectrum depends only on the magnitude, k. In these cases the angular integrals can usually be simply performed. In two dimensions, this is

$$\xi(x) = \frac{2\pi V}{(2\pi)^2} \int_0^\infty k dk P(k) J_0(kx), \tag{120}$$

where $J_0(z)$ is the ordinary Bessel function. In three dimensions, this is instead,

$$\xi(x) = \frac{4\pi V}{(2\pi)^3} \int_0^\infty k^2 dk P(k) j_0(kx), \tag{121}$$

where $j_0(z) = \sin(z)/z$ is a spherical Bessel function. The variance of the field in either case is given by

$$\sigma^2 = \xi(0) = \frac{V}{(2\pi)^N} \int d^N\mathbf{k} P(k) \propto \int k^{N-1} dk P(k). \tag{122}$$

Often we consider power spectra which follow a power law in k, $P(k) \propto k^n$, which are called *scale independent*. Two are of special significance: the *white noise* spectrum where the power is constant ($n = 0$), corresponding to a delta function for the correlation function, and the *scale invariant* spectrum ($n = -N$), corresponding to equal contributions per log interval of k. The scale invariant spectrum is both infra-red and ultra-violet log divergent. Redder spectra ($n < -N$) are infra-red divergent, while bluer spectra ($n > -N$) are ultra-violet divergent. Note that in large scale structure, we usually have near scale invariance in the potential field ($n = -3$), which corresponds to a bluer index for the density fluctuations ($n = 1$).

The analog to the three point correlation function is called the *bispectrum*, defined as

$$\langle \delta_{\mathbf{k_1}} \delta_{\mathbf{k_2}} \delta_{\mathbf{k_3}} \rangle \equiv \frac{(2\pi)^N}{V} \delta_D(\mathbf{k_1}+\mathbf{k_2}+\mathbf{k_3}) B(\mathbf{k_1}, \mathbf{k_2}), \tag{123}$$

where the Dirac delta function comes from homogeneity as above. The bispectrum is the transform of $\zeta(\mathbf{x_1}, \mathbf{x_2})$, and given isotropy is a function only of

the shape of the triangle defined between $\mathbf{k_1}, \mathbf{k_2}$ and $\mathbf{k_3}$. Thus it is sometimes written as $B(k_1, k_2, k_3)$. Higher order correlations can be similarly defined.

Note that the power spectrum can be calculated from the magnitudes of the $\delta_{\mathbf{k}}$ alone, without any phase information. Since all the information of a Gaussian field is in the power spectrum, there should be none in the phases. Any phase correlations are a sign of non-Gaussianity.

10.3 Window functions

In real life its rarely possible to sample a field at an infinitesimal point; because of observational constraints like resolution issues, we actually measure a field smoothed over some scale. The kind of smoothing depends on the experiment, but in general the observed field is the underlying convolved with some window function, $W(\mathbf{y})$:

$$\delta^{obs}(\mathbf{x}) = \frac{1}{V}\int d^N\mathbf{y} W(\mathbf{y})\delta(\mathbf{x}+\mathbf{y}) \tag{124}$$

The observed two point function can be found with a double convolution of the underlying one:

$$\xi^{obs}(\mathbf{x}) = \frac{1}{V^2}\int d^N\mathbf{y} d^N\mathbf{y}' W(\mathbf{y})W(\mathbf{y}')\xi(\mathbf{x}+\mathbf{y}-\mathbf{y}'). \tag{125}$$

This can be a bit hard to evaluate directly, but it simplifies significantly in Fourier space, where convolutions become products. Thus,

$$\delta^{obs}_{\mathbf{k}} = W_{\mathbf{k}}\delta_{\mathbf{k}}, \tag{126}$$

where $W_{\mathbf{k}}$ is the Fourier transform of $W(\mathbf{x})$. Assuming an isotropic window function, the observed power spectrum is $P^{obs}(k) = W^2(k)P(k)$; so, for example, the observed correlation function in three dimensions is given by,

$$\xi^{obs}(x) = \frac{4\pi V}{(2\pi)^3}\int_0^\infty k^2 dk W^2(k)P(k)j_0(kx). \tag{127}$$

Typical window functions include Gaussian functions and tophat functions, and these have the effect of damping any structure smaller than the smoothing length; that is, the high k modes are suppressed. Conventionally, the window function is usually defined such that

$$\frac{1}{V}\int d^N\mathbf{y} W(\mathbf{y}) = 1. \tag{128}$$

For example, the 3-d Gaussian window function in real space is

$$W(\mathbf{x}) = \frac{V}{(2\pi)^{3/2}R^3}e^{-r^2/2R^2} \tag{129}$$

and its Fourier transform is simply $W_{\mathbf{k}} = e^{-k^2R^2/2}$ which $\to 1$ for $k \to 0$.

10.4 Selection function

It is also impossible to sample a field at every point in space. The above approach can also be used to take into account limits to the geometry of the sample. Here though, it is a product in real space, and a convolution in Fourier space. These constraints make the measurements less sensitive to large scale (low k) power. The observations should not be sensitive to wavenumbers much smaller than the inverse of the characteristic scale (D) of the survey.

The sample geometry is characterised by the selection function, $f(\mathbf{x})$, which is normalised similarly to the window function. For galaxy surveys, this is often the product of a sky mask times the redshift selection function. The observed structure is simply the product of the true structure times the selection function, $\delta^{obs}(\mathbf{x}) \equiv \delta(\mathbf{x})f(\mathbf{x})$. In Fourier space, this becomes a convolution,

$$\delta^{obs}_{\mathbf{k}} = \frac{V}{(2\pi)^N}\int d^N\mathbf{k}'\delta_{\mathbf{k}'}f_{\mathbf{k}-\mathbf{k}'}. \tag{130}$$

Note that the selection function explicitly breaks breaks translation invariance, meaning that the two-point statistics are no longer diagonal in Fourier space. However, their expectations can be calculated similarly,

$$\langle\delta^{obs}_{\mathbf{k}}\delta^{obs*}_{\mathbf{k}'}\rangle = \frac{V}{(2\pi)^N}\int d^N\mathbf{k}''P_{\mathbf{k}''}f_{\mathbf{k}-\mathbf{k}''}f^*_{\mathbf{k}'-\mathbf{k}''}. \tag{131}$$

Rather than trying to invert this expression directly for the true power spectrum, it is more common to use search some parameterisation of the underlying power spectrum for one which can reproduce the observed two-point statistics.

Yet one more complication exists when sampling a field over a finite patch, which is that we usually do not observe the fractional over density, but rather the total density. To construct overdensity, we would need to know the true background density of the field, whereas usually we only know the density observed locally. If our inferred density contrast is $\tilde{\delta}(\mathbf{x})$, it must be related to the true density contrast through, $\rho(\mathbf{x}) = \bar{\rho}(1+\delta(\mathbf{x})) = \bar{\rho}_{local}(1+\tilde{\delta}(\mathbf{x}))$, where

$$\bar{\rho}_{local} = \bar{\rho} + \frac{1}{V}\int d^N\mathbf{x}f(\mathbf{x})(\bar{\rho}\delta(\mathbf{x})). \tag{132}$$

Thus, the inferred density is given by,

$$\tilde{\delta}(\mathbf{x}) = \frac{\bar{\rho}}{\bar{\rho}_{local}} \times \left(\delta(\mathbf{x}) - \frac{1}{V}\int d^N\mathbf{x}f(\mathbf{x})\delta(\mathbf{x})\right) \tag{133}$$

The prefactor will give a small correction in the normalisation of the inferred power spectrum, and we will drop it for simplicity.

Again, the observed field is only measured over a finite region, so the true observable is given by $\tilde{\delta}^{obs}(\mathbf{x}) = \tilde{\delta}(\mathbf{x})f(\mathbf{x})$. In Fourier space, it can be written as

$$\begin{aligned}\tilde{\delta}^{obs}_{\mathbf{k}} &= \frac{V}{(2\pi)^N}\int d^N\mathbf{k}'\delta_{\mathbf{k}'}f_{\mathbf{k}-\mathbf{k}'} - f_{\mathbf{k}}\frac{1}{V}\int d^N\mathbf{x}f(\mathbf{x})\delta(\mathbf{x}) \\ &= \frac{V}{(2\pi)^N}\int d^N\mathbf{k}'\delta_{\mathbf{k}'}\left(f_{\mathbf{k}-\mathbf{k}'} - f_{\mathbf{k}}f_{\mathbf{k}'}\right).\end{aligned} \tag{134}$$

By definition, we are subtracting off the local mean and so the average of this new field should be zero; in Fourier space, this corresponds to $\tilde{\delta}^{obs}_{\mathbf{0}} = 0$, which

follows from above since $f_{\mathbf{0}} = 1$. This is known as the *integral constraint,* and the above equation demonstrates that the observables will not depend on modes much larger than the region which is probed.

Note that the effective power spectrum window is somewhat different from that implied by Peacock (eq. 16.127). There, the effective window for $\langle \tilde{\delta}^{obs}_{\mathbf{k}} \tilde{\delta}^{obs*}_{\mathbf{k}} \rangle$ is $|f_{\mathbf{k}-\mathbf{k}'}|^2 - |f_{\mathbf{k}}|^2 |f_{\mathbf{k}'}|^2$, whereas squaring the aboove expression gives $|f_{\mathbf{k}-\mathbf{k}'} - f_{\mathbf{k}} f_{\mathbf{k}'}|^2$. These do not appear to be equilavent and evaluating them for exponential selection functions they seem to differ by about a factor of two for small k and k', though they both obey the integral constraint. (Perhaps I have erred, or pehaps the Peacock analysis ignores the fact that the mean is not any number, but is itself a function of the density field and so is correlated with it.)

10.5 Wavelets and other transformations

$\delta(\mathbf{x})$ is localised in real space, and $\delta_{\mathbf{k}}$ is localised in Fourier space. We can also consider transformations of the data which are localised in neither, and these are generically called wavelets. One can then perform the same statistical measures on the amplitudes in the wavelet basis.

For example, one can smooth the field with a many Gaussian window functions, centred at different positions and using a range of smoothing lengths, and use these amplitudes to represent the data instead. Ideally the wavelet basis should contain the same information as field, with the same number of amplitudes so there is no redundant information.

In practise wavelets are often chosen on the basis of how fast they are to compute. There are a number of fast wavelet transforms for discrete data which are typically used. While one can easily measure the probability distributions in a wavelet basis, its generally not particularly a good basis for finding analytic predictions. Generally they are used for comparing to simulated data sets.

Wavelets are often a good way of compressing data; one simply keeps those which have an amplitude greater than a particular threshold. This is necessarily a *lossy* compression (some information is lost); but with the right wavelet choice you can ensure that the important information is stored more compactly.

10.6 Measures of non-Gaussianity

Often one wants to know if a particular map is Gaussian. To do this, one chooses some (sometimes arbitrary) statistics of the map and compares them to a collection of Gaussian maps with the same power spectrum. Its best to find a statistic which will be sensitive to the form of non-Gaussianity you might expect to see. However, beware of falling into the trap of defining the statistic based on the map you wish to test; such posteriorly defined statistics are nearly impossible to interpret.

It can be useful to look at topological measures of the field. For example, consider all regions where the field is above (or below) a given threshold. What is the distribution of volumes for the disconnected regions? How are their total surface areas distributed? How many handles (holes) does a typical region have? These measures are known as *Minkowski functionals.* While useful measures, they tend to emphasises small scale features where noise is often dominant.

Other interesting statistics focus on the peaks or minima of the fields. For example, one can make a discrete map of the maxima above a given threshold,

and look at their density and higher order correlation functions.

11 Advanced topics

11.1 Future advanced topics

- Resampling techniques
- Kolmogorov-Smirnov tests, extensions
- Diffusion, Martingales
- Independent component analysis
- Tests of isotropy

11.2 Interesting problems

- The Monty Hall problem

  Monty Hall is a game show host, who offers you the choice of what lies behind door number 1, door number 2 or door number 3; behind one of these is a brand new car, while the other two have consolation prizes. Once you have chosen your door, Monty shows you that one of the other doors has a goat behind it. He then offers you a chance to change your choice to the third door. Do you swap prizes? Does it make a difference? (Assume he always gives you this choice, whether or not you originally chose correctly.)

- Doomsday and Sleeping Beauty problems

  The Doomsday argument says that humanity is near an end. On the basis that you are a typical human observer, you might expect roughly as many humans to exist after you as have existed before you (60 billion.) However, if the population is growing exponentially, e.g. doubling every 30 years, it will reach 60 billion in 100 years time and we have a high probability of being extinct. Obviously exponential growth is not sustainable, but the argument is made that even if population stabilises, humanity only has a few thousand years left.

  The Sleeping Beauty problem also relates to anthropic arguments. On Sunday she sleeps, and on the basis of a coin flip, she is awoken either only on Monday (heads) or on both Monday and Tuesday (tails). When awoken, she is interviewed and then given an amesia-inducing sleep drug and is finally awoken Wednesday to end the experiment. When she is interviewed, she is asked what she views as the probability that the coin came up heads. What should her answer be?

- Simpson's paradox

  Simpson's paradox is an apparent inconsistency where a trend seen in two subgroups of data is reversed when these subgroups are combined. An example taken from Wikipedia: 13 students apply for maths scholarships (8 boys, 5 girls) and 13 students apply for physics scholarships (8 girls

and 5 boys). Of the maths scholarships, 2 are won by boys (2/8 = 25%) and 1 is won by a girl (1/5 = 20%). Of the physics scholarships, 6 are won by girls (6/8 = 75%) and 4 are won by boys (4/5 = 80%.) Despite achieving a higher success rate in both types of scholarships, the boys are less successsful on the whole (6/13 boys win scholarships compared to 7/13 girls.) This occurs because the boys preferentially applied for the kinds of scholarships which proved harder to obtain.

This also can occur with continuous data, where two subsets of data have a particular trend, but the opposite trend is seen when they are combined. Random splits of data sets can sometimes yield the Simpson's paradox, but its more likely to be a true casual effect if the splitting is well motivated.

- Lindley's paradox

  This is an apparent disagreement between Bayesian and frequentist approaches for hypothesis testing, resulting from the fact that the approaches are asking fundamentally different questions. For example, a frequentist test might exclude some hypothesis as unlikely, while a Bayesian approach shows it is preferred to some other, less constraining hypothesis.

- Stein's paradox

  This phenomenon shows that when three or more parameters are estimated together, there exist combined estimators which are more accurate on average than if the parameters are estimated separately; *this is true even when the parameters and the data are completely independent of each other.*

  For example, suppose you estimate n parameters θ_i based on one measurement of each, x_i, where the errors are Gaussian and independent. Stein's paradox says that the estimator which minimises a risk function based on the mean squared error, $\langle \sum_i (\theta_i - W_i)^2 \rangle$ is not simply the vector $W_i = x_i$, even though for individual parameter x_i is the optimal least squared estimator. The James-Stein estimator is one which shrinks the naive estimator towards the origin (or apparently in any random direction!) [I can't say I understand this; I suspect it reflects an implicit prior in the assumed risk function with a preferred point in parameter space. Frequentists are funny like that.]

References

[1] P. R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences,* McGraw-Hill (1969).

[2] G. Casella, R. L. Berger, *Statistical Inference,* Duxbury (1990). An introduction to orthodox statistics.

[3] S. Dodelson, *Modern Cosmology* Academic Press (2003). See particularly chapter 11 on data analysis.

[4] E. T. Jaynes, *Probability Theory– the Logic of Science,* Cambridge 2003. An excellent discussion of the foundations of statistics.

[5] M. G. Kendall, A. Stuart, (A. O'Hagan, J. K. Ord), *Advanced Theory of Statistics,* Arnold Publishers (1998). Classic text, now includes a volume on Bayesian approach.

[6] R. Lupton, *Statistics in Theory and Practice* Princeton University Press (1993). General book written by a barefoot practising astronomer.

[7] D. J. C. MacKay, *Information theory, inference, and learning algorithms* Cambridge (2003). A good how-to book for MCMC, neural networks, data compression.

[8] J. A. Peacock, *Cosmological Physics,* Cambridge (1999).

[9] P. J. E. Peebles, *The Large Scale Structure of the Universe,* Princeton (1980). Defined the subject, but can be a hard read.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes – The Art of Scientific Computing* Cambridge (1992). A good, concise introduction to a number of numerical techniques written by astronomers; includes Fourier and wavelet transforms, Wiener filtering, K-S tests, etc.

[11] D. S. Sivia, *Data Analysis: A Bayesian Tutorial*, Oxford University Press (1996). Short, very clear introduction to the subject.

[12] E. Vanmarke, *Random Fields: Analysis and Synthesis*, MIT Press (1983).

Appendix A: Glossary of some statistical terms

boot-strapping - a resampling technique where a set of simulated data sets are created from an observed data set by use of resampling with replacement. This is done either by random sampling or by exhausting all ways the resampling could be done.

Gibbs' sampling - a Monte Carlo method where the random steps are taken in a single dimension, alternating in turn through every dimension.

independent component analysis - a method of blind source separation; it typically assumes the observations are an unknown linear transformation of a number of independent non-Gaussian sources. The method attempts to find the inverse transformation which maximises the inferred non-Gaussianity of the inferred sources.

jack-knife - a specific resampling technique where properties such as the variance are derived from the ensemble of smaller data sets where one or more points (or regions) of the data are omitted.

Kahunen-Loeve transform or decomposition - effectively another name for principal component analysis, it describes the transformation of data into uncorrelated components.

Kolmogorov-Smirnov test - a method of comparing whether two distributions are consistent by finding the largest value of the difference between their cumulative distribution functions. See Numerical Recipes for a nice discussion of it and extensions.

Martingale series - a random process where the expectation value of the next observation is simply given by the present observation, though possibly in a way that involves past observations.

Markov series - a random process where the next observation depends only on the present observation and not on the past.

non-parametric models - while any statistical description requires some choice of parameters to be quantified, in astronomy this generally refers to a choice which does not assume some particular model; for example, inferring bins of a power spectrum.

robust statistics - a set of statistics which may not be optimal, but is more robust against a small contimation by outliers which might be non-Gaussian or have a large variance. For example, the median may be a more robust estimator of the centre of a distribution than the mean, which could be thrown off by a single large outlier.

Appendix B: Useful Relations of Special Functions

Spherical harmonic relations:

- Definition

$$Y_{lm}(\theta,\phi) = \left(\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}\right)^{1/2} P_l^m(\cos\theta)e^{im\phi}$$

- Orthogonality -

$$\int Y_{\ell m}(\mathbf{n})Y^*_{\ell' m'}(\mathbf{n})d\Omega_{\mathbf{n}} = \delta_{\ell\ell'}\delta_{mm'}$$

- Completeness -

$$\sum_{\ell m} Y_{\ell m}(\mathbf{n})Y^*_{\ell m}(\mathbf{n}') = \delta(\phi-\phi')\delta(\cos\theta-\cos\theta')$$

- Plane wave expansion -

$$e^{i\mathbf{k}\cdot\mathbf{r}} = 4\pi\sum_{\ell m} i^\ell j_\ell(kr)Y_{\ell m}(\widehat{\mathbf{k}})Y^*_{\ell m}(\widehat{\mathbf{r}}) = \sum_\ell (2\ell+1)i^\ell j_\ell(kr)P_\ell(\cos\theta),$$

  where $j_\ell(x)$ are the spherical Bessel functions.

Legendre relations (note $P_l(x) = P_l^0(x)$):

- Legendre equation

$$(1-x^2)\frac{d^2}{dx^2}P_l^m(x) - 2x\frac{d}{dx}P_l^m(x) + \left[l(l+1) - \frac{m^2}{1-x^2}\right]P_l^m(x) = 0$$

- Orthogonality

$$\int_{-1}^{1} P_l^m(x)P_{l'}^m(x)dx = \frac{2}{2l+1}\frac{(l+m)!}{(l-m)!}\delta_{ll'}$$

- Rodrigues' formula

$$P_l^m(x) = \frac{(-1)^m}{2^l l!}(x^2-1)^{m/2}\frac{d^{l+m}}{dx^{l+m}}(x^2-1)^l$$

- Derivative relation

$$(1-x^2)\frac{d}{dx}P_l^m(x) = (l-m+1)P_l^m(x) - (l+1)xP_l^m(x)$$

- Recurrence relation

$$(2l+1)xP_l^m(x) = (l-m+1)P_{l+1}^m(x) + (l+m)P_{l-1}^m(x)$$

- Relation to spherical harmonics

$$(2l+1)P_l(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) = 4\pi\sum_m Y_{lm}(\hat{\mathbf{k}})Y^*_{lm}(\hat{\mathbf{r}})$$

Bessel relations (useful for 2-d relations):

- Bessel equation

$$\frac{d^2}{dx^2}J_m(x) + \frac{1}{x}\frac{d}{dx}J_m(x) + \left(1 - \frac{m^2}{x^2}\right)J_m(x) = 0$$

- Orthogonality

$$\int kdkJ_m(k\rho)J_m(k\rho') = \frac{1}{\rho}\delta(\rho-\rho')$$

- Plane wave expansion

$$e^{ik\rho\cos\phi} = \sum_m i^m e^{im\phi}J_m(k\rho)$$

- Integral representation

$$J_m(x) = \frac{1}{2\pi i^m}\int_0^{2\pi} d\phi e^{ix\cos\phi - im\phi}$$

- Derivative relation

$$\frac{d}{dx}J_m(x) = \frac{1}{2}\left[J_{m-1}(x) - J_{m+1}(x)\right]$$

- Recurrence relation

$$J_{m+1}(x) = \frac{2m}{x}J_m(x) - J_{m-1}(x)$$

- Spherical Bessel functions

$$j_m(x) \equiv \left(\frac{\pi}{2x}\right)^{1/2} J_{m+\frac{1}{2}}(x)$$