Fisher Matrices and All That: Experimental Design and Data Compression

A. Heavens

Institute for Astronomy, University of Edinburgh, Blackford Hill,
Edinburgh EH9 3HJ
afh@roe.ac.uk

Abstract. 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 also take a look at data compression methods, such as Karhunen–Loève and MOPED, which can be extremely valuable for speeding up analysis of large data sets, and making the analysis more stable.

1 Introduction: Data Analysis

Most data analysis problems are *inverse problems*. You have a set of data $\{\boldsymbol{x}\}$, and you wish to interpret the data in terms of a model. The model will typically have some parameters $\{\boldsymbol{\Theta}\}$ in it, which you want to determine. Clearly, what one wants is the probability distribution of $\boldsymbol{\Theta}$, given the data $\boldsymbol{x}$, i.e.

$$p(\boldsymbol{\Theta}|\boldsymbol{x}) \; . \tag{1}$$

From this one can calculate the expectation values of the parameters, and their errors. What may be readily calculable is not this, rather the opposite, $p(\boldsymbol{x}|\boldsymbol{\Theta})$. For example, consider a model which is a Gaussian with mean μ and variance σ^2. The probability of a single variable x given the parameters $\boldsymbol{\Theta} = (\mu, \sigma)$ 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 $p(\boldsymbol{\Theta}|\boldsymbol{x})$ using Bayes' Theorem:

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

- $p(\boldsymbol{x}|\boldsymbol{\Theta})$ is called the *likelihood* $L(\boldsymbol{x}; \boldsymbol{\Theta})$.

Heavens, A.: *Fisher Matrices and All That: Experimental Design and Data Compression.*
Lect. Notes Phys. **665**, 51–72 (2009)
DOI 10.1007/978-3-540-44767-2_2 © Springer-Verlag Berlin Heidelberg 2009

- $p(\boldsymbol{\Theta})$ 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"), but other choices can sometimes be justified.
- $p(\boldsymbol{x})$ is the *evidence*. In the end, we are interested in the relative probabilities of the parameters, and this term does not depend on them, so is often ignored. It can play a role when more than one theoretical model is being considered, and one wants to choose which model is most likely, whatever the parameters are. (A different model should be interpreted loosely, e.g. one may consider a flat CDM Universe as a model, or a CDM Universe with non-zero curvature. The curvature is another parameter, but the "model" in this context differs, as it has one more parameter.)

Thus for flat priors, we have simply

$$p(\boldsymbol{\Theta}|\boldsymbol{x}) \propto L(\boldsymbol{x};\boldsymbol{\Theta}) \; . \tag{4}$$

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 $p(\boldsymbol{x}|\boldsymbol{\Theta})$ is not the maximum-likelihood estimate.

My rule of thumb is that if the priors are assigned theoretically, and they influence the result significantly, the data are usually not good enough.

2 Fisher Matrix Analysis

This has been adapted from [14] (hereafter TTH).

How accurately can we estimate model parameters from a given data set? This question was basically answered 60 years ago [1], 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 $x_1, x_2, \ldots, x_N$, which we arrange in an n-dimensional vector $\boldsymbol{x}$. 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, the first N coefficients of a Fourier–Bessel 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 $\boldsymbol{x}$ as a random variable with some probability distribution $L(\boldsymbol{x};\boldsymbol{\Theta})$, which depends in some known way on a vector of M model parameters $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_M)$.

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 $\boldsymbol{x}$, it too is a random variable. For it to be a good estimate, we would of course like it to be unbiased, i.e.

$$\langle \boldsymbol{\Theta} \rangle = \boldsymbol{\Theta}_0 \ , \tag{5}$$

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

$$\Delta\theta_i \equiv \left(\left\langle \theta_i^2 \right\rangle - \langle \theta_i \rangle^2 \right)^{1/2} \ . \tag{6}$$

In statistics jargon, we want the BUE θ_i, which stands for the "best unbiased estimator".

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

$$\mathsf{F}_{ij} \equiv \left\langle \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} \right\rangle \ , \tag{7}$$

where $\mathcal{L} \equiv -\ln L$. Another key quantity is the *maximum-likelihood estimator*, or *ML-estimator* for brevity —indexML-estimator defined as the parameter vector $\boldsymbol{\Theta}_{\rm ML}$ that maximizes the likelihood function $L(\boldsymbol{x}; \boldsymbol{\Theta})$.

Using this notation, a number of powerful theorems have been proved (see, e.g. [8, 7]):

1. For any unbiased estimator, $\Delta\theta_i \geq 1/\sqrt{\mathsf{F}_{ii}}$.
2. If there is a BUE $\boldsymbol{\Theta}$, then it is the ML-estimator or a function thereof.
3. The ML-estimator is asymptotically BUE.

The first of these theorems, known as the Cramér–Rao inequality, 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. This is called the *conditional error*, and is the minimum error bar attainable on θ_i if all the other parameters are known. *It is rarely relevant and should almost never be quoted.*

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

$$\Delta\theta_i \geq \left(\mathsf{F}^{-1} \right)_{ii}^{1/2} \ . \tag{8}$$

This is called the *marginal error*, and 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. 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

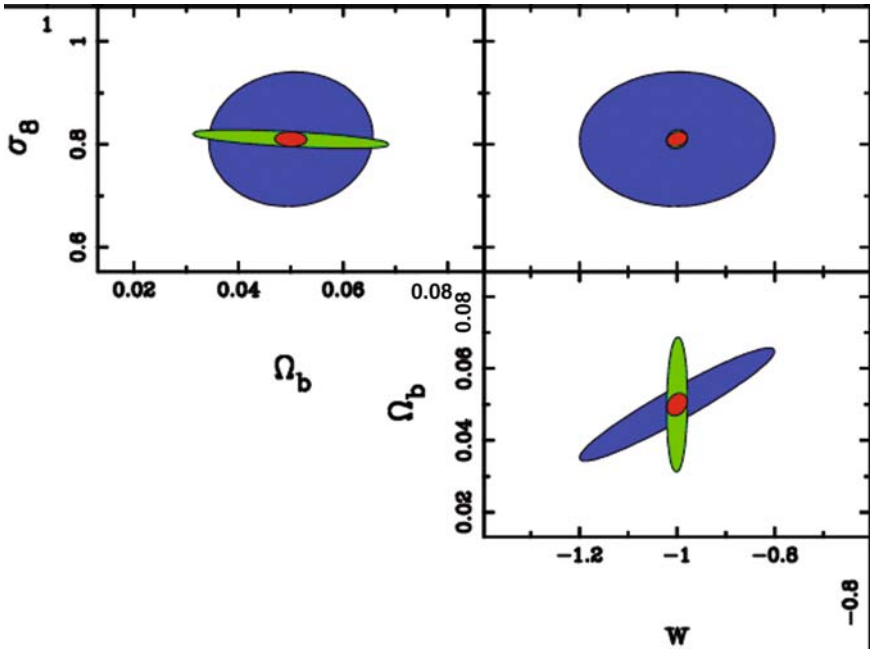

Fig. 1. The expected error ellipses for cosmological parameters (σ_8, baryon density parameter Ω_b, and dark energy equation of state $w \equiv p/\rho c^2$) from a 3D weak lensing survey of 1000 square degrees, with a median redshift of 1 and a photometric redshift error of 0.15. Probabilities are marginalized 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 (Figure courtesy Tom Kitching)

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.

2.1 Fisher Information Matrix

Although the proof of the Cramér–Rao inequality is rather lengthy, it is quite easy to acquire some intuition for why the Fisher information matrix has the form that it does. This is the purpose of the present section.

Let us Taylor expand $\mathcal{L}$ around the ML-estimate $\mathbf{\Theta}$. By definition, all the first derivatives $\partial\mathcal{L}/\partial\theta_i$ will vanish at the ML-point, since the likelihood function has its maximum there, so the local behaviour will be dominated by the quadratic terms. Since $L = \exp[-\mathcal{L}]$, we thus see that the likelihood function will be approximately Gaussian near the ML-point. If the error bars are quite small, L usually drops sharply before third-order terms have become

important, so that this Gaussian is a good approximation to L everywhere. Interpreting L as a Bayesian probability distribution in parameter space, the covariance matrix T is thus given simply by the second derivatives at the ML-point, as the inverse of the Hessian matrix:

$$(\mathsf{T}^{-1})_{ij} \equiv \frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} \; . \tag{9}$$

Note that the Fisher information matrix F is simply the expectation value of this quantity at the point $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$ (which coincides with the ML-point on average if the ML-estimate is unbiased). It is basically a measure of how fast (on average) the likelihood function falls off around the ML-point, i.e. a measure of the width and shape of the peak. From this discussion, it is also clear that we can use its inverse, F^{-1}, as an estimate of the covariance matrix

$$\mathsf{T} \equiv \langle \boldsymbol{\Theta}\boldsymbol{\Theta}^t \rangle - \langle \boldsymbol{\Theta} \rangle \langle \boldsymbol{\Theta} \rangle^t \tag{10}$$

of our parameter estimates when we use the ML-method.

2.2 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 $n \ln[2\pi]$)

$$2\mathcal{L} = \ln\det \mathsf{C} + (\boldsymbol{x} - \boldsymbol{\mu})\, \mathsf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})^t \; , \tag{11}$$

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

$$\mathsf{C} = \langle (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^t \rangle \tag{12}$$

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

$$\mathsf{D} \equiv (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^t \tag{13}$$

and using the matrix identity (see exercises) $\ln\det \mathsf{C} = \mathrm{Trace} \ln \mathsf{C}$, we can re-write (11) as

$$2\mathcal{L} = \mathrm{Trace}\left[\ln \mathsf{C} + \mathsf{C}^{-1}\mathsf{D}\right] \; . \tag{14}$$

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

$$\mathsf{C}_{,i} \equiv \frac{\partial}{\partial\theta_i}\mathsf{C} \; . \tag{15}$$

Since C is a symmetric matrix for all values of the parameters, it is easy to see that all the derivatives $\mathsf{C}_{,i}$, $\mathsf{C}_{,ij}$, will also be symmetric matrices. Using the

matrix identities $(\mathsf{C}^{-1})_{,i} = -\mathsf{C}^{-1}\mathsf{C}_{,i}\,\mathsf{C}^{-1}$ and $(\ln \mathsf{C})_{,i} = \mathsf{C}^{-1}\mathsf{C}_{,i}$ (see exercises), we find

$$2\mathcal{L}_{,i} = \mathrm{Trace}\left[\mathsf{C}^{-1}\mathsf{C}_{,i} - \mathsf{C}^{-1}\mathsf{C}_{,i}\,\mathsf{C}^{-1}\mathsf{D} + \mathsf{C}^{-1}\mathsf{D}_{,i}\right] . \tag{16}$$

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

$$\begin{cases} \langle \mathsf{D} \rangle & = \mathsf{C} , \\ \langle \mathsf{D}_{,i} \rangle & = 0 , \\ \langle \mathsf{D}_{,ij} \rangle & = \boldsymbol{\mu}_{,i}\,\boldsymbol{\mu}^t_{,j} + \boldsymbol{\mu}_{,j}\,\boldsymbol{\mu}^t_{,i} . \end{cases} \tag{17}$$

Using this and (16), we obtain $\langle \mathcal{L}_{,i} \rangle = 0$. 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 (16), we obtain

$$\begin{aligned} 2\mathcal{L}_{,ij} = \mathrm{Trace}[& - \mathsf{C}^{-1}\mathsf{C}_{,i}\,\mathsf{C}^{-1}\mathsf{C}_{,j} + \mathsf{C}^{-1}\mathsf{C}_{,ij} \\ & + \mathsf{C}^{-1}(\mathsf{C}_{,i}\,\mathsf{C}^{-1}\mathsf{C}_{,j} + \mathsf{C}_{,j}\,\mathsf{C}^{-1}\mathsf{C}_{,i})\mathsf{C}^{-1}\mathsf{D} \\ & - \mathsf{C}^{-1}(\mathsf{C}_{,i}\,\mathsf{C}^{-1}\mathsf{D}_{,j} + \mathsf{C}_{,j}\,\mathsf{C}^{-1}\mathsf{D}_{,i}) \\ & - \mathsf{C}^{-1}\mathsf{C}_{,ij}\,\mathsf{C}^{-1}\mathsf{D} + \mathsf{C}^{-1}\mathsf{D}_{,ij}] . \end{aligned} \tag{18}$$

Substituting this and (17) into (7) and using the trace identity $\mathrm{Trace}[\mathsf{AB}] = \mathrm{Trace}[\mathsf{BA}]$, many terms drop out and the Fisher information matrix reduces to simply

$$\mathsf{F}_{ij} = \langle \mathcal{L}_{,ij} \rangle = \frac{1}{2}\mathrm{Trace}[\mathsf{C}^{-1}\mathsf{C}_{,i}\mathsf{C}^{-1}\mathsf{C}_{,j} + \mathsf{C}^{-1}\mathsf{M}_{ij}] , \tag{19}$$

where we have defined the matrix $\mathsf{M}_{ij} \equiv \langle \mathsf{D}_{,ij} \rangle = \boldsymbol{\mu}_{,i}\,\boldsymbol{\mu}^t_{,j} + \boldsymbol{\mu}_{,j}\,\boldsymbol{\mu}^t_{,i}$.

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. Figure 1 shows an example, where a prior from CMB experimental results has been added to a hypothetical 3D weak lensing survey.

For a good discussion of how to interpret likelihood plots in more than one dimension, see the Numerical Recipes books (e.g. [12]; available online at `www.nr.com`).