Bayesian Statistics in Cosmology

Alan Heavens

Institute for Astronomy, University of Edinburgh, Blackford Hill, Edinburgh EH9 3HJ,
afh@roe.ac.uk
Lectures and workshops at SCMA V, Penn State June 2011

1 Introduction

In these lectures I cover a number of topics in cosmological data analysis. I concentrate on general techniques which are common in cosmology, or techniques which have been developed in a cosmological context. In fact they have very general applicability, for problems in which the data are interpreted in the context of a theoretical model, and thus lend themselves to a Bayesian treatment.

We consider the general problem of estimating parameters from data, and consider how one can use Fisher matrices to analyse survey designs before any data are taken, to see whether the survey will actually do what is required. We outline numerical methods for estimating parameters from data, including Monte Carlo Markov Chains and the Hamiltonian Monte Carlo method. We also look at Model Selection, which covers various scenarios such as whether an extra parameter is preferred by the data, or answering wider questions such as which theoretical framework is favoured, using General Relativity and braneworld gravity as an example. These notes are not a literature review, so there are relatively few references.

After this introduction, the sections are:

- Parameter Estimation
- Fisher Matrix analysis
- Numerical methods for Parameter Estimation
- Model Selection

1.1 Notations:

- *Data* will be called $\mathbf{x}$, or x, or x_i, and is written as a vector, even if it is a 2D image.
- *Model parameters* will be called $\boldsymbol{\theta}$, or θ or θ_α

1.2 Inverse problems

Most data analysis problems are *inverse problems*. You have a set of data $\mathbf{x}$, and you wish to interpret the data in some way. Typical classifications are:

- Hypothesis testing
- Parameter estimation
- Model selection

Cosmological examples of the first type include

- Are CMB data consistent with the hypothesis that the initial fluctuations were gaussian, as predicted (more-or-less) by the simplest inflation theories?
- Are large-scale structure observations consistent with the hypothesis that the Universe is spatially flat?

Cosmological examples of the second type include

- In the Big Bang model, what is the value of the matter density parameter?
- What is the value of the Hubble constant?

Model selection can include slightly different types (but are mostly concerned with larger, often more qualitative questions):

- Do cosmological data favour the Big Bang theory or the Steady State theory?
- Is the gravity law General Relativity or higher-dimensional?
- Is there evidence for a non-flat Universe?

Note that the notion of a *model* can be a completely different paradigm (the first two examples), or basically the same model, but with a different parameter set. In the third example, we are comparing a flat Big Bang model with a non-flat one. The latter has an additional parameter, and is considered to be a different model. Similar problems exist elsewhere in astrophysics, such as how many absorption-line systems are required to account for an observed feature?

These lectures will principally be concerned with questions of the last two types, *parameter estimation* and *model selection*, but will also touch on experimental design and error forecasting. Hypothesis testing can be treated in a similar manner.

2 Parameter estimation

We collect some data, and wish to interpret them in terms of a *model*. A model is a theoretical framework which we assume is true. It will typically have some parameters $\boldsymbol{\theta}$ in it, which you want to determine. The goal of parameter estimation is to provide estimates of the parameters, and their errors, or ideally the whole probability distribution of $\boldsymbol{\theta}$, given the data $\mathbf{x}$. This is called the *posterior probability distribution* i.e. it is the probability that the parameter takes certain values, *after* doing the experiment (as well as assuming a lot of other things):

$$p(\boldsymbol{\theta}|\mathbf{x}). \tag{1}$$

This is an example of RULE 1[1]: Start by thinking about what it is you want to know, and write it down mathematically.

From $p(\boldsymbol{\theta}|\mathbf{x})$ one can calculate the expectation values of the parameters, and their errors. Note that we are immediately taking a Bayesian view of probability, as a *degree of belief*, rather than a frequency of occurrence in a set of trials.

2.1 Forward modelling

Often, what may be easily calculable is not this, rather the opposite, $p(\mathbf{x}|\boldsymbol{\theta})$[2]. The opposite is sometimes referred to as *forward modelling* - i.e. if we know what the parameters are, we can compute the expected distribution of the data. Examples of forward modelling distributions

[1] There is no rule n: $n > 1$.

[2] If you are confused about $p(A|B)$ and $p(B|A)$ consider if A=pregnant and B=female. $p(A|B)$ is a few percent, $p(B|A)$ is unity

include the common ones - binomial, Poisson, gaussian etc. or may be more complex, such as the predictions for the CMB power spectrum as a function of cosmological parameters. As a concrete example, consider a model which is a gaussian with mean μ and variance σ^2. The model has two parameters, $\boldsymbol{\theta} = (\mu, \sigma)$, and the probability of a single variable x given the parameters is

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

but this is not what we actually want. However, we can relate this to $p(\boldsymbol{\theta}|x)$ using Bayes' Theorem, here written for a more general data vector $\mathbf{x}$:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\boldsymbol{\theta}, \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \tag{3}$$

- $p(\boldsymbol{\theta}|\mathbf{x})$ is the *posterior* probability for the parameters.
- $p(\mathbf{x}|\boldsymbol{\theta})$ is called the *Likelihood* and given its own symbol $L(\mathbf{x}; \boldsymbol{\theta})$.
- $p(\boldsymbol{\theta})$ is called the *prior*, and expresses what we know about the parameters prior to the experiment being done. This may be the result of previous experiments, or theory (e.g. some parameters, such as the age of the Universe, may have to be positive). In the absence of any previous information, the prior is often assumed to be a constant (a 'flat prior').
- $p(\mathbf{x})$ is the *evidence*.

For parameter estimation, the evidence simply acts to normalise the probabilities,

$$p(\mathbf{x}) = \int d\boldsymbol{\theta}\, p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{4}$$

and the *relative* probabilities of the parameters do not depend on it, so it is often ignored and not even calculated.

However, the evidence does play an important role in *model selection*, when more than one theoretical model is being considered, and one wants to choose which model is most likely, whatever the parameters are. We turn to this later.

Actually all the probabilities above should be conditional probabilities, given any prior information I which we may have. For clarity, I have omitted these for now. I may be the result of previous experiments, or may be a theoretical prior, in the absence of any data. In such cases, it is common to adopt the *principle of indifference* and assume that all values of the parameter(s) is (are) equally likely, and take $p(\boldsymbol{\theta})$=constant (perhaps within some finite bounds, or if infinite bounds, set it to some arbitrary constant and work with an unnormalised prior). This is referred to as a *flat prior*. Other choices can be justified.

Thus for flat priors, we have simply

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto L(\mathbf{x}; \boldsymbol{\theta}). \tag{5}$$

Although we may have the full probability distribution for the parameters, often one simply uses the peak of the distribution as the estimate of the parameters. This is then a *Maximum Likelihood* estimate. Note that if the priors are not flat, the peak in the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ is not necessarily the maximum likelihood estimate.

A 'rule of thumb' is that if the priors are assigned theoretically, and they influence the result significantly, the data are usually not good enough. (If the priors come from previous experiment, the situation is different - we can be more certain that we really have some prior knowledge in this case).

Finally, note that this method does not generally give a goodness-of-fit, only relative probabilities. It is still common to compute χ^2 at this point to check the fit is sensible.

2.2 Updating the probability distribution for a parameter

One will often see in the literature forecasts for a new survey, where it is assumed that we will know quite a lot about cosmological parameters from another experiment. Typically these days it is *Planck*, which is predicted to constrain many cosmological parameters very accurately. Often people 'include a Planck prior'. What does this mean, and is it justified? Essentially, what is assumed is that by the time of the survey, Planck will have happened, and we can combine results. We can do this in two ways: regard Planck+survey as new data, or regard the survey as the new data, but our prior information has been set by what we know from Planck. If Bayesian statistics makes sense, it should not matter which we choose. We show this now.

If we obtain some more information, from a new experiment, then we can use Bayes' theorem to update our estimate of the probabilities associated with each parameter. The problem reduces to that of showing that adding the results of a new experiment to the probability of the parameters is the same as doing the two experiments first, and then seeing how they both affect the probability of the parameters. In other words it should not matter how we gain our information, the effect on the probability of the parameters should be the same.

We start with Bayes' expression for the posterior probability of a parameter (or more generally of some hypothesis), where we put explicitly that all probabilities are conditional on some prior information I.

$$p(\boldsymbol{\theta}|\mathbf{x}I) = \frac{p(\boldsymbol{\theta}|I)p(\mathbf{x}|\boldsymbol{\theta}I)}{p(\mathbf{x}|I)}. \tag{6}$$

Let say we do a new experiment with new data, $\mathbf{x}'$. We have two ways to analyse the new data:

- Interpretation 1: we regard $\mathbf{x}'$ as the dataset, and $\mathbf{x}I$ (means $\mathbf{x}$ *and* I) as the new prior information.
- Interpretation 2: we put all the data together, and call it $\mathbf{x}'\mathbf{x}$, and interpret it with the old prior information I.

If Bayesian inference is to be consistent, it should not matter which we do.

Let us start with interpretation 1. We rewrite Bayes' theorem, equation (**??**) by changing datasets $\mathbf{x} \to \mathbf{x}'$, and letting the old data become part of the prior information $I \to I' = \mathbf{x}I$. Bayes' theorem is now

$$p(\boldsymbol{\theta}|\mathbf{x}'I') = \frac{p(\boldsymbol{\theta}|\mathbf{x}I)p(\mathbf{x}'|\boldsymbol{\theta}\mathbf{x}I)}{p(\mathbf{x}'|\mathbf{x}I)}. \tag{7}$$

We now notice that the new prior in this expression is just the old posteriori probability from equation (**??**), and that the new likelihood is just

$$p(\mathbf{x}'|\mathbf{x}\boldsymbol{\theta}I) = \frac{p(\mathbf{x}'\mathbf{x}|\boldsymbol{\theta}I)}{p(\mathbf{x}|\boldsymbol{\theta}I)}. \tag{8}$$

Substituting this expression for the new likelihood:

$$p(\boldsymbol{\theta}|\mathbf{x}I') = \frac{p(\boldsymbol{\theta}|\mathbf{x}I)p(\mathbf{x}'\mathbf{x}|\boldsymbol{\theta}I)}{p(\mathbf{x}'|\mathbf{x}I)p(\mathbf{x}|\boldsymbol{\theta}I)}. \tag{9}$$

Using Bayes' theorem again on the first term on the top and the second on the bottom, we find

$$p(\boldsymbol{\theta}|\mathbf{x}I') = \frac{p(\boldsymbol{\theta}|I)p(\mathbf{x}'\mathbf{x}|\boldsymbol{\theta}I)}{p(\mathbf{x}'|\mathbf{x}I)p(\mathbf{x}|I)}, \tag{10}$$

and simplifying the bottom gives finally

$$p(\boldsymbol{\theta}|\mathbf{x}I') = \frac{p(\boldsymbol{\theta}|I)p(\mathbf{x}'\mathbf{x}|\boldsymbol{\theta}I)}{p(\mathbf{x}'\mathbf{x}|I)} = p(\boldsymbol{\theta}|([\mathbf{x}\mathbf{x}']I) \tag{11}$$

which is Bayes' theorem in Interpretation 2. i.e. it has the same form as equation (**??**), the outcome from the initial experiment, but now with the data $\mathbf{x}$ replaced by $\mathbf{x}'\mathbf{x}$. In other words, we have shown that $\mathbf{x} \to \mathbf{x}'$ and $I \to \mathbf{x}I$ is equivalent to $\mathbf{x} \to \mathbf{x}'\mathbf{x}$. This shows us that it doesn't matter how we add in new information. Bayes' theorem gives us a natural way of improving our statistical inferences as our state of knowledge increases.

2.3 Errors

Let us assume we have a posterior probability distribution, which is single-peaked. Two common estimators (indicated by a hat: $\hat{\boldsymbol{\theta}}$) of the parameters are the peak (most probable) values, or the mean,

$$\hat{\boldsymbol{\theta}} = \int d\boldsymbol{\theta}\, \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{x}). \tag{12}$$

An estimator is *unbiased* if its expectation value is the true value $\boldsymbol{\theta}_0$:

$$\langle \hat{\boldsymbol{\theta}} \rangle = \boldsymbol{\theta}_0. \tag{13}$$

Let us assume for now that the prior is flat, so the posterior is proportional to the likelihood. This can be relaxed. Close to the peak, a Taylor expansion of the log likelihood implies that locally it is a mutivariate gaussian *in parameter space*:

$$\ln L(\mathbf{x};\boldsymbol{\theta}) = \ln L(\mathbf{x};\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_{0\alpha})\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}_\alpha \partial \boldsymbol{\theta}_\beta}(\boldsymbol{\theta}_\beta - \boldsymbol{\theta}_{0\beta}) + \dots \tag{14}$$

or

$$L(\mathbf{x};\boldsymbol{\theta}) = L(\mathbf{x};\boldsymbol{\theta}_0) \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_{0\alpha})H_{\alpha\beta}(\boldsymbol{\theta}_\beta - \boldsymbol{\theta}_{0\beta})\right]. \tag{15}$$

The Hessian matrix $\mathsf{H}_{\alpha\beta} \equiv -\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta}$ controls whether the estimates of θ_α and θ_β are correlated or not. If it is diagonal, the estimates are uncorrelated. Note that this is a statement about *estimates* of the quantities, not the quantities themselves, which may be entirely independent, but if they have a similar effect on the data, their estimates may be correlated. Note that in cases of practical interest, the likelihood may not be well described by a multivariate gaussian at levels which set the interesting credibility levels (e.g. 68%). We turn later to how to proceed in such cases.

2.4 Conditional and marginal errors

If we fix all the parameters except one, then the error is given by the curvature along a line through the likelihood (posterior, if prior is not flat):

$$\sigma_{\mathrm{conditional},\alpha} = \frac{1}{\sqrt{\mathsf{H}_{\alpha\alpha}}}. \tag{16}$$

This is called the *conditional error*, and is the minimum error bar attainable on θ_α if all the other parameters are known. *It is rarely relevant and should almost never be quoted.*

2.5 Marginalising over a gaussian likelihood

The marginal distribution of $\boldsymbol{\theta}_1$ is obtained by integrating over the other parameters:

$$p(\boldsymbol{\theta}_1) = \int d\boldsymbol{\theta}_2 \dots d\boldsymbol{\theta}_N p(\boldsymbol{\theta}) \tag{17}$$

a process which is called *marginalisation*. Often one sees marginal distributions of all parameters in pairs, as a way to present some complex results. In this case two variables are left out of the integration.

If you plot such error ellipses, you *must* say what contours you plot. If you say you plot 1σ and 2σ contours, I don't know whether this is for the joint distribution (i.e. 68% of the probability lies within the inner contour), or whether 68% of the probability of a single parameter lies within the bounds projected onto a parameter axis. The latter is a 1σ, single-parameter error contour (and corresponds to $\Delta\chi^2 = 1$), whereas the former is a 1σ contour for the joint distribution, and corresponds to $\Delta\chi^2 = 2.3$.

Note that $\Delta\chi^2 = \chi^2 - \chi^2(minimum)$, where

$$\chi^2 = \sum_i \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{18}$$

for data x_i with $\mu_i = \langle x_i \rangle$ and variance σ_i^2. If the data are correlated, this generalises to

$$\chi^2 = \sum_{ij} (x_i - \mu_i) C_{ij}^{-1} (x_j - \mu_j) \tag{19}$$

where $C_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$.

For other dimensions, see Table 1, or read...

The Numerical Recipes bible, chapter 15.6 [?]

Read it. Then, when you need to plot some error contours, read it again.

Table 1. $\Delta\chi^2$ for joint parameter estimation for 1, 2 and 3 parameters.

σ	p	M=1	M=2	M=3
1σ	68.3%	1.00	2.30	3.53
2σ	95.4%	4.00	6.17	8.02
3σ	99.73%	9.00	11.8	14.2

Note that some of the results I give assume the likelihood (or posterior) is well-approximated by a multivariate gaussian. This may not be so. If your posterior is a single peak, but is not well-approximated by a multivariate gaussian, label your contours with the enclosed probability. If the likelihood is complicated (e.g. multimodal), then you may have to plot it and leave it at that - reducing it to a maximum likelihood point and error matrix is not very helpful. Not that in this case, the mean of the posterior may be unhelpful - it may lie in a region of parameter space with a very small posterior.

A multivariate gaussian likelihood is a common assumption, so it is useful to compute marginal errors for this rather general situation. The simple result is that the marginal error on parameter θ_α is

$$\sigma_\alpha = \sqrt{(\mathsf{H}^{-1})_{\alpha\alpha}}. \tag{20}$$

Note that we invert the Hessian matrix, and then take the square root of the diagonal components. Let us prove this important result. In practice it is often used to estimate errors for a future experiment, where we deal with the expectation value of the Hessian, called the *Fisher Matrix*:

$$\mathsf{F}_{\alpha\beta} \equiv \langle \mathsf{H}_{\alpha\beta} \rangle = \left\langle -\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}_\alpha \partial \boldsymbol{\theta}_\beta} \right\rangle. \tag{21}$$

We will have much more to say about Fisher matrices later. The expected error on θ_α (marginalising over other variables) is thus

$$\sigma_\alpha = \sqrt{(\mathsf{F}^{-1})_{\alpha\alpha}}. \tag{22}$$

It is always at least as large as the expected conditional error. Note: this result applies for gaussian-shaped likelihoods, and is useful for experimental design. For real data, you would do the marginalisation a different way - see later.

To prove this, we will use characteristic functions.

Characteristic functions

In probability theory the Fourier Transform of a probability distribution function is known as the *characteristic function*. For a multivariate distribution with N parameters, it is defined by

$$\phi(\mathbf{k}) = \int d^N\boldsymbol{\theta}\, p(\boldsymbol{\theta}) e^{-i\mathbf{k}\cdot\boldsymbol{\theta}} \tag{23}$$

with reciprocal relation

$$p(\boldsymbol{\theta}) = \int \frac{d^N\mathbf{k}}{(2\pi)^N}\, \phi(\mathbf{k}) e^{i\mathbf{k}\cdot\boldsymbol{\theta}} \tag{24}$$

(note the choice of where to put the factors of 2π is not universal). Hence the characteristic function is also the expectation value of $e^{-i\mathbf{k}.\boldsymbol{\theta}}$:

$$\phi(\mathbf{k}) = \langle e^{-i\mathbf{k}.\boldsymbol{\theta}} \rangle. \tag{25}$$

Part of the power of characteristic functions is the ease with which one can generate all of the moments of the distribution by differentiation:

$$\langle \boldsymbol{\theta}_\alpha^{n_\alpha} \dots \boldsymbol{\theta}_\beta^{n_\beta} \rangle = \left[\frac{\partial^{n_\alpha + \dots + n_\beta} \phi(\mathbf{k})}{\partial(-i\mathbf{k}_\alpha)^{n_\alpha} \dots \partial(-i\mathbf{k}_\beta)^{n_\beta}} \right]_{\mathbf{k}=\mathbf{0}}. \tag{26}$$

This can be seen if one expands $\phi(\mathbf{k})$ in a power series, using

$$\exp(\alpha) = \sum_{i=0}^{\infty} \frac{\alpha^n}{n!}, \tag{27}$$

giving

$$\phi(\mathbf{k}) = 1 - i\mathbf{k}\cdot\langle\boldsymbol{\theta}\rangle - \frac{1}{2}\sum_{\alpha\beta} \mathbf{k}_\alpha \mathbf{k}_\beta \langle \boldsymbol{\theta}_\alpha \boldsymbol{\theta}_\beta \rangle + \dots. \tag{28}$$

Hence for example we can compute the mean

$$\langle \boldsymbol{\theta}_\alpha \rangle = \left[\frac{\partial \phi(\mathbf{k})}{\partial(-i\mathbf{k}_\alpha)} \right]_{\mathbf{k}=\mathbf{0}} \tag{29}$$

and the covariances, from

$$\langle \boldsymbol{\theta}_\alpha \boldsymbol{\theta}_\beta \rangle = \left[\frac{\partial^2 \phi(\mathbf{k})}{\partial(-i\mathbf{k}_\alpha)\partial(-i\mathbf{k}_\beta)} \right]_{\mathbf{k}=\mathbf{0}}. \tag{30}$$

(Putting $\alpha = \beta$ yields the variance of $\boldsymbol{\theta}_\alpha$ after subtracting the square of the mean).

2.6 The expected marginal error on $\boldsymbol{\theta}_\alpha$ is $\sqrt{(\mathbf{F}^{-1})_{\alpha\alpha}}$

The likelihood is here assumed to be a multivariate gaussian, with expected hessian given by the Fisher matrix. Thus (suppressing ensemble averages)

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{M/2}\sqrt{\det \mathsf{F}}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathsf{F} \boldsymbol{\theta} \right), \tag{31}$$

where T indicates transpose, and for simplicity I have assumed the parameters have zero mean (if not, just redefine $\boldsymbol{\theta}$ as the difference between $\boldsymbol{\theta}$ and the mean). We proceed by diagonalising the quadratic, then computing the characteristic function, and compute the covariances using equation (**??**). This is achieved in the standard way by rotating the parameter axes:

$$\Psi = \mathsf{R}\boldsymbol{\theta} \tag{32}$$

for a matrix R. Since F is real and symmetric, R is orthogonal, $\mathsf{R}^{-1} = \mathsf{R}^T$. Diagonalising gives

$$\boldsymbol{\theta}^T \mathsf{F} \boldsymbol{\theta} = \Psi^T \mathsf{R} \mathsf{F} \mathsf{R}^T \Psi, \tag{33}$$

and the diagonal matrix composed of the eigenvalues of F

$$\boldsymbol{\Lambda} = \mathsf{R} \mathsf{F} \mathsf{R}^T, \tag{34}$$

Note that the eigenvalues of F are positive, as F must be positive-definite.

The characteristic function is

$$\phi(\mathbf{k}) = \frac{1}{(2\pi)^{M/2}\sqrt{\det \mathsf{F}}} \int d^M \Psi \exp\left(-\frac{1}{2}\Psi^T \boldsymbol{\Lambda} \mathsf{F} \Psi\right) \exp(-i\mathbf{k}^T \mathsf{R}^T \Psi) \tag{35}$$

where we exploit the fact that the rotation has unit Jacobian to change $d^M\boldsymbol{\theta}$ to $d^M\Psi$. If we define $\mathbf{K} \equiv \mathsf{R}\mathbf{k}$,

$$\phi(\mathbf{k}) = \frac{1}{(2\pi)^{M/2}\sqrt{\det \mathsf{F}}} \int d^M \Psi \exp\left(-\frac{1}{2}\Psi^T \boldsymbol{\Lambda} \Psi\right) \exp(-i\mathbf{K}^T \Psi) \tag{36}$$

and since $\boldsymbol{\Lambda}$ is diagonal, the first exponential is a sum of squares, which we can integrate separately, using

$$\int_{-\infty}^{\infty} d\psi \exp(-\Lambda\psi^2/2) \exp(-iK\psi) = \sqrt{2\pi/\Lambda} \exp[-K^2/(2\Lambda)]. \tag{37}$$

All multiplicative factors cancel (since the rotation preserves the eigenvalues, so $\det(\mathsf{F}) = \prod \boldsymbol{\Lambda}_\alpha$), and we obtain

$$\phi(\mathbf{k}) = \exp\left(-\sum_i K_i^2/(2\Lambda_i)\right) = \exp\left(-\frac{1}{2}\mathbf{K}^T \boldsymbol{\Lambda}^{-1} \mathbf{K}\right) = \exp\left(-\frac{1}{2}\mathbf{k}^T \mathsf{F}^{-1} \mathbf{k}\right) \tag{38}$$

where the last result follows from $\mathbf{K}^T \boldsymbol{\Lambda}^{-1} \mathbf{K} = \mathbf{k}^T (\mathsf{R}^T \boldsymbol{\Lambda}^{-1} \mathsf{R}) \mathbf{k} = \mathbf{k}^T \mathsf{F}^{-1} \mathbf{k}$.

Having obtained the characteristic function, the result (**??**) follows immediately from equation (**??**).

2.7 Marginalising over 'amplitude' variables

It is not uncommon to want to marginalise over a nuisance parameter which is a simple scaling variable. Examples include a calibration uncertainty, or perhaps an unknown gain in an electronic detector. This can be done analytically for a gaussian data space:

$$L(\boldsymbol{\theta}; \mathbf{x}) = \frac{1}{(2\pi)^{N/2}\sqrt{\det \mathsf{C}}} \exp\left[-\frac{1}{2}\sum_{ij} (\mathbf{x}_i - \mathbf{x}_i^{th}) \mathsf{C}_{ij}^{-1} (\mathbf{x}_j - \mathbf{x}_j^{th})\right] \tag{39}$$

where we have N data points with covariance matrix $\mathsf{C}_{ij} \equiv \langle (\mathbf{x}_i - \mathbf{x}_i^{th})(\mathbf{x}_j - \mathbf{x}_j^{th}) \rangle$, and the model data values $\mathbf{x}_i^{th}$ and C depend on the parameters. Let us now assume that the covariance matrix does not depend on the parameters, so the parameter dependence is only through $\mathbf{x}^{th}(\boldsymbol{\theta})$, and further we assume that one of the parameters simply scales the theoretical signal. i.e. $\mathbf{x}^{th}(\boldsymbol{\theta}) = A\mathbf{x}^{th}(\boldsymbol{\theta}|A=1)$. If we want the likelihood of all the other parameters (call this $L'(\boldsymbol{\theta}'; \mathbf{x})$, with one fewer parameter), marginalised over the unknown A, then we can integrate:

$$\begin{aligned} L(\boldsymbol{\theta}'; \mathbf{x}) &= \int dA L(\boldsymbol{\theta}; \mathbf{x}) p(A) \\ &= \frac{1}{(2\pi)^{N/2}\sqrt{\det \mathsf{C}}} \int dA \exp\left[-\frac{1}{2}\sum_{ij} (\mathbf{x}_i - A\mathbf{x}_i^{th}) \mathsf{C}_{ij}^{-1} (\mathbf{x}_j - A\mathbf{x}_j^{th})\right] p(A) \end{aligned} \tag{40}$$

where $p(A)$ is the prior for A. If we take a uniform (unnormalised!) prior on A between limits $\pm\infty$, and the theoretical $\mathbf{x}$ are now at $A = 1$, then we can integrate the quadratic. You can do this.

Exercise: do this.

3 Fisher Matrix Analysis

This has been adapted from [**?**] (hereafter TTH).

How accurately can we estimate model parameters from a given data set? This question was basically answered 60 years ago [**?**], and we will now summarize the results, which are both simple and useful.

Suppose for definiteness that our data set consists of N real numbers x_1, $x_2, ..., x_N$, which we arrange in an N-dimensional vector $\mathbf{x}$. These numbers could for instance denote the measured temperatures in the N pixels of a CMB sky map, the counts-in-cells of a galaxy redshift survey, N coefficients of a Fourier expansion of an observed galaxy density field, or the number of gamma-ray bursts observed in N different flux bins. Before collecting the data, we think of $\mathbf{x}$ as a random variable with some probability distribution $L(\mathbf{x};\boldsymbol{\theta})$, which depends in some known way on a vector of M model parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_M)$.

Such model parameters might for instance be the spectral index of density fluctuations, the Hubble constant h, the cosmic density parameter Ω or the mean redshift of gamma-ray bursts. We will let $\boldsymbol{\theta}_0$ denote the true parameter values and let $\boldsymbol{\theta}$ refer to our estimate of $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ is some function of the data vector $\mathbf{x}$, it too is a random variable. For it to be a good estimate, we would of course like it to be unbiased, *i.e.*,

$$\langle \boldsymbol{\theta} \rangle = \boldsymbol{\theta}_0, \tag{41}$$

and give as small error bars as possible, *i.e.*, minimize the standard deviations

$$\Delta\theta_\alpha \equiv \left(\langle \theta_\alpha^2 \rangle - \langle \theta_\alpha \rangle^2 \right)^{1/2}. \tag{42}$$

In statistics jargon, we want the BUE θ_α, which stands for the "Best Unbiased Estimator".

A key quantity in this context is the so-called *Fisher information matrix*, defined as

$$\mathsf{F}_{\alpha\beta} \equiv \left\langle \frac{\partial^2 \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle, \tag{43}$$

where

$$\mathcal{L} \equiv -\ln L. \tag{44}$$

Another key quantity is the *maximum likelihood estimator*, or *ML-estimator* for brevity, defined as the parameter vector $\boldsymbol{\theta}_{\rm ML}$ that maximizes the likelihood function $L(\mathbf{x};\boldsymbol{\theta})$.

Using this notation, a number of powerful theorems have been proven (see *e.g.* [**?**],[**?**]):

1. For any unbiased estimator, $\Delta\theta_\alpha \geq 1/\sqrt{\mathsf{F}_{\alpha\alpha}}$ (the *Cramér-Rao* inequality).
2. If an unbiased estimator attaining ("saturating") the Cramér-Rao bound exists, it is the ML estimator (or a function thereof).
3. The ML-estimator is asymptotically BUE.

The first of these theorems thus places a firm lower limit on the error bars that one can attain, regardless of which method one is using to estimate the parameters from the data. You won't do better, but you might do worse.

The normal case is that the other parameters are estimated from the data as well, in which case, as we have seen, the minimum standard deviation rises to

$$\Delta\theta_\alpha \geq (\mathsf{F}^{-1})^{1/2}_{\alpha\alpha}. \tag{45}$$

This is called the *marginal error*, and I reemphasise that this is normally the relevant error to quote.

The second theorem shows that maximum-likelihood (ML) estimates have quite a special status: if there is a best method, then the ML-method is the one. Finally, the third result

basically tells us that in the limit of a very large data set, the ML-estimate for all practical purposes is the best estimate, the one that for which the Cramér-Rao inequality becomes an equality[3]. It is these nice properties that have made ML-estimators so popular.

Note that conditional and marginal errors coincide if F is diagonal. If it is not, then the *estimates* of the parameters are correlated (even if the parameters themselves are uncorrelated). e.g. in the example shown in Fig. 1, estimates of the baryon density parameter Ω_b and the dark energy equation of state $w \equiv p/\rho c^2$ are strongly correlated with WMAP data alone, so we will tend to overestimate both Ω_b and w, or underestimate both. However, the value of Ω_b in the Universe has nothing obvious to do with w - they are independent.

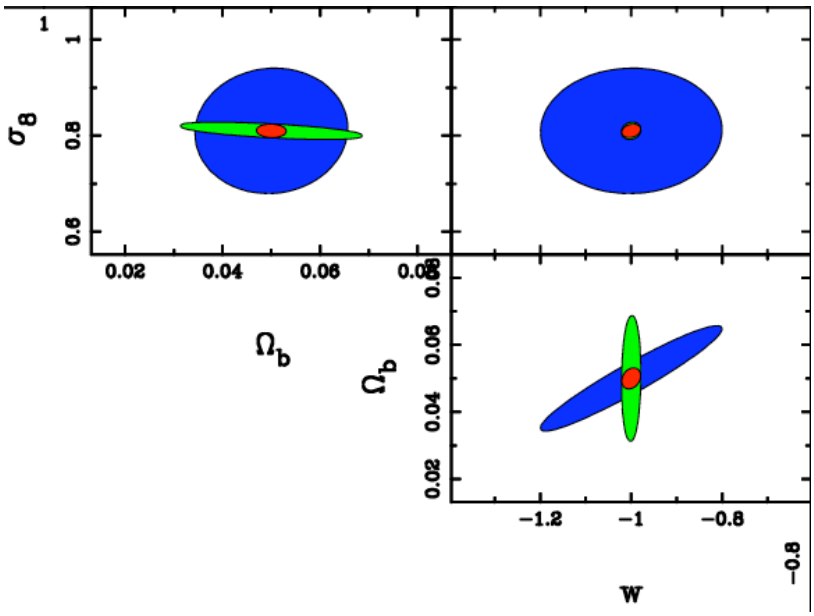

Fig. 1. The expected error ellipses for cosmological parameters (σ_8, baryon density parameter Ω_b, and dark energy equation of state $w \equiv p/\rho c^2$) from a 3D weak lensing survey of 1000 square degrees, with a median redshift of 1 and a photometric redshift error of 0.15. Probabilities are marginalised over all other parameters, except that $n = 1$ and a flat Universe are assumed. Dark ellipses represent a prior from WMAP, pale represents the 3D lensing survey alone, and the central ellipses show the combination (from Kitching T., priv. comm.).

3.1 Cramér-Rao inequality: proof for simple case

This discussion follows closely Andrew Hamilton's lectures to the Valencia Summer School in 2005 [**?**].

At the root of the Cramér-Rao inequality is the Schwarz inequality. If we have estimators for two parameters θ_1 and θ_2, then it is clear that the expectation value of

$$\left\langle \left(\Delta\hat{\theta}_1 + \lambda \Delta\hat{\theta}_2 \right)^2 \right\rangle \geq 0 \tag{46}$$

where $\Delta\hat{\theta}_\alpha \equiv \hat{\theta}_\alpha - \theta_{0\alpha}$ is the difference between the estimate and the true value. Evidently equation **??** holds for any λ. It is easy to show that the left-hand-side is a minimum if we choose $\lambda = -\langle \Delta\hat{\theta}_1 \Delta\hat{\theta}_2 \rangle / \langle (\Delta\hat{\theta}_2)^2 \rangle$, from which we obtain the *Schwarz inequality*:

$$\left\langle \left(\Delta\hat{\theta}_1 \right)^2 \right\rangle \left\langle \left(\Delta\hat{\theta}_2 \right)^2 \right\rangle \geq \langle \Delta\hat{\theta}_1 \, \Delta\hat{\theta}_2 \rangle^2 \tag{47}$$

[3] This is sometimes called 'saturating the Cramér-Rao bound'

Let us treat the simple case of one parameter, and data x. An unbiased estimator $\hat{\theta}$ of a parameter whose true value is θ_0 is the value θ which satisfies

$$\langle \theta - \theta_0 \rangle = \int (\theta - \theta_0) L(x;\theta) dx = 0. \tag{48}$$

Differentiating with respect to θ we get

$$\int (\theta - \theta_0) \frac{\partial L}{\partial \theta} dx + \int L(x;\theta) dx = 0. \tag{49}$$

The last integral is unity, and we can therefore write

$$\left\langle (\theta - \theta_0) \frac{\partial \ln L}{\partial \theta} \right\rangle = -1, \tag{50}$$

and the Schwarz inequality gives

$$\left\langle (\theta - \theta_0)^2 \right\rangle \geq \frac{1}{\left\langle \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right\rangle} \tag{51}$$

The final expression is obtained by differentiating $\int L(x;\theta) dx = 1$ twice with respect to θ to show

$$0 = \frac{\partial}{\partial \theta} \int \frac{\partial \ln L}{\partial \theta} L dx = \int \left[\frac{\partial^2 \ln L}{\partial \theta^2} + \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right] L(x;\theta) dx \tag{52}$$

so we obtain the *Cramér-Rao inequality*

$$\left\langle (\theta - \theta_0)^2 \right\rangle \geq -\frac{1}{\left\langle \frac{\partial^2 \ln L}{\partial \theta^2} \right\rangle} = \frac{1}{F_{\theta\theta}}. \tag{53}$$

Note that for a single variable the conditional error is the same as the marginal error - the Fisher 'matrix' has rank 1.

Combining experiments

If the experiments are independent, you can simply add the Fisher matrices (why?). Note that the marginal error ellipses (marginalising over all but two variables) in the combined dataset can be much smaller than you might expect, given the marginal error ellipses for the individual experiments, because the operations of adding the experimental data and marginalising do not commute.

3.2 The Gaussian Case

Let us now explicitly compute the Fisher information matrix for the case when the probability distribution is Gaussian, *i.e.*, where (dropping an irrelevant additive constant $N \ln[2\pi]$)

$$2\mathcal{L} = \ln \det \mathsf{C} + (\mathbf{x} - \boldsymbol{\mu}) \mathsf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu})^T, \tag{54}$$

where in general both the mean vector $\boldsymbol{\mu}$ and the covariance matrix

$$\mathsf{C} = \langle (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \rangle \tag{55}$$

depend on the model parameters $\boldsymbol{\theta}$. Although vastly simpler than the most general situation, the Gaussian case is nonetheless general enough to be applicable to a wide variety of problems in cosmology. Defining the data matrix

$$\mathsf{D} \equiv (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \tag{56}$$

and using the matrix identity (see exercises) $\ln\det \mathsf{C} = \mathrm{Tr}\ln \mathsf{C}$, where Tr indicates trace, we can re-write (**??**) as

$$2\mathcal{L} = \mathrm{Tr}\left[\ln \mathsf{C} + \mathsf{C}^{-1}\mathsf{D}\right]. \tag{57}$$

We will use the standard comma notation for derivatives, where for instance

$$\mathsf{C}_{,\alpha} \equiv \frac{\partial}{\partial\theta_\alpha}\mathsf{C}. \tag{58}$$

Since C is a symmetric matrix for all values of the parameters, it is easy to see that all the derivatives $\mathsf{C}_{,\alpha}$, $\mathsf{C}_{,\alpha\beta}$, will also be symmetric matrices. Using the matrix identities $(\mathsf{C}^{-1})_{,\alpha} = -\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}$ and $(\ln \mathsf{C})_{,\alpha} = \mathsf{C}^{-1}\mathsf{C}_{,\alpha}$ (see exercises), we find

$$2\mathcal{L}_{,\alpha} = \mathrm{Tr}\left[\mathsf{C}^{-1}\mathsf{C}_{,\alpha} - \mathsf{C}^{-1}\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}\mathsf{D} + \mathsf{C}^{-1}\mathsf{D}_{,\alpha}\right]. \tag{59}$$

When evaluating C and $\boldsymbol{\mu}$ at the true parameter values, we have $\langle \mathbf{x}\rangle = \boldsymbol{\mu}$ and $\langle \mathbf{x}\mathbf{x}^T\rangle = \mathsf{C} + \boldsymbol{\mu}\boldsymbol{\mu}^T$, which gives

$$\begin{cases} \langle \mathsf{D}\rangle & = \mathsf{C}, \\ \langle \mathsf{D}_{,\alpha}\rangle & = 0, \\ \langle \mathsf{D}_{,\alpha\beta}\rangle & = \boldsymbol{\mu}_{,\alpha}\,\boldsymbol{\mu}_{,\beta}^T + \boldsymbol{\mu}_{,\beta}\,\boldsymbol{\mu}_{,\alpha}^T. \end{cases} \tag{60}$$

Using this and equation (**??**), we obtain $\langle \mathcal{L}_{,\alpha}\rangle = 0$. In other words, the ML-estimate is correct on average in the sense that the average slope of the likelihood function is zero at the point corresponding to the true parameter values. Applying the chain rule to equation (**??**), we obtain

$$\begin{aligned} 2\mathcal{L}_{,\alpha\beta} = \mathrm{Tr}[& -\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}\mathsf{C}_{,\beta} + \mathsf{C}^{-1}\mathsf{C}_{,\alpha\beta} \\ & + \mathsf{C}^{-1}(\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}\mathsf{C}_{,\beta} + \mathsf{C}_{,\beta}\,\mathsf{C}^{-1}\mathsf{C}_{,\alpha})\mathsf{C}^{-1}\mathsf{D} \\ & - \mathsf{C}^{-1}(\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}\mathsf{D}_{,\beta} + \mathsf{C}_{,\beta}\,\mathsf{C}^{-1}\mathsf{D}_{,\alpha}) \\ & - \mathsf{C}^{-1}\mathsf{C}_{,\alpha\beta}\,\mathsf{C}^{-1}\mathsf{D} + \mathsf{C}^{-1}\mathsf{D}_{,\alpha\beta}]. \end{aligned} \tag{61}$$

Substituting this and equation (**??**) into equation (**??**) and using the trace identity $\mathrm{Tr}[\mathsf{AB}] = \mathrm{Tr}[\mathsf{BA}]$, many terms drop out and the Fisher information matrix reduces to simply

$$\mathsf{F}_{\alpha\beta} = \langle \mathcal{L}_{,\alpha\beta}\rangle = \frac{1}{2}\mathrm{Tr}[\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\mathsf{C}^{-1}\mathsf{C}_{,\beta} + \mathsf{C}^{-1}\mathsf{M}_{\alpha\beta}], \tag{62}$$

where we have defined the matrix $\mathsf{M}_{\alpha\beta} \equiv \langle \mathsf{D}_{,\alpha\beta}\rangle = \boldsymbol{\mu}_{,\alpha}\,\boldsymbol{\mu}_{,\beta}^T + \boldsymbol{\mu}_{,\beta}\,\boldsymbol{\mu}_{,\alpha}^T$.

The Fisher matrix requires no data.

This result is extremely powerful. If the data have a (multivariate) gaussian distribution (and the errors can be correlated; C need not be diagonal), and you know how the means $\boldsymbol{\mu}$ and the covariance matrix C depend on the parameters, you can calculate the Fisher Matrix *before you do the experiment*. The Fisher Matrix gives you the expected errors, so you know how well you can expect to do if you do a particular experiment, and you can then design an experiment to give you, for example, the best (marginal) error on the parameter you are most interested in.

Note that if the prior is not uniform, then you can simply add a 'prior matrix' to the Fisher matrix before inversion. Fig. **??** shows an example, where a prior from CMB experimental results has been added to a hypothetical 3D weak lensing survey.

Treat the Fisher errors as a one-way test: you might not achieve errors which are this small, but you won't do better. So if you want to measure some quantity with an accuracy of a metre, and a Fisher analysis tells you the error bar is the size of Belgium, give up.

Reparametrisation

Finally we mention that sometimes the gaussian approximation for the likelihood surface is not a very good approximation. With a good theoretical model, it is possible to make nonlinear transformations of the parameters to make the likelihood more gaussian. See [**?**] for more details.

iCosmo: a great resource

icosmo.org has a web-based calculator for Fisher matrices for cosmology. You can download results, or even the source code, to compute Fisher matrices for various experiments in lensing, BAOs, supernovae etc, or custom-design your own survey. It also gives many other useful things, such as lensing power spectra, angular diameter distances etc.

4 Numerical methods

If the problem has only two or three parameters, then it may be possible to evaluate the likelihood on a sufficiently fine grid to be able to locate the peak and estimate the errors. If the dimensionality of the parameter space is very large, then, as the number of grid points grows exponentially with dimension, it becomes rapidly unfeasible to do it this way. In fact, it's very inefficient to do this anyway, as typically most of the hypervolume has very small likelihood so is of little interest. There are various ways to sample the likelihood surface more efficiently, concentrating the points more densely where the likelihood is high.

The most common method in use is Monte Carlo Markov Chain (MCMC). We will also cover here a relatively new method (to cosmology), Hamiltonian Monte Carlo, which seems more efficient in cases studied.

4.1 Monte Carlo Markov Chain (MCMC) method

The aim of MCMC is to generate a set of points in the parameter space whose distribution function is the same as the *target density*, in this case the likelihood, or more generally the posterior. MCMC makes random drawings, by moving in parameter space in a Markov process - i.e. the next sample depends on the present one, but not on previous ones. By design, the resulting Markov Chain of points samples the posterior, such that the density of points is proportional to the target density (at least asymptotically), so we can estimate all the usual quantities of interest from it (mean, variance, etc). The number of points required to get good estimates is said to scale linearly with the number of parameters, so very quickly becomes much faster than grids as the dimensionality increases. In cosmology, we are often dealing with around 10-20 parameters, so MCMC has been found to be a very effective tool.

The target density is approximated by a set of delta functions (you may need to normalise)

$$p(\boldsymbol{\theta}) \simeq \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{\theta}-\boldsymbol{\theta}_i) \tag{63}$$

from which we can estimate any integrals (such as the mean, variance etc.):

$$\langle f(\boldsymbol{\theta})\rangle \simeq \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i). \tag{64}$$

The basic procedure to make the chain is to generate a new point $\boldsymbol{\theta}^*$ from the present point $\boldsymbol{\theta}$ (by taking some sort of step), and accepting it as a new point in the chain with a probability which depends on the ratio of the new and old target densities. The distribution of steps is called the *proposal distribution*. The most popular algorithm is the *Metropolis-Hastings* algorithm, where the probability of acceptance is

$$p(\text{acceptance}) = min\left[1, \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}{p(\boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}\right] \tag{65}$$

where the proposal distribution function is $q(\boldsymbol{\theta} * |\boldsymbol{\theta})$ for a move from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$.

If the proposal distribution is symmetric (as is often the case), the algorithm simplifies to the *Metropolis algorithm*:

$$p(\text{acceptance}) = min\left[1, \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})}\right]. \quad (66)$$

- Choose a random initial starting point in parameter space, and compute the target density.
- Repeat:
- Generate a step in parameter space from a *proposal distribution*, generating a new trial point for the chain.
- Compute the target density at the new point, and accept it (or not) with the Metropolis-Hastings algorithm.
- If the point is not accepted, *the previous point is repeated in the chain*[4].
- End Repeat:

The easy bits: this is trivial to code - you might just take a top-hat proposal distribution in each parameter direction, and it should work. The harder parts are (even in the tophat case): choosing an efficient proposal distribution; dealing with *burn-in* and *convergence*.

Proposal distribution

If the proposal distribution is small, in the sense that the typical jump is small, then the chain may take a very long time to explore the target distribution, and it will be very inefficient. Since the target density hardly changes, almost all points are accepted, but it still takes forever. This is an example of poor *mixing*. If the proposal distribution is too large, on the other hand, then the parameter space is explored, but the trial points are often a long way from the peak, at places where the target density is low. This is very inefficient as well, since they are almost always rejected by the Metropolis-Hastings algorithm. So what is best? You might expect the chain to do well if the proposal distribution is 'about the same size as the peak in the target density', and you would be right. In fact, a very good option is to draw from a multivariate gaussian with the Fisher matrix as Hessian. However, having said that, if you want something quick to code (but sub-optimal), a top-hat of about the right dimensions will do a decent job. You might have to run a preliminary chain or two to get an idea of the size of the target, but the rules say you are not allowed to change the proposal distribution in a chain, so once you have decided you must throw away your chains and begin again.

Burn-in and convergence

Theory indicates that the chain should fairly sample the target distribution once it has converged to a stationary distribution. This means that the early part of the chain (the 'burn-in' are ignored, and the dependence on the starting point is lost. *It is vitally important to have a convergence test.* Be warned that the points in a MCMC chain are correlated, and the chain can appear to have converged when it has not (one can reduce this problem by evaluating the correlation function of the points, and 'thinning' them by (say) taking only every third point (or whatever is suggested by the correlation analysis). Fig. **??** shows one such example. The classic test is the *Gelman-Rubin* (1992) convergence criterion. Start M chains, each with $2N$ points, starting at well-separated parts of parameter space. In this example, the first N are discarded as burn-in.

The idea is that you have two ways to estimate the mean of the parameters - either treat the combined chains as a single dataset, or look at the means of each chain. If the chains have converged, these should agree within some tolerance.

[4] It is a common mistake to neglect to do this

Following [**?**], let θ_i^J represent the point in parameter space in position i of chain J. Each chain has N points. Compute the mean of each chain (J):

$$\bar{\theta}^J \equiv \frac{1}{N}\sum_{i=1}^{N}\theta_i^J \tag{67}$$

and the mean of all the chains

$$\bar{\theta} \equiv \frac{1}{NM}\sum_{i=1}^{N}\sum_{J=1}^{M}\theta_i^J. \tag{68}$$

The variance of the means of each chain (the chain-to-chain variance) B/N (the N appears in Gelman-Rubin for convenience, it seems) is

$$\frac{B}{N} = \frac{1}{(M-1)}\sum_{J=1}^{M}(\bar{\theta}^J - \bar{\theta})^2 \tag{69}$$

and the average variance of each chain is

$$W = \frac{1}{M(N-1)}\sum_{i=1}^{N}\sum_{J=1}^{M}(\theta_i^J - \bar{\theta}^J)^2. \tag{70}$$

Under convergence, W and B/N should agree.

We consider the weighted estimate of the variance,

$$\sigma^2 = \frac{N-1}{N}W + \frac{B}{N}. \tag{71}$$

In the limit of large N this is an unbiased estimate of the variance of the target distribution (W and B should both tend to this). σ^2 overestimates the true variance if the starting distribution is overdispersed (i.e. starting points are more widely spread than the width of the distribution).

Including the sampling variability of the mean yields a variance estimate of

$$V = \sigma^2 + \frac{B}{MN}. \tag{72}$$

V is an overestimate of the variance of the distribution, and W is an underestimate initially. The latter arises because each chain won't have explored the target distribution adequately. The ratio of the two estimates is

$$\hat{R} = \frac{\frac{(N-1)}{N}W + B\left(1+\frac{1}{M}\right)\frac{1}{N}}{W}. \tag{73}$$

$\hat{R}$ should approach unity as convergence is achieved. How close to $\hat{R} = 1$? Opinions differ; I have seen suggestions to run the chain until the values of $\hat{R}$ are always < 1.03, or < 1.2, but a proof would be nice.

For convergence of chains with multiple parameters, Brooks & Gelman[**?**] generalised this to

$$\hat{R}^p = \frac{N-1}{N} + \frac{M+1}{M}\lambda \tag{74}$$

where λ is the largest eigenvalue of the symmetric matrix $W^{-1}B/N$, and W and B are generalisations from variances to covariances:

$$\frac{B_{\alpha\beta}}{N} = \frac{1}{(M-1)}\sum_{J=1}^{M}(\bar{\theta}_\alpha^J - \bar{\theta}_\alpha)(\bar{\theta}_\beta^J - \bar{\theta}_\beta) \tag{75}$$

and similarly for $W_{\alpha\beta}$.

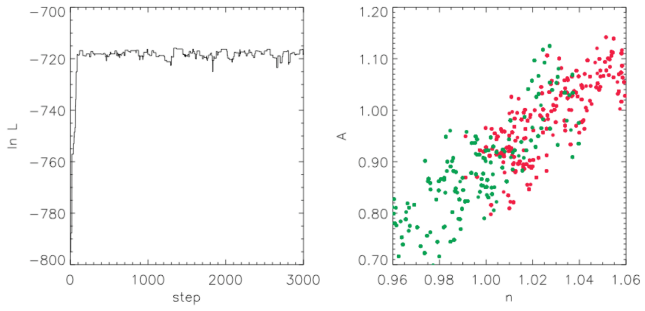

Fig. 2. Examples of unconverged chains. The left panel suggests the chain has converged, but the right panel shows the chain and a second one, also apparently converged, but showing clearly different distributions. From Verde et al. ApJS, 148, 195 (2003).

CosmoMC

Excellent resource. MCMC sampler with cosmological data (CMB + support for LSS, SNe). http://cosmologist.info/cosmomc/

For more details on MCMC in general, see [**?**, **?**, **?**].

4.2 Hamiltonian Monte Carlo

As we've seen, the proposal distribution has to be reasonably finely tuned to ensure good mixing, but not be too inefficient. What we would really like is to be able to take rather big jumps, so the chain is well-mixed, but in such a way that the probability of acceptance of each point is still high. Hamiltonian (or Hybrid) Monte Carlo (HMC) tries to do this, via a rather clever trick. In practice, it seems that it can be about M times faster than MCMC in M dimensions. Typical applications report that 4 times shorter chains give the same accuracy as MCMC. Originally developed for particle physics [**?**], there is a nice exposition in astrophysics by Hajian [**?**].

HMC works by sampling from a *larger* parameter space than we want to explore, by introducing M *auxiliary variables*, one for each parameter in the model. To see how this works, imagine each of the parameters in the problem as a coordinate. HMC regards the target distribution which we seek as an effective potential in this coordinate system, and for each coordinate it generates a generalised momentum. i.e. it expands the parameter space from its original 'coordinate space' to a phase space, in which there is a so-called 'extended' target distribution. It tries to sample from the extended target distribution in the $2M$ dimensions. It explores this space by treating the problem as a dynamical system, and evolving the phase space coordinates by solving the dynamical equations. Finally, it ignores the momenta (marginalising, as in MCMC), and this gives a sample of the original target distribution.

There are several advantages to this approach. One is that the probability of acceptance of a new point is high - close to unity; secondly, the system can make big jumps, so the mixing is better and the convergence faster. In doing the big jumps, it does some addi-

tional calculations, but these do not involve computing the likelihood, which is typically computationally expensive.

Let us see how it works. If the target density in M dimensions is $p(\boldsymbol{\theta})$, then we define a potential

$$U(\boldsymbol{\theta}) \equiv -\ln p(\boldsymbol{\theta}). \tag{76}$$

For each coordinate θ_α, we generate a momentum u_α, conveniently from a normal distribution with zero mean and unit variance, so the M-dimensional momentum distribution is a simple multivariate gaussian which we denote $\mathcal{N}(\mathbf{u})$. We define the kinetic energy

$$K(\mathbf{u}) \equiv \frac{1}{2}\mathbf{u}^T\mathbf{u}, \tag{77}$$

and the Hamilitonian is

$$H(\boldsymbol{\theta}, \mathbf{u}) \equiv U(\boldsymbol{\theta}) + K(\mathbf{u}). \tag{78}$$

The trick is that we generate chains to sample the *extended target density*

$$p(\boldsymbol{\theta}, \mathbf{u}) = \exp\left[-H(\boldsymbol{\theta}, \mathbf{u})\right]. \tag{79}$$

Since this is separable,

$$p(\boldsymbol{\theta}, \mathbf{u}) = \exp\left[-U(\boldsymbol{\theta})\right]\exp\left[-K(\mathbf{u})\right] \propto p(\boldsymbol{\theta})\mathcal{N}(\mathbf{u}) \tag{80}$$

and if we then marginalise over $\mathbf{u}$ by simply ignoring the $\mathbf{u}$ coordinates attached to each point in the chain (just as in MCMC), the resulting marginal distribution samples the desired target distribution $p(\boldsymbol{\theta})$. This is really very neat.

The key is that if we *exactly* solve the Hamiltonian equations

$$\begin{aligned} \dot{\theta}_\alpha &= u_\alpha \\ \dot{u}_\alpha &= -\frac{\partial H}{\partial \theta_\alpha} \end{aligned} \tag{81}$$

then H remains invariant, so the extended target density is always the same, and the acceptance is unity. Furthermore, we can integrate the equations for a long time if we wish, decorrelating the points in the chain.

There are several issues to consider.

- We seem to need the target density to define the potential, but this is what we are looking for. We need to approximate it.
- The aim is to do this fast, so we do not want to do many operations before generating a new sample. An easy way to achieve this is to employ a simple integrator (e.g. leap-frog; even Euler's method might do) and take several fairly big steps before generating a new point in the chain.
- We need to ensure we explore the extended space carefully.

The result of the two approximations is that H will not be quite constant. We deal with this as in MCMC by using the Metropolis algorithm, accepting the new point $(\boldsymbol{\theta}^*, \mathbf{u}^*)$ with a probability

$$min\left\{1, \exp\left[-H(\boldsymbol{\theta}^*, \mathbf{u}^*) + H(\boldsymbol{\theta}, \mathbf{u})\right]\right\}, \tag{82}$$

otherwise we repeat the old point $(\boldsymbol{\theta}, \mathbf{u})$ as usual.

How do we approximate? We want the gradients to be cheap to compute. It is usual to generate an approximate analytical potential by running a relatively short MCMC chain, computing the covariance of the points in the chain, and approximating the distribution by a multivariate gaussian with the same covariance. The gradients are then consequently easy to compute analytically.

The last point is that if we change the momentum only with Hamilton's equations of motion, we will restrict ourselves to a locus in phase space, and the target distribution will

not be properly explored. To avoid this, a new momentum is generated randomly when each point in the chain is generated. The art is to choose a good step in the integration, and the number of steps to take before generating a new point. Perhaps unsurprisingly, choosing these such that the new point differs from the previous one by about the size of the target peak works well. Thinning can be performed, and a convergence test must still be applied. Fig. **??** shows a comparison between HMC and MCMC for a simple case of a 6D gaussian target distribution, from [**?**].

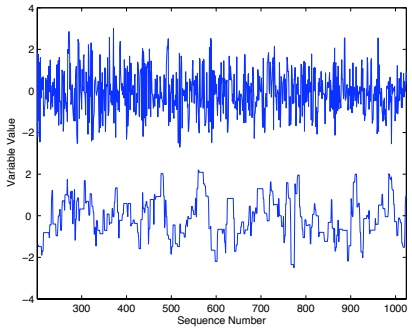

Fig. 3. A comparison of HMC sampling (top) and MCMC sampling (bottom). Note that the computer time required to generate each point in the HMC sampling will be larger than that of MCMC, so the actual gains are less than appears. From [**?**].

5 Model Selection

Model selection is in a sense a higher-level question than parameter estimation. In parameter estimation, one assumes a theoretical model within which one interprets the data, whereas in model selection, one wants to know which theoretical framework is preferred, given the data (regardless of the parameter values). The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example), or variants of the same idea. E.g. comparing a simple cosmological model where the Universe is assumed flat and the perturbations are strictly scale-invariant ($n = 1$), with a more general model where curvature is allowed to vary and the spectrum is allowed to deviate from scale-invariance. The sort of question asked here is essentially 'Do the data require a more complex model?'. Clearly in the latter type of comparison χ^2 itself will be of no use - it will always reduce if we allow more freedom. There are frequentist ways to try and answer these questions, but we are all by now confirmed Bayesians[5], so will approach it this way.

5.1 Bayesian evidence

The Bayesian method to select between models (e.g. General Relativity, & modified gravity) is to consider the Bayesian evidence ratio. Essentially we want to know if, given the data, there is evidence that we need to expand the space of gravity models beyond GR. Assuming non-commital priors for the models (i.e. the same a priori probability), the probability of the models given the data is simply proportional to the evidence.

We denote two competing models by M and M'. We assume that M' is a simpler model, which has fewer ($n' < n$) parameters in it. We further assume that it is *nested* in Model M', i.e. the n' parameters of model M' are common to M, which has $p \equiv n - n'$ extra parameters in it. These parameters are fixed to fiducial values in M'.

We denote by $\mathbf{x}$ the data vector, and by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ the parameter vectors (of length n and n').

[5] If not, please leave the room

Apply Rule 1: Write down what you want to know. Here it is $p(M|\mathbf{x})$ - the probability of the model, given the data.

The posterior probability of each model comes from Bayes' theorem:

$$p(M|\mathbf{x}) = \frac{p(\mathbf{x}|M)p(M)}{p(\mathbf{x})} \tag{83}$$

and similarly for M'. By marginalisation $p(\mathbf{x}|M)$, known as the *Evidence*, is

$$p(\mathbf{x}|M) = \int d\boldsymbol{\theta}\, p(\mathbf{x}|\boldsymbol{\theta}M)p(\boldsymbol{\theta}|M), \tag{84}$$

which should be interpreted as a multidimensional integration. Hence the posterior relative probabilities of the two models, regardless of what their parameters are[6], is

$$\frac{p(M'|\mathbf{x})}{p(M|\mathbf{x})} = \frac{p(M')}{p(M)} \frac{\int d\boldsymbol{\theta}'\, p(\mathbf{x}|\boldsymbol{\theta}'M')p(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta}\, p(\mathbf{x}|\boldsymbol{\theta}M)p(\boldsymbol{\theta}|M)}. \tag{85}$$

With non-committal priors on the models, $p(M') = p(M)$, this ratio simplifies to the ratio of evidences, called the *Bayes Factor*,

$$B \equiv \frac{\int d\boldsymbol{\theta}'\, p(\mathbf{x}|\boldsymbol{\theta}'M')p(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta}\, p(\mathbf{x}|\boldsymbol{\theta}M)p(\boldsymbol{\theta}|M)}. \tag{86}$$

Note that the a complicated model M will (if M' is nested) inevitably lead to a higher likelihood (or at least as high), but the evidence will favour the simpler model if the fit is nearly as good, through the smaller prior volume.

We assume uniform (and hence separable) priors in each parameter, over ranges $\Delta\boldsymbol{\theta}$ (or $\Delta\boldsymbol{\theta}'$). Hence $p(\boldsymbol{\theta}|M) = (\Delta\boldsymbol{\theta}_1 \ldots \Delta\boldsymbol{\theta}_n)^{-1}$ and

$$B = \frac{\int d\boldsymbol{\theta}'\, p(\mathbf{x}|\boldsymbol{\theta}', M')}{\int d\boldsymbol{\theta}\, p(\mathbf{x}|\boldsymbol{\theta}, M)} \frac{\Delta\boldsymbol{\theta}_1 \ldots \Delta\boldsymbol{\theta}_n}{\Delta\boldsymbol{\theta}'_1 \ldots \Delta\boldsymbol{\theta}'_{n'}}. \tag{87}$$

Note that if the prior ranges are not large enough to contain essentially all the likelihood, then the position of the boundaries would influence the Bayes factor. In what follows, we will assume the prior range is large enough to encompass all the likelihood.

In the nested case, the ratio of prior hypervolumes simplifies to

$$\frac{\Delta\boldsymbol{\theta}_1 \ldots \Delta\boldsymbol{\theta}_n}{\Delta\boldsymbol{\theta}'_1 \ldots \Delta\boldsymbol{\theta}'_{n'}} = \Delta\boldsymbol{\theta}_{n'+1} \ldots \Delta\boldsymbol{\theta}_{n'+p}, \tag{88}$$

where $p \equiv n - n'$ is the number of extra parameters in the more complicated model.

Here we see the problem. The evidence requires a multidimensional integration over the likelihood and prior, and this may be *very* expensive to compute. There are various ways to simplify this. One is analytic - follow the Fisher approach and assume the likelihood is a multivariate gaussian, others are numerical, such as nested sampling, where one tries to sample the likelihood in an efficient way. There are others, but we will focus on these. Note that shortcuts with names such as AIC and BIC may be unreliable as they are based on the best-fit χ^2, and from a Bayesian perspective we want to know how much parameter space would give the data with high probability. See [**?**] for more discussion.

5.2 Laplace approximation

The Bayes factor in equation (**??**) still depends on the specific dataset $\mathbf{x}$. For future experiments, we do not yet have the data, so we compute the expectation value of the Bayes factor, given the statistical properties of $\mathbf{x}$. The expectation is computed over the distribution of $\mathbf{x}$

[6] If a model has no parameters, then the integral is simply replaced by $p(\mathbf{x}|M)$

for the correct model (assumed here to be M). To do this, we make two further approximations: first we note that B is a ratio, and we approximate $\langle B\rangle$ by the ratio of the expected values, rather than the expectation value of the ratio. This should be a good approximation if the evidences are sharply peaked.

We also make the Laplace approximation, that the expected likelihoods are given by multivariate Gaussians. For example,

$$\langle p(\mathbf{x}|\boldsymbol{\theta},M)\rangle = L_0 \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)_\alpha \mathsf{F}_{\alpha\beta}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)_\beta\right] \tag{89}$$

and similarly for $\langle p(\mathbf{x}|\boldsymbol{\theta}',M')\rangle$. This assumes that a Taylor expansion of the likelihood around the peak value to second order can be extended throughout the parameter space. $F_{\alpha\beta}$ is the Fisher matrix, given for Gaussian-distributed data by equation (**??**):

$$\mathsf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\mathsf{C}^{-1}\mathsf{C}_{,\beta} + \mathsf{C}^{-1}(\mu_{,\beta}\mu^T_{,\alpha} + \mu_{,\alpha}\mu^T_{,\beta})\right]. \tag{90}$$

C is the covariance matrix of the data, and μ its mean (no noise). Commas indicate partial derivatives with respect to the parameters. For the correct model M, the peak of the expected likelihood is located at the true parameters $\boldsymbol{\theta}_0$. Note, however, that for the incorrect model M', the peak of the expected likelihood is not in general at the true parameters (see Fig. **??** for an illustration of this). This arises because the likelihood in the numerator of equation (**??**) is the probability of the dataset $\mathbf{x}$ given incorrect model assumptions.

The Laplace approximation is routinely used in forecasting marginal errors in parameters, using the Fisher matrix. Clearly the approximation may break down in some cases, but for Planck, the Fisher matrix errors are reasonably close to (within 30% of) those computed with Monte Carlo Markov Chains.

If we assume that the posterior probability densities are small at the boundaries of the prior volume, then we can extend the integrations to infinity, and the integration over the multivariate Gaussians can be easily done. This gives, for M, $(2\pi)^{n/2}(\det \mathsf{F})^{-1/2}$, so for nested models,

$$\langle B\rangle = (2\pi)^{-p/2}\frac{\sqrt{\det \mathsf{F}}}{\sqrt{\det \mathsf{F}'}}\frac{L'_0}{L_0}\Delta\boldsymbol{\theta}_{n'+1}\ldots\Delta\boldsymbol{\theta}_{n'+p}. \tag{91}$$

An equivalent expression was obtained, using again the Laplace approximation by [**?**]. The point here is that with the Laplace approximation, one can compute the L'_0/L_0 ratio from the Fisher matrix. To compute this ratio of likelihoods, we need to take into account the fact that, if the true underlying model is M, in M' (the incorrect model), the maximum of the expected likelihood will not in general be at the correct values of the parameters (see Fig. **??**). The n' parameters shift from their true values to compensate for the fact that, effectively, the p additional parameters are being kept fixed at incorrect fiducial values. If in M', the additional p parameters are assumed to be fixed at fiducial values which differ by $\delta\psi_\alpha$ from their true values, the others are shifted on average by an amount which is readily computed under the assumption of the multivariate Gaussian likelihood:

$$\delta\boldsymbol{\theta}'_\alpha = -(\mathsf{F}'^{-1})_{\alpha\beta}\mathsf{G}_{\beta\zeta}\delta\psi_\zeta \qquad \alpha,\beta = 1\ldots n', \zeta = 1\ldots p \tag{92}$$

where

$$\mathsf{G}_{\beta\zeta} = \frac{1}{2}\mathrm{Tr}\left[\mathsf{C}^{-1}\mathsf{C}_{,\beta}\mathsf{C}^{-1}\mathsf{C}_{,\zeta} + \mathsf{C}^{-1}(\mu_{,\zeta}\mu^T_{,\beta} + \mu_{,\beta}\mu^T_{,\zeta})\right], \tag{93}$$

which we recognise as a subset of the Fisher matrix. For clarity, we have given the additional parameters the symbol ψ_ζ; $\zeta = 1\ldots p$ to distinguish them from the parameters in M'.

With these offsets in the maximum likelihood parameters in model M', the ratio of likelihoods is given by

$$L'_0 = L_0 \exp\left(-\frac{1}{2}\delta\boldsymbol{\theta}_\alpha \mathsf{F}_{\alpha\beta}\delta\boldsymbol{\theta}_\beta\right) \tag{94}$$

where the offsets are given by $\delta\boldsymbol{\theta}_\alpha = \delta\boldsymbol{\theta}'_\alpha$ for $\alpha \le n'$ (equation **??**), and $\delta\boldsymbol{\theta}_\alpha = \delta\psi_{\alpha-n'}$ for $\alpha > n'$.

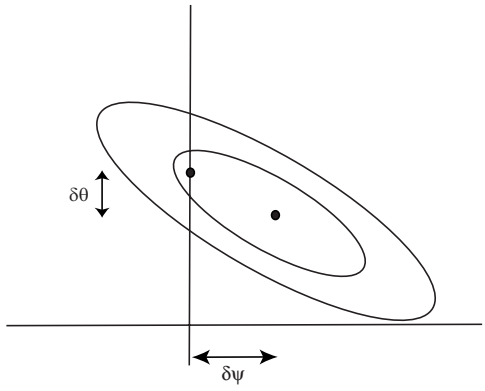

Fig. 4. Illustrating how assumption of a wrong parameter value can influence the best-fitting value of other model parameters. Ellipses represent iso-likelihood surfaces, and here in the simpler model, the parameter on the horizontal axis is assumed to take the value given by the vertical line. Filled circles show the true parameters in the more complicated model, and the best-fit parameters in the simpler model. From [**?**].

The final expression for the expected Bayes factor is then

$$\langle B \rangle = (2\pi)^{-p/2} \frac{\sqrt{\det \mathsf{F}}}{\sqrt{\det \mathsf{F}'}} \exp\left(-\frac{1}{2}\delta\boldsymbol{\theta}_\alpha \mathsf{F}_{\alpha\beta}\delta\boldsymbol{\theta}_\beta\right) \prod_{q=1}^{p} \Delta\boldsymbol{\theta}_{n'+q}. \tag{95}$$

Note that F and F^{-1} are $n \times n$ matrices, F' is $n' \times n'$, and G is an $n' \times p$ block of the full $n \times n$ Fisher matrix F. The expression we find is a specific example of the Savage-Dickey ratio (see e.g. [**?**]). For a nested model with a single additional parameter $\boldsymbol{\theta}_i$,

$$\langle B \rangle = \frac{p(\boldsymbol{\theta}_i|\mathbf{x})}{p(\boldsymbol{\theta}_i)}. \tag{96}$$

Here we explicitly use the Laplace approximation to compute the offsets in the parameter estimates which accompany the wrong choice of model, and compute the evidence ratio explicitly. Finally, note that this is the expected evidence ratio (nearly); it does not address the issue of what the distribution of evidence ratios should be.

Note that the 'Occam's razor' term, common in evidence calculations, is to some extent encapsulated in the term $(2\pi)^{-p/2}\frac{\sqrt{\det \mathsf{F}}}{\sqrt{\det \mathsf{F}'}}$, multiplied by the prior product: models with more parameters are penalised in favour of simpler models, unless the data demand otherwise. In cases where the Laplace approximation is not a good one, other techniques must be used, at more computational expense.

It is perhaps worth remarking that Occam's razor appears not to be fully incorporated into this term, as can be seen by considering a situation where the data do not depend at all on an additional parameter. In this case, the Bayesian evidence ratio is unity, so disappointingly no preference is shown at all.

As an example, consider testing General Relativity against other gravity theories, which predict a different growth rate of perturbations, $d\ln\delta/d\ln\Omega_m = \gamma$, where $\gamma = 0.55$ for GR, and (for example) $\gamma = 0.68$ for a flat DGP braneworld model. This can be probed with weak lensing, and we ask the question do the data favour a model where γ is a free parameter, rather than being fixed at 0.55?

We take a prior range $\Delta\gamma = 1$, and we ask the question of how different the growth rate of a modified-gravity model would have to be for these experiments to be expected to favour a relaxation of the gravity model from General Relativity. This is shown in Fig.**??**. It shows how the expected evidence ratio changes with progressively greater differences from the General Relativistic growth rate. We see that a next-generation weak lensing survey could even distinguish 'strongly' $\delta\gamma = 0.048$. Note that changing the prior range $\Delta\gamma$ by a

factor 10 changes the numbers by ~ 0.012, so the dependence on the prior range is rather small.

If one prefers to ask a frequentist question, then a combination of WL+*Planck*+BAO+SN should be able to distinguish $\delta\gamma = 0.13$, at 10.6σ. Alternatively, one can calculate the expected error on γ [**?**] within the extended model M. In this section, we are asking a slightly different question of whether the data demand that a wider class of models needs to be considered at all, rather than estimating a parameter within that wider class.

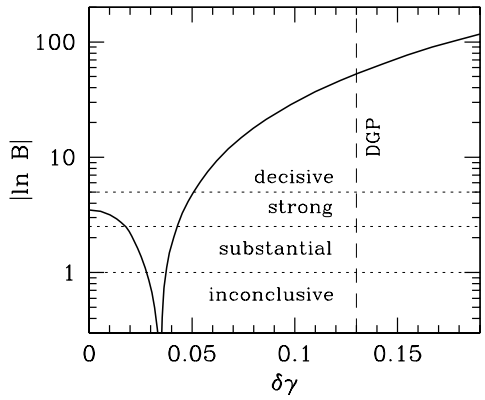

Fig. 5. The expected value of $|\ln\langle B\rangle|$ from a future large-scale deep weak lensing survey, as might be done with Euclid, JDEM or LSST, in combination with CMB constraints from *Planck*, as a function of the difference in the growth rate between the modified-gravity model and General Relativity. The crossover at small $\delta\gamma$ occurs because Occam's razor will favour the simpler (General Relativity) model unless the data demand otherwise. To the left of the cusp, GR would be likely to be preferred by the data. The dotted vertical line shows the offset of the growth factor for the flat DGP model. The descriptors are the terminology of Jeffreys (1961) [**?**]. From [**?**]

The case for a large, space-based 3D weak lensing survey is strengthened, as it offers the possibility of conclusively distinguishing Dark Energy from at least some modified gravity models.

5.3 Numerical sampling methods

In order to compute the evidence numerically, the prior volume needs to be sampled, in much the same way as in parameter estimation, except the requirements are slightly more stringent. MCMC could be used, although there are claims that it is not good at exploring the parameter space adequately. I find these claims puzzling, as in evidence calculations we are doing an N-dimensional integral, which is not so different from doing the $(N-1)$-dimensional integral in MCMC to get marginal errors. However, here are a couple of other sampling techniques.

The VEGAS algorithm, with rotation

This is suitable for single-peaked likelihoods. It is in Numerical Recipes, but needs a modification for efficiency. Essentially, one seeks to sample from a distribution which is close to the target distribution (sampling from a different distribution is called *importance sampling*), but one does not know what it is yet. One can do this iteratively, sampling from the prior first, then estimating the posterior to get a first guess at the posterior, and using that to refine the sampling distribution.

Now, one does not want to draw randomly from a *non-separable* function of the parameters, as a moment's thought will tell you that this is computationally very expensive, so one seeks a separable function, so one can then draw the individual parameters one after the

other from N distributions. This works well if the target distribution is indeed separable in the parameters, but not otherwise.

The key extra step [**?**] is to rotate the parameter axes, which can be done by computing (for example) the moments of the distribution (essentially the moment-of-inertia) after any step, and diagonalising it to find the eigenvectors.

The probability to be sampled is

$$p(\boldsymbol{\theta}) \propto g_1(\theta_1) g_2(\theta_2) \dots g_M(\theta_M). \tag{97}$$

where it can be shown that

$$g_\alpha(\theta_\alpha) \propto \sqrt{\int_{\beta \neq \alpha} d^{M-1}\theta_\beta \frac{f^2(\boldsymbol{\theta})}{\prod_{\beta \neq \alpha} g_\beta(\theta_\beta)}} \tag{98}$$

where f is the desired target distribution. Note that g depends on all other gs. g can be improved iteratively.

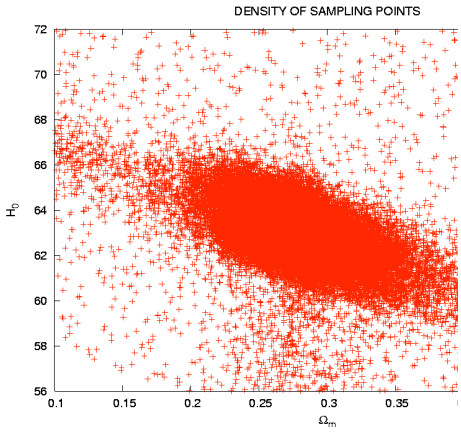

Fig. 6. The VEGAS Sampling algorithm, applied to supernova data (from [**?**]

Nested sampling

Other sampling methods can also be very effective (see e.g. [**?**, **?**, **?**, **?**]). Nested sampling was introduced by Skilling [**?**]. One samples from the prior volume, and gradually concentrates more points near the peak of the likelihood distribution, by repeatedly replacing the point with the lowest target density by one drawn strictly from the prior volume with higher target density. This has proved effective for cosmological model selection. I will not go into details here, except to say that the key is in drawing the new point from a suitable subset of the prior volume. This must be increasingly smaller as the points get more confined, otherwise the trial points will be rejected with increasing frequency and it becomes very inefficient, but it must also be a large enough subset that the entire prior volume above the lowest target density is almost certainly included inside. For further details, see the original papers. Note that, for multimodal target distributions, a modification called MultiNest exists [**?**]. CosmoNest and MultiNest are publicly-available additions to CosmoMC.

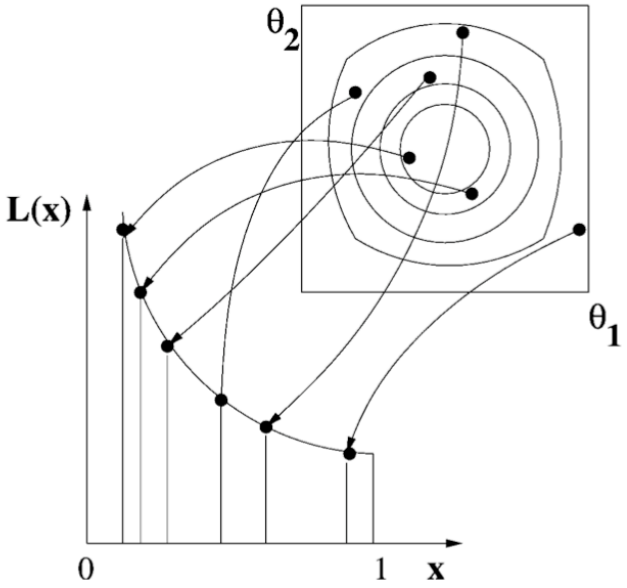

Fig. 7. The Nested Sampling algorithm. From [**?**]

6 Summary

We have explored the use of (principally) Bayesian methods for parameter estimation and model selection, and looked at some common numerical methods for performing these tasks. We have also shown how it is possible to work out the (minimum) errors on parameters, and on model probabilities, in advance of doing an experiment, to see if it is worthwhile. These are the common Fisher matrix approach for parameter estimation, and the expected evidence for model selection, which requires no more than the Fisher matrix to compute.

7 Exercises

1. A Bayesian Exercise. This is a variant of the famous *Monty Hall* problem, named after a game show host. For this exercise, you must formulate the problem in a fully Bayesian way.

You are participating in a game show where there are three small doors presented to you. You are told that behind two of them there is a bottle of fine italian wine, and behind the third is a can of the famous scottish soft drink, Irn Bru. You will win one of them, and naturally you prefer the Irn Bru. The game works as follows:

a) You point to, *but do not open*, one of the doors.
b) The game show host opens one of the other doors, revealing a bottle of wine.
c) You now either open the door which you originally selected, or switch to the third door. Which should you open?

2. This is a model selection problem, and will get you thinking about suitable priors. You draw lottery ticket number 1475567, and it wins. What is the probability that the lottery was rigged, so that the winning ticket was predetermined?

3. An exercise in linear algebra
a) Prove that $(\mathsf{C}^{-1})_{,\alpha} = -\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}$
b) Prove that $(\ln \mathsf{C})_{,\alpha} = \mathsf{C}^{-1}\mathsf{C}_{,\alpha}$.
c) Prove that $\ln\det \mathsf{C} = \mathrm{Tr}\ln \mathsf{C}$.

Hints:
a) What is $\mathsf{C}\mathsf{C}^{-1}$?
b) For matrices, $\exp(\mathsf{A}) \equiv \mathsf{I} + \mathsf{A} + \mathsf{A}^2/2! + \ldots$. ln is the inverse of exp.

c) If you rotate coordinate axes (in parameter space in this case), the determinant and trace don't change.

4. A survey is proposed, to determine the mean number density of a certain type of astronomical object, whose positions are random. The survey measures the number of objects in each of N independent cells of the same size and shape, such that the mean number per cell, $\bar{n}$, is $\gg 1$. If the cells are observed to have n_i objects in them ($i = 1, \ldots N$), show that the Fisher matrix (a scalar in this case) for $\bar{n}$ is

$$\mathsf{F} = \frac{N}{2\bar{n}^2} + \frac{N}{\bar{n}}. \tag{99}$$

Where is most of the information coming from, the dependence of $\boldsymbol{\mu}$ on $\bar{n}$, or C?

5. Binomial drinking - an interesting paradox. Precursor question. If we have an experiment with some discrete outcomes $n = 1, \ldots, \infty$, each occurring with probability P_n. Argue that the probability of the sum of two drawings $z = m + n$ from the distribution is

$$p_z(z) = \sum_{n=1}^{z-1} P_n P_{z-n}.$$

Every minute, as the second hand on a large clock on the wall reaches the top, I think about drinking a bottle of Irn Bru. I drink one with a probability p. Show that the probability of the next drink being taken at the M^{th} opportunity is

$$P_M = pq^{M-1}$$

where $q = 1 - p$.

Show that the expectation value of the number of time steps between outbursts is $\bar{M} = 1/p$. (Hint: expand $(1-q)^{-2}$ in a Taylor expansion).

Now, my friend comes in at a random time (not necessarily as the second hand reaches the top of the clock). Argue that the probability that I took my last drink m time steps previously was

$$P'_m = pq^{m-1}.$$

Show that the probability for the variable $S = M + m$ ($S \geq 2$) is

$$P_S = \sum_{M=1}^{S-1} p^2 q^{S-2}.$$

By expanding a different power of $1 - q$, show that the expectation value of S is

$$\bar{S} = \frac{2}{p},$$

and hence that the average time between the last output and the next one is

$$\bar{t} = \frac{2}{p} - 1.$$

For $p = 0.1$, $\bar{t} = 19$, compared with 10 for the mean time between drinks. How is the paradox resolved?

References

1. Amendola L., Kunz M., Sapone D., JCAP, 04, 013 (2008)
2. Beltran M., Garcia-Bellido J., Lesgourgues J., Liddle A. R., Slosar A., 2005, Phys, Rev., D71, 063532

3. Brooks S.P., Gelman A., Journal of Computational and Graphical Statistics, 7, 434 (1998)
4. Duan et al., Phys. Lett. B195, 216 (1987)
5. Feroz F., Hobson M., Bridges M., astroph/0809.3437 (2008)
6. Fisher R. A., J. Roy. Stat. Soc., 98, 39 (1935)
7. Gilks W.R, Richardson S., Spiegelhalter D.J., *Markov chain Monte Carlo in practice*, Chapman & Hall, London (1996)
8. Hajian A., astroph/0608679 (2006)
9. Hamilton A.J.S., astroph/0503603 (2005)
10. Heavens A.F., Kitching T., Verde L., MNRAS, 380, 1029 (2007).
11. Hobson M.P., Bridle S.L., Lahav O., 2002, MNRAS, 335, 377
12. Jeffreys H., 1961, Theory of Probability, Oxford University Press (Oxford, UK)
13. Kendall M. G., Stuart A., *The Advanced Theory of Statistics, Volume II*, Griffin, London (1969)
14. Kenney, J. F. & Keeping, E. S., *Mathematics of Statistics, Part II*, 2nd ed. (Van Nostrand, New York) (1951)
15. Kosowsky A., Milosavljevic M., Jimenez R., Phys. Rev. D66, 3007 (2002).
16. Lazarides, G., Ruiz de Austri, R., Trotta, R., Phys. Rev. D70, 123527
17. Lewis A., Bridle S., PRD, 66, 103511 (2002)
18. Liddle A., MNRAS, 377, 74 (2007).
19. Mukherjee P., Parkinson D., Liddle A.R., 2006, ApJ, 638, L51
20. Press W., et al, *Numerical Recipes in Fortran*, CUP, Cambridge, U.K. (1992)
21. Serra P., Heavens A., Melchiorri A., MNRAS 379, 169 (2007)
22. Skilling J., 2004, avaliable at http://www.inference.phy.cam.ac.uk/bayesys
23. Tegmark M., Taylor, A.N., Heavens A.F., ApJ, 480, 22 (1997) (TTH)
24. Trotta R., 2007, astroph/0703063
25. Verde L., astroph/0712.3028 (2007)
26. Verde L., et al, ApJS, 148, 195 (2003)

8 Solutions to selected exercises

3. (a) Prove that $(\mathsf{C}^{-1})_{,\alpha} = -\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\,\mathsf{C}^{-1}$

Since $\mathsf{C}\mathsf{C}^{-1} = \mathsf{I}$, its derivative is zero. Hence $\mathsf{C}(\mathsf{C}^{-1})_{,\alpha} + \mathsf{C}_{,\alpha}\mathsf{C}^{-1} = 0$. Result follows after premultiplication by C^{-1}.

(b) Prove that $(\ln \mathsf{C})_{,\alpha} = \mathsf{C}^{-1}\mathsf{C}_{,\alpha}$.

Let $\mathsf{A} = \ln \mathsf{C}$, so

$$\mathsf{C} = \exp \mathsf{A} = \mathsf{I} + \mathsf{A} + \frac{\mathsf{A}^2}{2!} + \ldots = \sum_{n=0}^{\infty} \frac{\mathsf{A}^n}{n!}$$

Hence

$$\mathsf{C}_{,\alpha} = \sum_{n=1}^{\infty} \frac{\mathsf{A}^{n-1}}{(n-1)!}\,\mathsf{A}_{,\alpha} = \sum_{m=0}^{\infty} \frac{\mathsf{A}^m}{m!}(\ln \mathsf{C})_{,\alpha} = \mathsf{C}(\ln \mathsf{C})_{,\alpha}$$

and result follows.

(c) Prove that $\ln\det \mathsf{C} = \mathrm{Tr} \ln \mathsf{C}$.

C is a symmetric matrix and therefore be diagonalised. In the new basis (where we have a new set of parameters which are linear combinations of the old ones), the diagonal components of C are strictly positive (they are variances of the new parameters).

Since the trace and determinant are unchanged by the diagonalisation, we can prove the result in the rotated system. If C is diagonal, then $\ln \mathsf{C}$ is diagonal[7], with components $\ln \mathsf{C}_{11},\ \ln \mathsf{C}_{22}, \ldots$. So $\mathrm{Tr} \ln \mathsf{C} = \sum_n \ln \mathsf{C}_{nn}$. Since $\det \mathsf{C} = \prod_n \mathsf{C}_{nn}$,

$$\ln\det \mathsf{C} = \sum_n \ln \mathsf{C}_{nn} = \mathrm{Tr} \ln \mathsf{C}.$$

[7] Let us diagonalise A to $\mathsf{A}' = \mathsf{R}^T\mathsf{A}\mathsf{R}^T$, so $\mathsf{A}'^n = \mathsf{R}^T\mathsf{A}^n\mathsf{R}$ (since $\mathsf{R}\mathsf{R}^T = \mathsf{I}$), so A^n is diagonalised by the same matrix as diagonalises A. Further if we consider $\mathsf{C}' = \mathsf{R}^T\mathsf{C}\mathsf{R} = \mathsf{R}e^{\mathsf{A}}\mathsf{R} = \sum_{n=0}^{\infty} \mathsf{R}^T\mathsf{A}^n\mathsf{R}/n! = \sum_{n=0}^{\infty}(\mathsf{R}^T\mathsf{A}\mathsf{R})^n/n! = \sum_{n=0}^{\infty}(\mathsf{A}')^n/n! = e^{\mathsf{A}'}$ so C is diagonalised by the same rotation which diagonalises A.

This proof is rigorous, but there may be a neater solution without diagonalisation.

4. For $\bar{n} \gg 1$, we can approximate the Poisson distribution by a gaussian, with $\langle n_i \rangle = \bar{n}$, and $\sigma_i^2 = \bar{n}$. Hence $\boldsymbol{\mu}^T = \bar{n}(1, 1, \ldots 1)$ and $\mathsf{C} = \bar{n}\,\mathrm{diag}(1, 1, \ldots, 1)$. Hence $\mathsf{C}_{,1} = \mathrm{diag}(1, 1, \ldots, 1)$, and M_{ij} is an $N \times N$ matrix filled with 2s. The result

$$\mathsf{F} = \frac{N}{2\bar{n}^2} + \frac{N}{\bar{n}} \tag{100}$$

follows. The first term arises from $\mathsf{C}_{,1}$, and the second from $\boldsymbol{\mu}_{,1}$. The second dominates since $\bar{n} \gg 1$.