B Analysis Methods

Phil Marshall, Licia Verde, Hu Zhan

This chapter describes the statistical analysis methods that are used in previous chapters either to forecast LSST performance or as suggested to analyze LSST data. We start with an introductory review before moving on to some practical examples.

B.1 Basic Parameter Estimation

Very readable introductions to probabilistic data analysis are given by Sivia (1996), MacKay (2003) and Jaynes (2003); an introduction to the basics is given in this section. A single piece of experimental data is often presented in the form $x = x_0 \pm \sigma_0$, with x_0 being the result of the measurement and σ_0 the estimate of its uncertainty. This is shorthand for something like the statement "I believe that the quantity I am trying to measure, x, is most likely from my experiments to be x_0, but I could also believe that it was actually less than or greater than this by roughly σ_0." That is, the relation $x = x_0 \pm \sigma_0$ is a compressed version of the probability distribution (or probability density function, PDF) $\Pr(x_0|x, H)$, to be read as the probability of getting x_0 given assumptions about x and H. When written as a function of the model parameters, this PDF is referred to as the likelihood. Since our observed data come in probability distribution form, any conclusions we draw from them will necessarily be probabilistic in nature as well.

Traditionally there are two interpretations of probability: "Frequentist" and "Bayesian." For frequentists, probabilities are just frequencies of occurrence: $\mathcal{P} = n/N$ where n denotes the number of successes and N the total number of trials. Probability is then defined as the limit for the number of independent trials going to infinity. In the example above if one were to repeat the experiment an infinite number of times, then x will fall in the range $[x_0 - \sigma_0, x_0 + \sigma_0]$, say, 68% of the time. Bayesians instead interpret probability as a degree of belief in a hypothesis – a quantified version of the original statement above.

In cosmology, statistical analysis tends to be carried out in the Bayesian framework. It is easy to understand why: cosmic variance makes cosmologists only too aware of the limited information available to them. Only one realization of the CMB anisotropy and the large scale structure is accessible to our telescopes, and so while this is not a technical barrier to our happily simulating large numbers of fictitious universes in order to either compute or interpret our uncertainties, it is perhaps something of a psychological one, promoting the acceptance of the Bayesian notion of probability.

Bayesian cosmologists, seeking a steady point for their lever, assume that the observable universe is just one particular realization of a true underlying stochastic model of the Universe: the cosmological parameters of this model can be inferred from this one realization via the rules of probability.

Only if we could average all possible (unobservable) realizations of the underlying model could we recover the true values of the parameters with certainty – but since we can only observe one of the infinite possible realizations of it, we have to settle for probability distributions for the parameters of the underlying model instead.

This mental approach has the distinct advantage that it keeps cosmologists honest about the assumptions they are making, not only about the underlying world model, but also every other aspect of the data set they are attempting to model: systematic errors should, in principle, be already at the forefront of her mind! The catch is that to interpret probability as a degree of belief in an hypothesis, Bayesian cosmologists have to assume a probability distribution for the hypothesis itself. This step can be somewhat arbitrary and thus subjective: this is the age-old point of friction between frequentists and Bayesians. In this appendix we will use the Bayesian framework, in keeping with the tradition of cosmology. We will nevertheless try to point out where the "subjectivity" of being Bayesian is introduced (and where it is not).

After this aside, let us return to practicalities. The precise functional form of the likelihood is always unknown, and so an assumption must be made about it before any interpretation of the data can occur. This assumption forms part of a model for the data, which we denote by H, whilst x is a variable parameter of this model. More often than not, the physical nature of the object being studied is best understood in terms of some different parameter, θ, rather than x: in this case the model still allows the datum x_0 to be predicted, and describes how its probability is distributed through $\Pr(x_0|\theta, H)$. If more than one datum is available, and they came from independent attempted measurements of x, we can write the joint likelihood as

$$\Pr(x_0, x_1, x_2, \ldots|\theta, H) = \Pr(x_0|\theta, H)\Pr(x_1|\theta, H)\Pr(x_2|\theta, H)\ldots, \tag{B.1}$$

the product rule for combining independent probabilities. This makes clearer the distinction between the parameter θ and the data (which can be conveniently packaged into the vector $\mathbf{d}$ having components x_i). Indeed, a more complicated model for the data would make use of more than one parameter when predicting the data; these can be described by the parameter vector $\boldsymbol{\theta}$. The generalization of Equation B.1 to N_d independent data sets, $\{\mathbf{d}_j\}$, is then:

$$\Pr(\mathbf{d}|\boldsymbol{\theta}, H) = \prod_{j=1}^{N_d} \Pr(\mathbf{d}_j|\boldsymbol{\theta}, H). \tag{B.2}$$

Within a given model then, the likelihood $\Pr(\mathbf{d}|\boldsymbol{\theta}, H)$ can be calculated for any values of the model parameters $\boldsymbol{\theta}$. However, as outlined above, cosmologists want statistical inferences, i.e., we want to learn more about our model and its parameters from the data, by calculating the posterior distribution $\Pr(\boldsymbol{\theta}|\mathbf{d}, H)$. This distribution contains all the information about the model supplied by the data, as well as all the information we had about the model other than that provided by the data: this can be seen by applying the product rule of conditional probability to give Bayes' theorem,

$$\Pr(\boldsymbol{\theta}|\mathbf{d}, H) = \frac{\Pr(\mathbf{d}|\boldsymbol{\theta}, H)\Pr(\boldsymbol{\theta}|H)}{\Pr(\mathbf{d}|H)}. \tag{B.3}$$

The prior $\Pr(\boldsymbol{\theta}|H)$ encodes the additional information (this is where the subjectivity of the Bayesian approach comes in), and is a PDF normalized over the parameter space. The likelihood is also a frequentist quantity (without dependence on the prior) while the posterior is a

Bayesian construct. In practical applications of Bayesian parameter inference it is good practice therefore to check how much the reported result depend on the choice of prior: reliable results depend very weakly on the prior chosen. This is in fact a form of model comparison: for Bayesians, a complete data model consists of a parameter set *and* the prior PDF for those parameters: some priors are more appropriate than others. We discuss quantitative model comparison below: in this context it provides a way of recovering some objectivity in Bayesian analysis.

B.2 Assigning and Interpreting PDFs

As Equation B.3 shows, computing the probability distribution for a parameter (and hence measuring it) necessarily involves the assignment of a prior PDF for that parameter. There are two types of prior we can assign:

- **Uninformative priors,** such as uniform distributions in the parameter or its logarithm (the Jeffreys prior) are often assumed. Sometimes we genuinely know very little, and so minimizing the average information content of a prior PDF (or maximizing its entropy) makes sense. In other situations we do know something about a model parameter, but choose to assign an uninformative prior in order to investigate cleanly the information content of the data. Sometimes the reason given is to "give an unbiased result." This makes less sense, given that Bayesian inferences are biased by design – biased towards what is already known about the system.
- **Informative priors:** it is very rare to know *nothing* about a model and its parameters. An experiment has usually been carried out before, with different data! The best kind of prior PDF is the posterior PDF of a previous experiment – this is exactly equivalent to combining data sets in a joint analysis (Equation B.2 above).

Given suitably assigned priors and likelihoods then, the posterior distribution gives the probability of the parameter vector lying between $\boldsymbol{\theta}$ and $(\boldsymbol{\theta} + d\boldsymbol{\theta})$. This is the answer to the problem, the complete inference within the framework of the model. However, we typically need to present some compressed version of the posterior PDF: what should we do?

The probability distribution for a single parameter θ_N is given by marginalization over the other parameters,

$$\Pr\left(\theta_N|\mathbf{d}, H\right) = \int \Pr\left(\boldsymbol{\theta}|\mathbf{d}, H\right) d^{N-1}\boldsymbol{\theta}. \tag{B.4}$$

This is the addition rule for probabilities, extended to the continuous variable case[1]. This single parameter, one-dimensional marginalized posterior is most useful when the parameter is the only one of interest. In other cases we need to represent the posterior PDF and the parameter constraints that it describes in higher dimensions – although beyond two dimensions the posterior PDF becomes very difficult to plot.

The one-dimensional marginalized posterior PDFs can be further compressed into their means, or medians, or confidence intervals containing some fraction of the total probability – confidence

[1] Sometimes Equation B.4 is used with the posterior $\Pr\left(\boldsymbol{\theta}|\mathbf{d}, H\right)$ substituted by the likelihood. Even in this case a Bayesian step has been taken: a uniform prior is "hidden" in the parameter space "measure" $d^{N-1}\boldsymbol{\theta}$.

intervals need to be defined carefully as the integrals can be performed a number of different ways. However, note that the set of one-dimensional marginalized posterior means (or medians, etc.) need not itself represent a model that is a good fit to the data. The "best-fit" point is the position in parameter space where the likelihood function has a global maximum. This point is of most interest when the prior PDF is uninformative: in the case of uniform prior PDFs on all parameters, the peak of the likelihood coincides with the peak of the posterior PDF, but in general it does not. Moreover, the maximum likelihood model is necessarily the model most affected by the noise in the data – if any model "over-fits" the data, it is that one! Graphical displays of marginalized posterior PDFs remain the most complete way to present inferences; propagating the full posterior PDF provides the most robust estimates of individual parameters.

One class of parameters that are invariably marginalized over in the final analysis are the so-called nuisance parameters. The model H is a model for the data, not just the physical system of interest: often there are aspects of the experimental setup that are poorly understood, and so best included in the model as functions with free parameters with estimated prior PDFs. This procedure allows the uncertainty to be propagated into the posterior PDF for the interesting parameters. Examples of nuisance parameters might include: calibration factors, unknown noise or background levels, point spread function widths, window function shapes, supernova dust extinctions, weak lensing mean source redshifts, and so on. If a systematic error in an experiment is identified, parametrized and then that nuisance parameter marginalized over, then it can be said to have been translated into a statistical error (seen as a posterior PDF width): a not unreasonable goal is to translate all systematic errors into statistical ones.

B.3 Model Selection

While the goal of parameter estimation is to determine the posterior PDF for a model's parameters, perhaps characterized simply by the most probable or best-fit values and their errors, model selection seeks to distinguish between different models, which in general will have different sets of parameters. Simplest is the case of *nested* models, where the more complicated model has additional parameters, in addition to those in the simpler model. The simpler model may be interpreted as a particular case for the more complex model, where the additional parameters are kept fixed at some fiducial values. The additional parameters may be an indication of new physics, thus the question one may ask is: "would the experiment provide data with enough statistical power to require additional parameters and therefore to signal the presence of new physics if the new physics is actually the true underlying model?" Examples of this type of question are: "do the observations require a modification to general relativity as well as a universe dominated by dark energy?"(§ 15.3), or, "do the observations require a new species of neutrino?" (§ 15.2). However, completely disparate models, with entirely different parameter sets can also be compared using the Evidence ratio. In this case, in is even more important to assign realistic and meaningful prior PDFs to all parameters.

These questions may be answered in a Bayesian context by considering the Bayesian Evidence ratio, or Bayes factor, B:

$$B = \frac{\Pr(\mathbf{d}|H_1)}{\Pr(\mathbf{d}|H_2)}, \tag{B.5}$$

where H_1 and H_2 represent the two models being compared. The Bayes factor is related to the perhaps more desirable posterior ratio

$$\frac{\Pr(H_1|\mathbf{d})}{\Pr(H_2|\mathbf{d})} = \frac{\Pr(\mathbf{d}|H_1)}{\Pr(\mathbf{d}|H_2)} \frac{\Pr(H_1)}{\Pr(H_2)}. \tag{B.6}$$

by the ratio of model prior probabilities. The latter is not, in general straightforward to assign with differences of opinion between analysts common. However, the Bayes factor itself can be calculated objectively once H_1 and H_2 have been defined, and so is the more useful quantity to present.

If there is no reason to prefer one hypothesis over another other than that provided by the data, the ratio of the probabilities of each of the two models being true is just given by the ratio of evidences. Another way of interpreting a value of B greater than unity is as follows: if models H_1 and H_2 are still to be presented as equally probable after the experiment has been performed, then proponents of the lower-evidence model H_2 must be willing to offer odds of B to one against H_1. In practice, the Bayesian Evidence ratio can be used simply to say that "the probability of getting the data would be B times higher if model H_2 were true than if H_2 were." Indeed, Jeffreys (1961) proposed that $1 < \ln B < 2.5$ be considered as "substantial" evidence in favor of a model, $2.5 < \ln B < 5$ as "strong," and $\ln B > 5$ as "decisive." Other authors have introduced different terminology (e.g., Trotta 2007a).

The evidence $\Pr(\mathbf{d}|H)$ is the normalization of the posterior PDF for the parameters, and so is given by integrating the product of the likelihood and the prior over all N parameters:

$$\Pr(\mathbf{d}|H) = \int \Pr(\mathbf{d}|\boldsymbol{\theta}, H) \Pr(\boldsymbol{\theta}|H)\, d^N\boldsymbol{\theta}. \tag{B.7}$$

There is ample literature on applications of Bayesian Evidence ratios in cosmology (e.g. Jaffe 1996; Hobson et al. 2002; Saini et al. 2004; Liddle et al. 2006; Marshall et al. 2006; Parkinson et al. 2006; Mukherjee et al. 2006a; Pahud et al. 2006; Szydłowski & Godłowski 2006b,a; Trotta 2007b; Pahud et al. 2007). The evidence calculation typically involves computationally expensive integration Skilling (2004); Trotta (2007a); Beltrán et al. (2005); Mukherjee et al. (2006a,b); Parkinson et al. (2006); however, it can often be approximated just as the model parameter posterior PDF can. For example, Heavens et al. (2007) shows how, by making simplifying assumptions in the same spirit of Fisher's analysis (Fisher 1935), one can compute the expected evidence for a given experiment, in advance of taking any data, and forecast the extent to which an experiment may be able to distinguish between different models. We implement this in § 15.2 and § 15.3. In § 15.2 we consider the issue of deviations from the standard number of three neutrino species. The simplest model has three neutrino species, but effectively this number can be changed by the existence of a light particle that does not couple to electrons, ions or photons, or by the decay of dark matter particles, or indeed any additional relativistic particle. Given the observables and errors achievable from a survey with given specifications, we use the evidence in order to address the issue of how much different from the standard value the number of neutrino species should be such that the alternative model should be favored over the reference model.

In § 15.3 we also employ the Bayesian evidence: this time the reference model is a cold dark matter + dark energy model, where gravity is described by General Relativity (GR). In the alternative model, GR is modified so that the growth of cosmological structure is different. Again, given the observables and errors achievable from a survey with given specifications, we use the evidence to quantify how different from the GR prediction the growth of structure would have to be such that the alternative model should be preferred.

B.4 PDF Characterization

The conceptually most straightforward way to carry out parameter inference (and model selection) is to tabulate the posterior PDF $\Pr(\boldsymbol{\theta}|\mathbf{d}, H)$ on a suitable grid, and normalize it via simple numerical integration. This approach is unlikely to be practical unless the number of parameters is very small and the PDF is very smooth. The number of function evaluations required increases exponentially with the dimensionality of the parameter space; moreover, the following marginalization integrals will all be correspondingly time-consuming. In this section we consider two more convenient ways to characterize the posterior PDF — the multivariate Gaussian (or Laplace) approximation, and Markov Chain Monte Carlo sampling.

B.4.1 The Laplace Approximation

By the central limit theorem, the product of a set of convex functions tends to the Gaussian functional form in the limit of large set size (Jaynes 2003); the posterior probability distribution of Equation B.3 often fits this bill, suggesting that the Gaussian distribution is likely to be a good approximation to the posterior density. Approximating probability distributions with Gaussians is often referred to as the Laplace approximation (see e.g. Sivia 1996; MacKay 2003).

In one dimension, a suitable Gaussian can be found by Taylor expansion about the peak position θ_0 of the logarithm of the posterior PDF $P(\theta)$ (where the conditioning on the data and the model have been dropped for clarity):

$$\log P(\theta) \approx \log P(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 \frac{d^2 P}{d\theta^2}\bigg|_{\theta_0}. \tag{B.8}$$

Exponentiating this expression gives the Gaussian approximation to the function,

$$g(\theta) \approx P(\theta_0) \exp\left[-\frac{(\theta - \theta_0)^2}{2\sigma^2}\right]. \tag{B.9}$$

The width σ of this Gaussian satisfies the following relation:

$$\frac{d^2 \log P}{d\theta^2}\bigg|_{\theta_0} = -\frac{1}{\sigma^2}. \tag{B.10}$$

The extension of this procedure to multivariate distributions is straightforward: instead of a single variance σ^2, an $N \times N$ covariance matrix C is required, such that the posterior approximation is

$$g(\boldsymbol{\theta}) = P(\boldsymbol{\theta}_0) \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathsf{C}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right], \tag{B.11}$$

and the covariance matrix has components

$$\left(\mathsf{C}^{-1}\right)_{ij} = -\left.\frac{\partial^2 \log P}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}_0} . \tag{B.12}$$

(This matrix is very unlikely to be diagonal – correlations, or degeneracies, between parameters are encoded in its off-diagonal elements.) The problem is now reduced to finding (numerically) the peak of the log-posterior, and its second derivatives at that point. When the data quality is good, one may expect the individual datum likelihoods to be already quite convex, giving a very peaky unimodal posterior: in this case the Gaussian approximation is likely to be both accurate, and more quickly and easily located. More commonly, techniques such as simulated annealing may be necessary when finding the maximum of complex functions of many parameters; in this case a Gaussian may not be such a good approximation anyway.

B.4.2 Fisher Matrices

The Fisher information matrix (Fisher 1935) is widely used for forecasting survey performance and for identifying dominant systematic effects (see e.g., Albrecht et al. 2009). The Fisher matrix formalism is very closely related to the Laplace approximation to the parameter posterior described above. The discussion that follows may seem unconventional to those more familiar with its frequentist origins and presentation: our aim is to show how the formalism has been adapted to modern Bayesian cosmology.

The Fisher matrix was originally defined to be

$$F_{\alpha\beta} = -\left\langle \frac{\partial^2 \ln P(\boldsymbol{x}|\boldsymbol{q})}{\partial q_\alpha \partial q_\beta} \right\rangle , \tag{B.13}$$

where $\boldsymbol{x}$ is a data vector, $\boldsymbol{q}$ is the vector of model parameters, and $\langle \ldots \rangle$ denotes an ensemble average. If the prior PDFs are non-uniform, we must replace the likelihood $P(\boldsymbol{x}|\boldsymbol{q})$ by the posterior $P(\boldsymbol{q}|\boldsymbol{x})$. In any case, we recognize the Laplace approximation and identify (by comparison with Equation B.12) the ensemble average covariance matrix of the inferred parameters as $\boldsymbol{F}^{-1}$.

This estimate of the forecast parameter uncertainties really corresponds to the best case scenario, as dictated by the Cramer-Rao theorem. More specifically, the estimated error of the parameter q_α is $\sigma(q_\alpha) \geq (F_{\alpha\alpha})^{-1/2}$ if all other parameters are known precisely, or $\sigma(q_\alpha) \geq [(\boldsymbol{F}^{-1})_{\alpha\alpha}]^{1/2}$ if all the parameters are estimated from the data.

Cosmological applications of the Fisher matrix were introduced by Jungman et al. (1996); Vogeley & Szalay (1996); Tegmark et al. (1997); Tegmark (1997). The key is to identify the correct likelihood function $P(\boldsymbol{x}|\boldsymbol{q})$ (which is anyway crucial for all inference techniques). However, the Fisher matrix analysis has a further limitation due to the Gaussian approximation of $P(\boldsymbol{q}|\boldsymbol{x})$ with respect to $\boldsymbol{q}$: the likelihood, priors and indeed choice of parametrization need to be such that this approximation is a good one. Usual practice is to approximate the likelihood function as Gaussian, and assert either Gaussian or uniform priors (both of which guarantee the Gaussianity of the posterior PDF).

If we approximate the likelihood function by a Gaussian distribution then,

$$P(\boldsymbol{x}|\boldsymbol{q}) = \frac{1}{(2\pi)^{N/2}\det[\boldsymbol{C}(\boldsymbol{q})]}\exp\left\{[\boldsymbol{x}-\bar{\boldsymbol{x}}(\boldsymbol{q})]^{\mathrm{T}}\,\frac{\boldsymbol{C}^{-1}(\boldsymbol{q})}{2}\,[\boldsymbol{x}-\bar{\boldsymbol{x}}(\boldsymbol{q})]\right\}, \tag{B.14}$$

where N is the dimension of the observables $\boldsymbol{x}$, $\bar{\boldsymbol{x}}(\boldsymbol{q})$ is the ensemble average of $\boldsymbol{x}$, $\boldsymbol{C}(\boldsymbol{q}) = \langle(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathrm{T}}\rangle$ is the covariance of $\boldsymbol{x}$. The Fisher matrix is then (Tegmark et al. 1997)

$$F_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{C}^{-1}\frac{\partial\boldsymbol{C}}{\partial q_\alpha}\boldsymbol{C}^{-1}\frac{\partial\boldsymbol{C}}{\partial q_\beta}\right) + \frac{\partial\bar{\boldsymbol{x}}}{\partial q_\alpha}\boldsymbol{C}^{-1}\frac{\partial\bar{\boldsymbol{x}}}{\partial q_\beta}, \tag{B.15}$$

where we have dropped the variables $\boldsymbol{q}$ in $\boldsymbol{C}$ and $\bar{\boldsymbol{x}}$ for clarity. To avoid confusion, we note that $\boldsymbol{C}$ is the covariance matrix of the data $\boldsymbol{x}$, whereas $\boldsymbol{F}^{-1}$ is the covariance matrix of the parameters $\boldsymbol{q}$.

In the Gaussian approximation, marginalization, and moment-calculating integrals are analytic. Independent Fisher matrices are additive; a Gaussian prior on q_α, $\sigma_{\mathrm{P}}(q_\alpha)$, can be introduced via $F^{\mathrm{new}}_{\alpha\alpha} = F_{\alpha\alpha} + \sigma_{\mathrm{P}}^{-2}(q_\alpha)$. A Fisher matrix of the parameters $\boldsymbol{q}$ can be projected onto a new set of parameters $\boldsymbol{p}$ via

$$F^{\mathrm{new}}_{\mu\nu} = \sum_{\alpha,\beta}\frac{\partial q_\alpha}{\partial p_\mu}F_{\alpha\beta}\frac{\partial q_\beta}{\partial p_\nu}. \tag{B.16}$$

Fixing a parameter is equivalent to striking out its corresponding row and column in the Fisher matrix. To obtain a new Fisher matrix after marginalizing over a parameter, one can strike out the parameter's corresponding row and column in the original covariance matrix (i.e., the inverse of the original Fisher matrix) and then invert the resulting covariance matrix[2].

B.4.3 Examples

At this point, we give two worked examples from observational cosmology, describing the data model and Fisher matrix forecasts of parameter uncertainties.

Example 1: Type Ia Supernovae
For SNe, the observables are their peak magnitudes in a certain band

$$m_i = \bar{m}(\boldsymbol{q}, z_i) + n_i, \tag{B.17}$$

where the subscript i labels each SN, $\bar{m}$ is the mean value of the SN peak magnitude at redshift z_i, the parameters $\boldsymbol{q}$ include both cosmological and nuisance parameters, and n_i represents the observational noise and intrinsic scatter of the peak magnitude. The mean peak magnitude is given by

$$\bar{m}_i = M + 5\log\left[D_{\mathrm{L}}(w_0, w_{\mathrm{a}}, \Omega_{\mathrm{m}}, \Omega_{\mathrm{k}}, h, \ldots, z)\right] + \text{evolution terms} + \text{const}, \tag{B.18}$$

where M[3] is the mean absolute peak magnitude at $z = 0$, D_{L} is the luminosity distance, and the evolution terms account for a possible drift of the mean absolute peak magnitude with time. In a number of forecasts, the evolution terms are simply model with a quadratic function $az + bz^2$ (e.g., Albrecht et al. 2006; Knox et al. 2006; Zhan et al. 2008).

[2] For a better numerical treatment, see Albrecht et al. (2009).

[3] M is degenerate with the Hubble constant.

We assume that the scatter n_i 1) does not depend on cosmology or redshift, 2) is uncorrelated with each other, and 3) is normally distributed, i.e.,

$$\begin{aligned} \langle (m_i - \bar{m}_i)(m_j - \bar{m}_j) \rangle &= \sigma_m^2 \delta_{ij}^{\mathrm{K}}, & \text{(B.19)} \\ P(\boldsymbol{m}|\boldsymbol{q}) = \Pi_i P(m_i|\boldsymbol{q}, \sigma_m) &= \Pi_i \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left[-\frac{(m_i - \bar{m}_i)^2}{2\sigma_m^2}\right], & \text{(B.20)} \end{aligned}$$

where δ_{ij}^{K} is the Kronecker delta function, and $\sigma_m \sim 0.15$, then Fisher matrix reduces to

$$F_{\alpha\beta} = \Sigma_i \frac{\partial \bar{m}_i}{\partial q_\alpha} \frac{1}{\sigma_m^2} \frac{\partial \bar{m}_i}{\partial q_\beta}. \tag{B.21}$$

With a photometric redshift SN sample, the Fisher matrix has to be integrated over the photometric redshift error distribution (Zhan et al. 2008):

$$\begin{aligned} F_{\alpha\beta} &= \frac{1}{\sigma_{\mathrm{m}}^2} \int n_{\mathrm{p}}(z_{\mathrm{p}}) \frac{\partial \bar{m}_{\mathrm{p}}(z_{\mathrm{p}})}{\partial q_\alpha} \frac{\partial \bar{m}_{\mathrm{p}}(z_{\mathrm{p}})}{\partial q_\beta} dz_{\mathrm{p}} & \text{(B.22)} \\ \bar{m}_{\mathrm{p}} &= \int [5 \log D_{\mathrm{L}}(w_0, w_{\mathrm{a}}, \Omega_{\mathrm{m}}, \Omega_{\mathrm{k}}, h, \ldots, z) + M + \text{evol. terms} + \text{const.}]\, p(z|z_{\mathrm{p}}) dz, \end{aligned}$$

the subscript p signifies photometric redshift space, $n_{\mathrm{p}}(z_{\mathrm{p}})$ is the SN distribution in photometric redshift space, and $p(z|z_{\mathrm{p}})$ is the probability density of a SN at z given its photometric redshift z_{p}.

Example 2: Gaussian Random Fields

The power spectrum is the covariance of the Fourier modes of the field. For an isotropic field, the modes are uncorrelated, i.e.,

$$\langle \hat{\delta}(\boldsymbol{k}) \hat{\delta}^*(\boldsymbol{k}') \rangle = P(k)(2\pi)^3 \delta^{\mathrm{D}}(\boldsymbol{k} - \boldsymbol{k}'), \tag{B.23}$$

where $P(k)$ is the power spectrum, and $\delta^{\mathrm{D}}(\boldsymbol{k} - \boldsymbol{k}')$ is the Dirac delta function. By definition, the modes $\hat{\delta}(\boldsymbol{k})$ have zero mean. Since surveys are limited by volume, the wavenumbers must be discrete. For a cubic survey of volume $V = L^3$, we have

$$\langle \hat{\delta}(\boldsymbol{k}) \hat{\delta}^*(\boldsymbol{k}') \rangle = P(k) V \delta_{\boldsymbol{n},\boldsymbol{n}'}^{\mathrm{K}}, \tag{B.24}$$

where $\boldsymbol{k} = 2\pi\boldsymbol{n}/L$, and $\boldsymbol{n} = (n_1, n_2, n_3)$ with integer n_is running from $-\infty$ to ∞. If the density field is discretized, e.g., on a grid, then n_is are limited by the Nyquist frequency. For convenience, we use $\boldsymbol{k}$ and $\boldsymbol{n}$ interchangeably, with the understanding that $\boldsymbol{k}$ is discrete. If the power spectrum is calculated based on discrete objects, e.g., galaxies, then we have

$$P_{\mathrm{g}}(k) = P(k) + n_{\mathrm{g}}^{-1}, \tag{B.25}$$

where n_{g} is the galaxy number density.

For a Gaussian random field sampled by galaxies, the modes are normally distributed and are completely characterized by the power spectrum $P_{\mathrm{g}}(k)$. Using the Fourier modes (rather than the power spectrum) as observables, we obtain the Fisher matrix using Equation B.15 and Equation B.24

$$F_{\alpha\beta} = \frac{1}{2} \Sigma_{\boldsymbol{n}} \frac{\partial \ln P_{\mathrm{g}}(n)}{\partial q_\alpha} \frac{\partial \ln P_{\mathrm{g}}(n)}{\partial q_\beta}, \tag{B.26}$$

where the summation runs over all available modes. When the survey volume is sufficiently large, one can replace the summation in Equation B.26 with an integral (Tegmark 1997)

$$F_{\alpha\beta} = \frac{V}{2}\int \frac{\partial \ln P_{\rm g}(k)}{\partial q_\alpha}\frac{\partial \ln P_{\rm g}(k)}{\partial q_\beta}\frac{k^2 dk}{4\pi^2}. \tag{B.27}$$

In terms of angular power spectra, the Fisher matrix becomes

$$F_{\alpha\beta} = f_{\rm sky}\sum_{\ell}\frac{2\ell+1}{2}{\rm Tr}\left[\boldsymbol{P}^{-1}(\ell)\frac{\partial \boldsymbol{P}(\ell)}{\partial q_\alpha}\boldsymbol{P}^{-1}(\ell)\frac{\partial \boldsymbol{P}(\ell)}{\partial q_\beta}\right], \tag{B.28}$$

where $f_{\rm sky}$ is the fraction of sky covered by the survey, and for each multipole ℓ, $\boldsymbol{P}(\ell)$ is a matrix of power spectra between pairs of redshift bins (see e.g, Equation 13.6 and Equation 15.3).

B.4.4 Sampling Methods

Probability distributions calculated on high-dimensional regular grids are rather unwieldy. A Gaussian approximation allows integrals over the posterior to be performed analytically – but may not provide sufficient accuracy especially if the PDF is not unimodal.

A far more useful characterization of the posterior PDF is a list of samples drawn from the distribution. By definition, the number density of these samples is proportional to the probability density, such that (given enough samples) a smoothed histogram is a good representation of the probability density function. Each histogram bin value is an integral over this PDF, as are all other inferences. By sampling from the distribution, these integrals are calculated by Monte Carlo integration (as opposed to the simple summation of a gridding algorithm). Marginal distributions are trivial to calculate – the histogram needs only to be constructed in the required dimensions, usually just one or two. Samples are also useful for the simple reason that they represent acceptable fits: visualization of the model corresponding to each sample can provide much insight into the information content of the data.

The problem is now how to draw samples from a general PDF $P(\boldsymbol{\theta})$. The Metropolis-Hastings algorithm (and various derivatives) provides a flexible solution to this problem: see the books by e.g. Gilks et al. (1996); Ó Ruanaidh & Fitzgerald (1996); Neal (1993); MacKay (2003) for good introductions. This is the basic Markov chain Monte Carlo method, and works as follows. A Markov chain is a series of random variables (specifically representing points in a parameter space) whose values are each determined only by the previous point in the series. Generation of a Markov chain proceeds as follows: a candidate sample point is drawn from a suitably chosen “proposal density” $S(\boldsymbol{\theta}', \boldsymbol{\theta})$, and then accepted with probability $A(\boldsymbol{\theta}', \boldsymbol{\theta})$ – if not accepted, the current sample is repeated to preserve the invariance of the target distribution. In the Metropolis-Hastings algorithm, the acceptance probability is given by

$$A(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left[1, \frac{P(\boldsymbol{\theta})}{P(\boldsymbol{\theta}')}\right], \tag{B.29}$$

provided the proposal distribution S is symmetric about the previous sample point.

In other words, we accept the new sample if the probability density at that point is higher, and otherwise accept it with probability equal to the ratio of new to old probability densities. Note that since the sampling procedure depends only on a probability ratio; the normalization of the PDF need not be known: this is just the situation we find ourselves in when analyzing data, able only to calculate the unnormalized product of likelihood and prior.

As seen in the previous paragraphs, the basic MCMC algorithm is very simple; whilst it guarantees that the output list of sample points will have been drawn from the target density P, that is not the same as fully sampling the distribution in a finite time. Consequently, the computational challenge lies in the choice of proposal density S. If S is too compact, the chains take too long to explore the parameter space; too broad, and the sample rejection rate becomes very high as too much time is spent testing regions of low likelihood.

Lewis & Bridle (2002) provide a useful primer to the use of MCMC in cosmological parameter estimation, and in particular CMB analysis. In the next sections we show some example sampled PDFs, and then outline some common problems encountered when sampling.

Example: WMAP5

The WMAP team provide their cosmological parameter inferences in Markov Chain form, downloadable from their website[4] (see the papers by Dunkley et al. 2009; Komatsu et al. 2009, for details). In Figure B.1 we display posterior PDFs in a 4-dimensional cosmological parameter space by plotting the WMAP team's MCMC samples, marginalizing by projecting the samples onto two different planes, and color-coding them by the samples' Hubble constant values in order to visualize this third dimension. In the top row, the likelihood function is for the WMAP 5th year data set alone, while in the second row the likelihood for the "SN all" combined supernova type Ia data set has been multiplied in.

MCMC Sampling Issues

MCMC sampling is often the preferred way to approximate a posterior PDF: the CPU time taken scales (in principle) only linearly with the number of parameter space dimensions (as opposed to exponentially in the case of brute-force gridding), it provides (in principle) accurate statistical uncertainties that take into account the various (often non-linear) parameter degeneracies, and avoids (in principle) the false maxima in the PDF that can cause Laplace approximation maximizers either to need restarting in multiple locations, or worse, to give misleading results.

However, MCMC sampling can be affected by a number of problems. We give a very brief overview here and refer the reader to the cited textbooks for more information. As the number of dimensions increases, finding isolated sharp peaks in the very large volumes involved is a particularly difficult problem. "Cooling" the process (starting by sampling from the prior, and gradually increasing the weight of the likelihood during a "burn-in" phase) can help alleviate this – this also allows evidence estimation via "thermodynamic integration" (Ó Ruanaidh & Fitzgerald 1996). These burn-in samples are to be discarded.

It also gets progressively more difficult to move away from a false maximum: proposal distributions that are too broad can lead to very high sample rejection rates and low chain mobility. Similarly

[4] http://lambda.gsfc.nasa.gov/

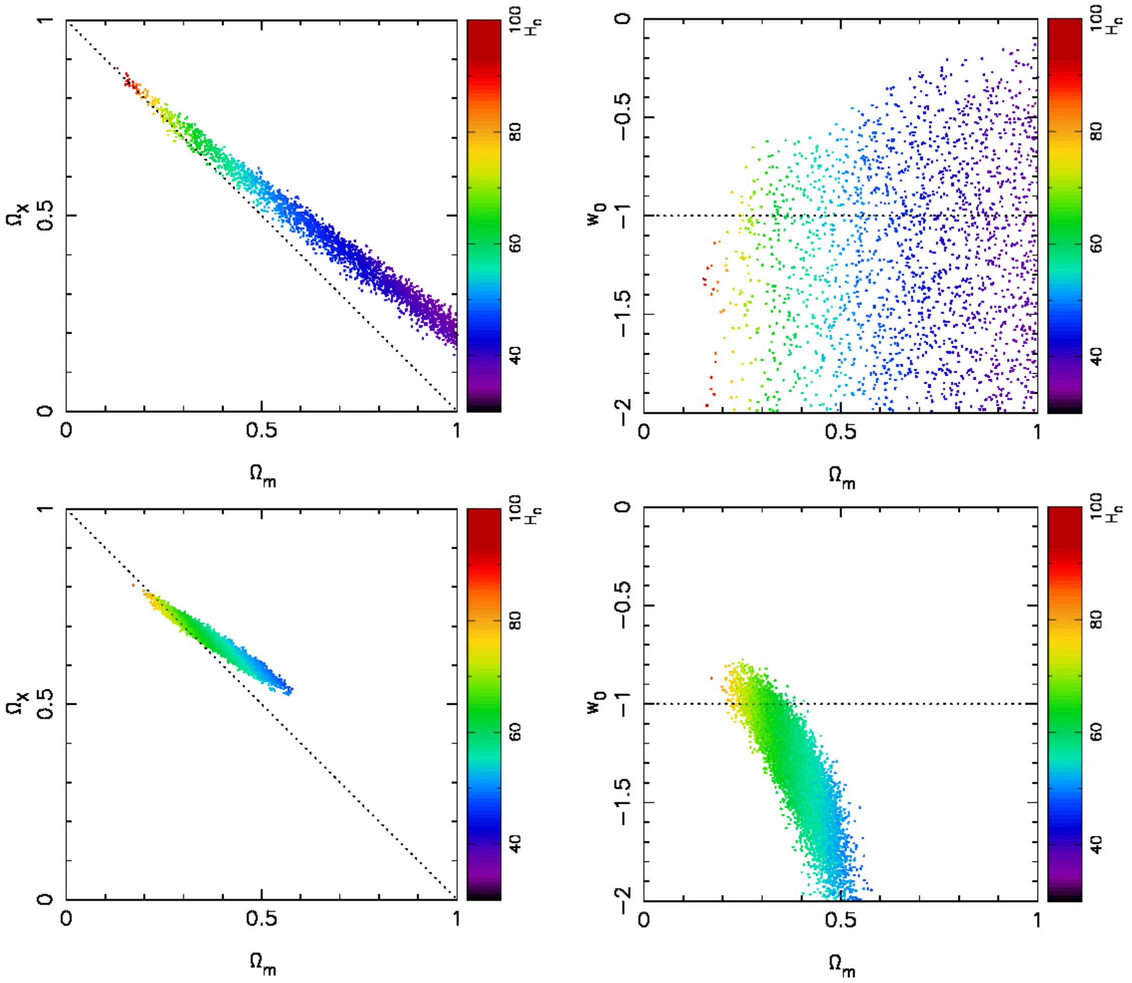

Figure B.1: Marginalized posterior PDFs for cosmological parameters given the WMAP 5 year data set, represented by MCMC sample density. *Top*: $\Pr(\Omega_{\rm m}, \Omega_X|\text{WMAP5 only})$ (left) and $\Pr(\Omega_{\rm m}, w_0|\text{WMAP5 only})$ (right). *Bottom*: $\Pr(\Omega_{\rm m}, \Omega_X|\text{WMAP5} + \text{SNe})$ (left) and $\Pr(\Omega_{\rm m}, w_0|\text{WMAP5} + \text{SNe})$ (right). In each row, uniform priors were assumed for $\Omega_{\rm m}$, Ω_X (the dark energy density parameter), w_0 (the non-evolving dark energy equation of state) and H_0. Samples are color-coded by H_0 to allow a third dimension to be visualized. The dashed lines are loci representing universes with flat geometry amd a cosmological constant. The Markov chains in the two rows contain different numbers of samples.

very narrow degeneracies also lead to high sample rejection rates. In these cases, the design of the proposal distribution is key! One partial solution is to re-parametrize such that the degeneracies are not so pronounced: a simple example is in CMB analysis, where working with $\omega_b = \Omega_b h^2$ instead of Ω_b removes a particularly pronounced “banana” degeneracy. However, the prior PDFs need especially careful attention in this case, since a uniform prior in A is never a uniform prior in B(A).

To increase efficiency, updating the covariance matrix of the proposal distribution as sampling proceeds has some considerable appeal (arising from the intuition that the best proposal distribution must be close to the target PDF itself), but the updating must be done carefully to preserve detailed balance in the chains. Finally, how do you know when you are finished? Various convergence tests on the chains have been proposed (e.g., Gelman & Rubin 1992); Dunkley et al. look at the chain power spectrum to check for unwanted correlations. In general, it is usually found that running multiple parallel Markov chains allow more tests and provide greater confidence.

Importance Sampling

We can incorporate new information into an MCMC inference by importance-sampling the posterior distribution (see e.g., Lewis & Bridle 2002, for an introduction). This procedure allows us to calculate integrals (such as means and confidence limits) over the updated posterior PDF, P_2, by re-weighting the samples drawn from the original PDF, P_1. For example, the posterior mean value of a parameter x:

$$\begin{aligned} \langle x \rangle_2 &= \int x \cdot P_2(x)\, dx, \\ &= \int x \frac{P_2(x)}{P_1(x)} \cdot P_1(x)\, dx. \end{aligned} \tag{B.30}$$

By weighting the samples from P_1 by the ratio $P_2(x)/P_1(x)$, we can emulate a set of samples drawn directly from P_2. It works most efficiently when P_1 and P_2 are quite similar, and fails if P_1 is zero-valued over some of the range of P_2 or if the sampling of P_1 is too sparse. In many cases though, it provides an efficient way to explore the effects of both additional likelihoods and alternative priors.

As an example, Figure B.2 shows the same marginalized posterior PDF from the WMAP 5-year data set as in Figure B.1, but after importance sampling using the Hubble constant measurement of $H_0 = 74.2 \pm 3.6 \mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ by Riess et al. (2009). To make this plot, we interpret the Riess et al. measurement as the Gaussian PDF $\Pr(H_0|\mathrm{Riess})$, and then write the updated posterior PDF arising from a joint analysis of the WMAP 5-year data and the Riess et al. data as follows:

$$\Pr(H_0, \boldsymbol{q}|\mathrm{WMAP5,\ Riess}) \propto \Pr(\mathrm{WMAP5}|H_0, \boldsymbol{q}) \cdot \Pr(H_0|\mathrm{Riess}) \Pr(\boldsymbol{q}) \tag{B.31}$$

$$= \Pr(H_0, \boldsymbol{q}|\mathrm{WMAP5\,only}) \cdot \frac{\Pr(H_0|\mathrm{Riess})}{\Pr(H_0)}. \tag{B.32}$$

Here, for clarity, the cosmological parameters other than H_0 are denoted by the “vector” $\boldsymbol{q} = \{Om, \Omega_X, w_0, \ldots\}$.

In the second line, we have substituted the original posterior PDF, $\Pr(H_0, \boldsymbol{q}|\mathrm{WMAP5\,only})$: from this equation it is clear that the weight, $P_2(x)/P_1(x)$, from Equation B.30 is just given by the value of the PDF, $\Pr(H_0|\mathrm{Riess})$, if (as was the case) the original prior on H_0 was uniform. To make the plots, each sample was added to a two-dimensional histogram according to its weight; the histogram was then minimally smoothed and contours computed.

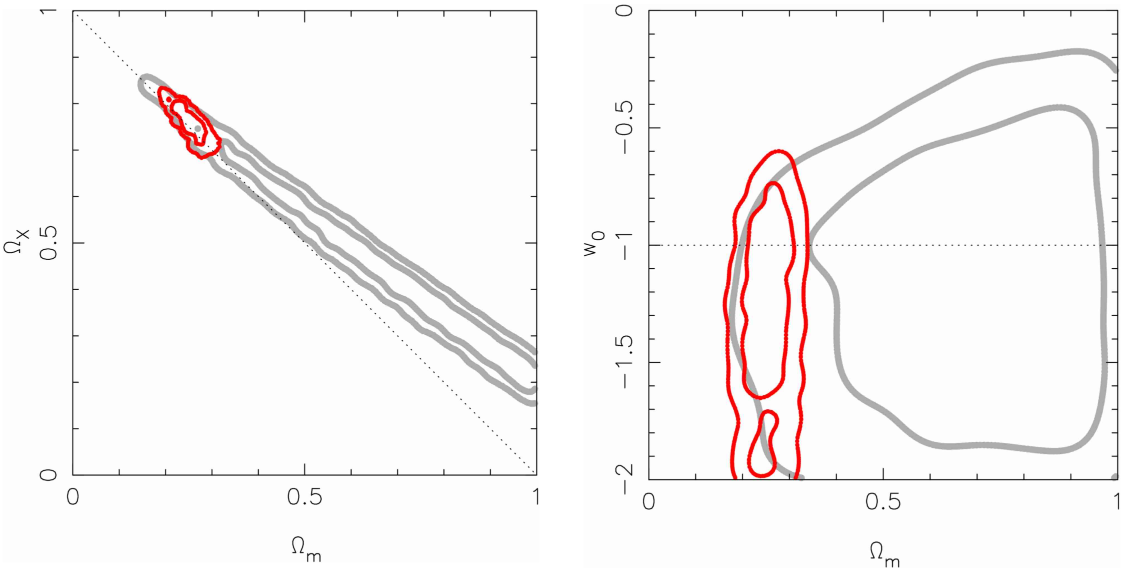

Figure B.2: Marginalized posterior PDFs for cosmological parameters given the WMAP 5 year data set and the Riess et al. (2009) Hubble constant measurement, obtained by importance-sampling (red curves). *Left*: $\Pr(\Omega_{\rm m}, \Omega_X|{\rm WMAP5, Riess})$ and *right*: $\Pr(\Omega_{\rm m}, w_0|{\rm WMAP5, Riess})$. The contours shown contain 68% and 95% of the integrated posterior probability: the gray contours in the background show the PDFs without the Riess et al Hubble constant constraint.

References

Albrecht, A. et al., 2009, ArXiv e-prints, 0901.0721

—, 2006, ArXiv Astrophysics e-prints, astro-ph/0609591

Beltrán, M., García-Bellido, J., Lesgourgues, J., Liddle, A. R., & Slosar, A., 2005, *Phys. Rev. D*, 71, 063532

Dunkley, J. et al., 2009, *ApJS*, 180, 306

Fisher, R. A., 1935, *J. Roy. Stat. Soc.*, 98, 39

Gelman, A., & Rubin, D. B., 1992, *Statistical Science*, 7, 457

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J., 1996, Markov-Chain Monte-Carlo In Practice. Cambridge: Chapman and Hall

Heavens, A. F., Kitching, T. D., & Verde, L., 2007, *MNRAS*, 380, 1029

Hobson, M. P., Bridle, S. L., & Lahav, O., 2002, *MNRAS*, 335, 377

Jaffe, A., 1996, *ApJ*, 471, 24

Jaynes, E., 2003, Probability Theory: The Logic of Science. Cambridge: CUP

Jeffreys, H., 1961, Theory of Probability. Oxford University Press

Jungman, G., Kamionkowski, M., Kosowsky, A., & Spergel, D. N., 1996, *Phys. Rev. D*, 54, 1332

Knox, L., Song, Y.-S., & Zhan, H., 2006, *ApJ*, 652, 857

Komatsu, E. et al., 2009, *ApJS*, 180, 330

Lewis, A., & Bridle, S., 2002, *Phys. Rev. D*, 66, 103511

Liddle, A. R., Mukherjee, P., Parkinson, D., & Wang, Y., 2006, *Phys. Rev. D*, 74, 123506

MacKay, D., 2003, Information Theory, Inference and Learning Algorithms. Cambridge: CUP

Marshall, P., Rajguru, N., & Slosar, A., 2006, *Phys. Rev. D*, 73, 067302

Mukherjee, P., Parkinson, D., Corasaniti, P. S., Liddle, A. R., & Kunz, M., 2006a, *MNRAS*, 369, 1725

Mukherjee, P., Parkinson, D., & Liddle, A. R., 2006b, *ApJL*, 638, L51

Neal, R., 1993, Probabilistic inference using markov chain monte carlo methods. Tech. Rep. CRG-TR-93-1, Department of Computer Science, University of Toronto

Ó Ruanaidh, J., & Fitzgerald, W., 1996, Numerical Bayesian Methods Applied to Signal Processing. New York: Springer-Verlag

Pahud, C., Liddle, A. R., Mukherjee, P., & Parkinson, D., 2006, *Phys. Rev. D*, 73, 123524

—, 2007, *MNRAS*, 381, 489

Parkinson, D., Mukherjee, P., & Liddle, A. R., 2006, *Phys. Rev. D*, 73, 123523

Riess, A. G. et al., 2009, *ApJ*, 699, 539

Saini, T. D., Weller, J., & Bridle, S. L., 2004, *MNRAS*, 348, 603

Sivia, D., 1996, Data Analysis: A Bayesian Tutorial. Oxford: OUP

Skilling, J., 2004, Available at. http://www.inference.phy.cam.ac.uk/bayesys

Szydłowski, M., & Godłowski, W., 2006a, *Phys. Lett. B*, 639, 5

—, 2006b, *Phys. Lett. B*, 633, 427

Tegmark, M., 1997, *Phys. Rev. Lett.*, 79, 3806

Tegmark, M., Taylor, A. N., & Heavens, A. F., 1997, *ApJ*, 480, 22

Trotta, R., 2007a, *MNRAS*, 378, 72

—, 2007b, *MNRAS*, 378, 819

Vogeley, M. S., & Szalay, A. S., 1996, *ApJ*, 465, 34

Zhan, H., Wang, L., Pinto, P., & Tyson, J. A., 2008, *ApJL*, 675, L1