Chapter 3

Bayesian cosmostatistics

Contents

> "A previous acquaintance with probability and statistics is not necessary; indeed, a certain amount of innocence in this area may be desirable, because there will be less to unlearn."
> — Edwin Thompson Jaynes (2003), *Probability Theory: The Logic of Science*

Abstract

In this chapter, essential concepts of Bayesian statistics in the context of cosmological data analysis are presented. We discuss motivations for seeing probabilistic calculations as an extension of ordinary logic and justify the use of a prior in an experimental learning process by referring to the "no-free lunch theorem". This chapters also reviews parameter inference, model comparison, and contains a brief introduction to the subject of Markov Chain Monte Carlo methods.

This chapter aims at introducing the necessary background in Bayesian probability theory for presenting the BORG algorithm in chapter 4 and applications in the following chapters. A much more complete picture can be found in the reference book of Gelman *et al.* (2013). For introductions to Bayesian statistics in a cosmological context, see Hobson (2010) and the reviews or lecture notes by Trotta (2008); Heavens (2009); Verde (2010); Leclercq, Pisani & Wandelt (2014).

This chapter is organized as follows. Section 3.1 is a general introduction on plausible reasoning. Basic concepts and definitions used in Bayesian statistics are presented in section 3.2. In section 3.3, we discuss standard statistical inference problems. Finally, section 3.4 is summarizes the basics of Markov Chain Monte Carlo methods.

3.1 Introduction: plausible reasoning

When discussing statistical data analysis, two different points of view are traditionally reviewed and opposed: the frequentist (see e.g. Kendall & Stuart, 1968) and the Bayesian approaches. In this author's experience,

arguments for or against each of them are generally on the level of a philosophical or ideological position, at least among cosmologists in 2015. Before criticizing this controversy, somewhat dated to the 20th century, and stating that more recent scientific work suppresses the need to appeal to such arguments, we report the most common statements encountered.

3.1.1 On the definition of probability

Frequentist and Bayesian statistics differ in the epistemological interpretation of probability and their consequences for testing hypotheses and comparing models. First and foremost, the methods differ on the understanding of the concept of the probability $\mathcal{P}(A)$ of an event A. In frequentist statistics, one defines the probability $\mathcal{P}(A)$ as the relative frequency with which the event A occurs in repeated experiments, i.e. the number of times the event occurs over the total number of trials, in the limit of a infinite series of equiprobable repetitions. As forcefully argued for example by Trotta (2008), this definition of probability has several shortcomings. Besides being useless in real life (as it assumes an infinite repetition of experiments with nominally identical test conditions, requirement that is never met in most practical cases), it cannot handle unrepeatable situations, which have a particular importance in cosmology, as we have exactly one sample of the Universe. More importantly, this definition is surprisingly circular, in the sense that it assumes that repeated trials are equiprobable, whereas it is the very notion of probability that is being defined in the first place.

On the other hand, in Bayesian statistics, the probability $\mathcal{P}(A)$ represents the degree of belief that any reasonable person (or machine) shall attribute to the occurrence of event A under consideration of all available information. This definition implies that in Bayesian theory, probabilities are used to quantify uncertainties independently of their origin, and therefore applies to any event. In other words, probabilities represent a state of knowledge in presence of partial information. This is the intuitive concept of probability as introduced by Laplace, Bayes, Bernoulli, Gauß, Metropolis, Jeffreys, etc. (see Jaynes, 2003).

3.1.2 On parameter determination

Translated to the measurement of a parameter in an experiment, the definitions of probabilities given in the previous section yield differences in the questions addressed by frequentist and Bayesian statistical analyses.

In the frequentist point of view, statements are of the form: "the measured value x occurs with probability $\mathcal{P}(x)$ if the measurand X has the true value $\mathcal{X}$". This means that the only questions that can be answered are of the form: "given the true value $\mathcal{X}$ of the measurand X, what is the probability distribution of the measured values x?". It also implies that statistical analyses are about building *estimators*, $\hat{X}$, of the truth, $\mathcal{X}$.

In contrast, Bayesian statistics allows statements of the form: "given the measured value x, the measurand X has the true value $\mathcal{X}$ with probability Q". Therefore, one can also answer the question: "given the observed measured value x, what is the probability that the true value of X is $\mathcal{X}$?", which arguably is the only natural thing to demand from data analysis. For this reason, Bayesian statistics offers a principled approach to the question underlying every measurement problem, of how to *infer* the true value of the measurand given all available information, including observations.

In summary, in the context of parameter determination, the fundamental difference between the two approaches is that frequentist statistics assumes the measurement to be uncertain and the measurand known, while Bayesian statistics assumes the observation to be known and the measurand uncertain. Similar considerations can be formulated regarding the problems of hypothesis testing and model comparison.

3.1.3 Probability theory as extended logic

As outlined in the seminal work of Cox (1946), popularized and generalized by Jaynes (in particular in his inspirational posthumous book, Jaynes, 2003),[1] neither the Bayesian nor the frequentist approach is universally applicable. It is possible to adopt a more general viewpoint that can simply be referred to as "probability theory", which encompasses both approaches. This framework automatically includes all Bayesian and frequentist calculations, but also contains concepts that do not fit into either category (for example, the principle of maximum entropy, which can be applied in the absence of a particular model, when very little is known beyond the raw data).

[1] At this point, the influence of Shannon (1948) and Pólya (1954a,b) should also be emphasized.

In the author's view, this approach is a breakthrough that remains shockingly unknown in astrophysics. As we believe that a conceptual understanding of these concepts are of interest for the purpose of this thesis, we now qualitatively describe the salient features of this way of thinking.

The Cox-Jaynes theorem (1946) states that there is only a single set of rules for doing plausible reasoning which is consistent with a set of axioms that is in qualitative correspondence with common sense. These axioms, or desiderata, are (Jaynes, 2003, section 1.7):

1. *Degrees of plausibility are represented by real numbers.* We denote by $w(A|B)$ the real number assigned to the plausibility of some proposition A, given some other proposition B.

2. *Plausible reasoning qualitatively agrees with human common sense with respect to the "direction" in which reasoning is to go.* Formally, we introduce a continuity assumption: $w(A)$ changes only infinitesimally if A changes infinitesimally. In addition, if some old information C gets updated to C' in such a way that the plausibility of A is increased, but the plausibility of A given B is unchanged, i.e. $w(A|C') > w(A|C)$ and $w(B|AC') = w(B|AC)$, we demand that the plausibility that A is false decrease, i.e. $w(\bar{A}|C') < w(\bar{A}|C)$, and that the plausibility of A and B can only increase, i.e. $w(AB|C') \geq w(AB|C)$.

3. *Plausible reasoning is performed consistently.* This is requiring the three common colloquial meanings of the word "consistent":

   (a) *If a conclusion can be reached in more than one way, then every possible way must lead to the same result.*

   (b) *Consistent plausible reasoning always takes into account all of the evidence it has relevant to a question. It does not arbitrarily ignore some of the available information, basing its conclusion on what remains. In other words, it is completely non-ideological.*

   (c) *Equivalent states of knowledge (up to the labeling of propositions) are represented by equal plausibility assignments.*

The Cox-Jaynes theorem demonstrates that the only consistent system to manipulate numerical "plausibilities" that respect these rules is isomorphic to probability theory,[2] and shows that this system consistently extends the two-valued Boolean algebra $\{0, 1\}$ to the continuum $[0, 1]$. This paradigm therefore introduces a "logical" interpretation of probabilities that can be deduced without any reference to frequencies.

In this perspective, statistical techniques that use Bayes' theorem or the maximum-entropy inference rule are fully as valid as any based on the frequentist interpretation of probability. In fact, they are the *unique* consistent generalization of logical deduction in the presence of uncertainty. As demonstrated by Jaynes, their introduction enables to broaden the scope of probability theory so that it includes various seemingly unrelated fields, such as communication theory of the maximum-entropy interpretation of thermodynamics. They also provides a rational basis to the mechanism of logical induction and therefore to machine learning.

3.2 Inverse problems and the mechanism of experimental learning

This section draws from section 3 in Leclercq, Pisani & Wandelt (2014).

The "plausible reasoning" framework described in section 3.1 can be formulated mathematically by introducing the concept of conditional probability $\mathcal{P}(A|B)$, which describes the probability that event A will occur given whatever information B is given on the right side of the vertical conditioning bar. To conditional probabilities applies the following famous identity, which allows to go from forward modeling to the inverse problem, by noting that if one knows how x arises from y, then one can use x to constrain y:

$$\mathcal{P}(y|x)\mathcal{P}(x) = \mathcal{P}(x|y)\mathcal{P}(y) = \mathcal{P}(x, y). \tag{3.1}$$

This observation forms the basis of Bayesian statistics.

[2] Formally, the theorem states that there exists an isomorphism f such that for any two propositions A, B, we have $f \circ w(A|B) = \mathcal{P}(A|B)$.

3.2.1 What is Bayesian analysis?

Bayesian analysis is a general method for updating the probability estimate for a theory in light of new data. It is based on Bayes' theorem,

$$\mathcal{P}(\theta|d) = \frac{\mathcal{P}(d|\theta)\mathcal{P}(\theta)}{\mathcal{P}(d)}. \tag{3.2}$$

In the previous formula, θ represents the set of model parameters for a particular theory and d the data (before it is known), written as a vector. Therefore,

- $\mathcal{P}(d|\theta)$ is the probability of the data *before it is known*, given the theory. It is called the *likelihood*;
- $\mathcal{P}(\theta)$ is the probability of the theory in the absence of data. It is called the prior probability distribution function or simply the *prior*;
- $\mathcal{P}(\theta|d)$ is the probability of the theory after the data is known. It is called the posterior probability distribution function or simply the *posterior*;
- $\mathcal{P}(d)$ is the probability of the data *before it is known*, without any assumption about the theory. It is called the *evidence*.

A simple way to summarize Bayesian analysis can be formulated by the following:

Whatever is uncertain gets a pdf.

This statement can be a little disturbing at first (e.g. the value of $\Omega_{\rm m}$ is a constant of nature, certainly not a random number of an experiment). What it means is that in Bayesian statistics, pdfs are used to quantify uncertainty of all kinds, not just what is usually referred to as "randomness" in the outcome of an experiment. In other words, the pdf for an uncertain parameter can be thought as a "belief distribution function", quantifying the degree of truth that one attributes to the possible values for some parameter (see the discussion in section 3.1.1). Certainty can be represented by a Dirac distribution, e.g. if the data determine the parameters completely.

The inputs of a Bayesian analysis are of two sorts:

- the *prior*: it includes modeling assumptions, both theoretical and experimental. Specifying a prior is a systematic way of quantifying what one assumes true about a theory before looking at the data.
- the *data*: in cosmology, these can include the temperature in pixels of a CMB map, galaxy redshifts, photometric redshifts pdfs, etc. Details of the survey specifications have also to be accounted for at this point: noise, mask, survey geometry, selection effects, biases, etc.

A key point is that the output of a Bayesian analysis is a pdf, the *posterior density*. Therefore, contrary to frequentist statistics, the output of the analysis is not an estimator for the parameters. The word "estimator" has a precise meaning in frequentist statistics: it is a function of the data which returns a number that is meant to be close to the parameter it is designed to estimate; or the left and right ends of a confidence interval, etc. The outcome of a Bayesian analysis is the posterior pdf, a pdf whose values give a quantitative measure of the relative degree of rational belief in different parameter values given the combination of prior information and the data.

3.2.2 Prior choice

The prior choice is a key ingredient of Bayesian statistics. It is sometimes considered problematic, since there is no unique prescription for selecting the prior. Here we argue that prior specification is not a limitation of Bayesian statistics and does not undermine objectivity as sometimes misstated.

The guiding principle is that there can be no inference without assumptions, that there does not exist an "external truth", but that science is building predictive models in certain axiomatic frameworks. In this regard, stating a prior in Bayesian probability theory becomes a systematic way to quantify one's assumptions and state of knowledge about the problem in question before the data is examined. While it is true that such probability assignment does not describe something that could be measured in a physical experiment, it is completely objective in the sense that it is independent of the "personal feelings" of the user. Anyone who has

the same information, but comes to a different conclusion, is necessarily violating one of Cox's desiderata (see the discussion in section 3.1.3).

Bayes' theorem gives an unequivocal procedure to update even different degrees of beliefs. As long as the prior has a support that is non-zero in regions where the likelihood is large (Cromwell's rule), the repeated application of the theorem will converge to a unique posterior distribution (Bernstein-von Mises theorem). Generally, objectivity is assured in Bayesian statistics by the fact that, if the likelihood is more informative than the prior, the posterior converges to a common function.

Specifying priors exposes assumptions to falsification and scientific criticism. This is a positive feature of Bayesian probability theory, because frequentists also have to make assumptions that may be more difficult to find within the analysis. An important theorem (Wolpert & Macready, 1997) states that there is "no-free lunch" for optimization problems: when searching for the local extremum of a target function (the likelihood in our case) in a finite space, the average performance of algorithms (that do not resample points) across all possible problems is identical. An important implication is that no universally good algorithm exists (Ho & Pepyne, 2002); prior information should always be used to match procedures to problems.

In many situations, domain knowledge is highly relevant and should be included in the analysis. For example, when trying to estimate a mass m from some data, one should certainly enforce it to be a positive quantity by setting a prior such that $\mathcal{P}(m) = 0$ for $m < 0$. Frequentist techniques based on the likelihood can give estimates and confidence intervals that include negative values. Taken at face value, this result is meaningless, unless special care is taken (e.g. the so-called "constrained likelihood" methods). The use of Bayes' theorem ensures that meaningless results are excluded from the beginning and that one knows how to place bets on values of the parameter given the actual data set at hand.

As discussed in the introduction, in cosmology, the current state-of-the-art is that previous data (COBE, WMAP, Planck, SDSS etc.) allowed to establish an extremely solid theoretical footing: the so-called ΛCDM model. Even when trying to detect deviations from this model in the most recent data, it is absolutely well-founded to use it as prior knowledge about the physical behaviour of the Universe. Therefore, using less informative priors would be refusing to "climb on the shoulder of giants".

It can happen that the data are not informative enough to override the prior (e.g. for sparsely sampled data or very high-dimensional parameter space), in which case care must be given in assessing how much of the final (first level, see section 3.3.1) inference depends on the prior choice. A good way to perform such a check is to simulate data using the posterior and see if it agrees with the observed data. This can be thought of as "calculating doubt" (Starkman, Trotta & Vaudrevange, 2008; March *et al.*, 2011) to quantify the degree of belief in a model given observational data in the absence of explicit alternative models. Note that even in the case where the inference strongly depends on prior knowledge, information has been gained on the constraining power (or lack thereof) of the data.

For model selection questions (second level analysis, see section 3.3.3), the impact of the prior choice is much stronger, since it is precisely the available prior volume that matters in determining the penalty that more complex models should incur. Hence, care should be taken in assessing how much the outcome would change for physically reasonable changes in the prior.

There exists a vast literature about quantitative prescriptions for prior choice that we cannot summarize here. An important topic concerns the determination of "ignorance priors" or "Jeffreys' priors": a systematic way to quantify a maximum level of uncertainty and to reflect a state of indifference with respect to symmetries of the problem considered. While the ignorance prior is unphysical (nothing is ever completely uncertain) it can be viewed as a convenient approximation to the problem of carefully constructing an accurate representation of weak prior information, which can be very challenging – especially in high dimensional parameter spaces.

For example, it can be shown that, if one is wholly uncertain about the position of the pdf, a "flat prior" should be chosen. In this case, the prior is taken to be constant (within some minimum and maximum value of the parameters so as to be proper, i.e. normalizable to unity). In this fashion, equal probability is assigned to equal states of knowledge. However, note that a flat prior on a parameter θ does not necessarily correspond to a flat prior on a non-linear function of that parameter, $\varphi(\theta)$. Since $\mathcal{P}(\varphi) = \mathcal{P}(\theta) \times |\mathrm{d}\theta/\mathrm{d}\varphi|$, a non-informative (flat) prior on θ can be strongly informative about φ. Analogously, if one is entirely uncertain about the width of the pdf, i.e. about the scale of the inferred quantity θ, it can be shown that the appropriate prior is $\mathcal{P}(\theta) \propto 1/\theta$, which gives the same probability in logarithmic bins, i.e. the same weight to all orders of magnitude.

3.3 Bayesian data analysis problems

This section draws from section 3 in Leclercq, Pisani & Wandelt (2014).

Bayesian data analysis problems can be typically classified as: parameter inference, model comparison, hypothesis testing. For example, cosmological questions of these three types, related to the large-scale structure, would be respectively

- What is the value of w, the equation of state of dark energy?
- Is structure formation driven by general relativity or by massive gravity?
- Are large-scale structure observations consistent with the hypothesis of a spatially flat universe?

In this section; we describe the methodology for questions of the first two types. Hypothesis testing, i.e. inference within an uncertain model, in the absence of an explicit alternative, can be treated in a similar manner.

3.3.1 First level analysis: Bayesian parameter inference

The general problem of Bayesian parameter inference can be stated as follows. Given a physical model $\mathcal{M}$,[3] a set of hypotheses is specified in the form of a vector of parameters, θ. Together with the model, priors for each parameter must be specified: $\mathcal{P}(\theta|\mathcal{M})$. The next step is to construct the likelihood function for the measurement, with a probabilistic, generative model of the data: $\mathcal{P}(d|\theta,\mathcal{M})$. The likelihood reflects how the data are obtained: for example, a measurement with Gaussian noise will be represented by a normal distribution.

Once the prior is specified and the data is incorporated in the likelihood function, one immediately obtains the posterior distribution for the model parameters, integrating all the information known to date, by using Bayes' theorem (equation (3.2)):

$$\mathcal{P}(\theta|d,\mathcal{M}) \propto \mathcal{P}(d|\theta,\mathcal{M})\mathcal{P}(\theta|\mathcal{M}). \tag{3.3}$$

Note that the normalizing constant $\mathcal{P}(d|\mathcal{M})$ (the Bayesian evidence) is irrelevant for parameter inference (but fundamental for model comparison, see section 3.3.3).

Usually, the set of parameters θ can be divided in some physically interesting quantities φ and a set of nuisance parameters ψ. The posterior obtained by equation (3.3) is the joint posterior for $\theta = (\varphi,\psi)$. The marginal posterior for the parameters of interest is written as (marginalizing over the nuisance parameters)

$$\mathcal{P}(\varphi|d,\mathcal{M}) \propto \int \mathcal{P}(d|\varphi,\psi,\mathcal{M})\mathcal{P}(\varphi,\psi|\mathcal{M})\,\mathrm{d}\psi. \tag{3.4}$$

This pdf is the final inference on φ from the joint posterior. The following step, to apprehend and exploit this information, is to explore the posterior. It is the subject of the next section.

3.3.2 Exploration of the posterior

The result of parameter inference is contained in the posterior pdf, which is the actual output of the statistical analysis. Since this pdf cannot always be easily represented, convenient communication of the posterior information can take different forms:

- a direct visualization, which is only possible if the parameter space has sufficiently small dimension (see figure 3.1).
- the computation of statistical summaries of the posterior, e.g. the mean, the median, or the mode of the distribution of each parameter, marginalizing over all others, its standard deviation; the means and covariance matrices of some groups of parameters, etc. It is also common to present the inference by plotting two-dimensional subsets of parameters, with the other components marginalized over (this is especially useful when the posterior is multi-modal or with heavy tails).

[3] In this section, we make explicit the choice of a model $\mathcal{M}$ by writing it on the right-hand side of the conditioning symbol of all pdfs.

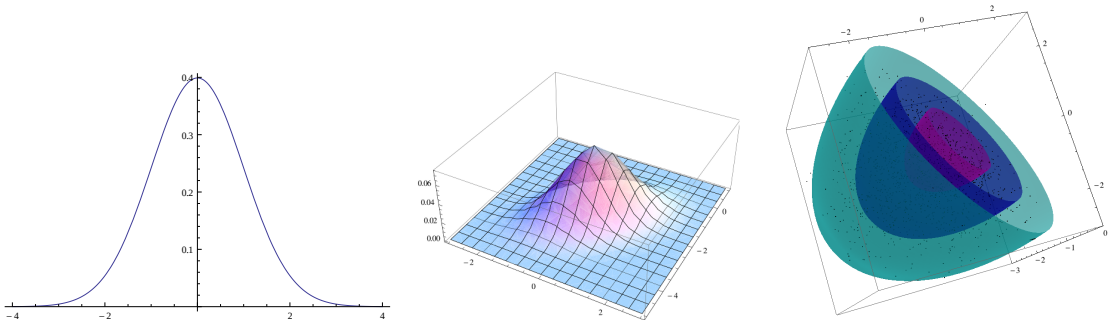

Figure 3.1: Example visualizations of posterior densities in low-dimensional parameter spaces (from left to right: one, two and three).

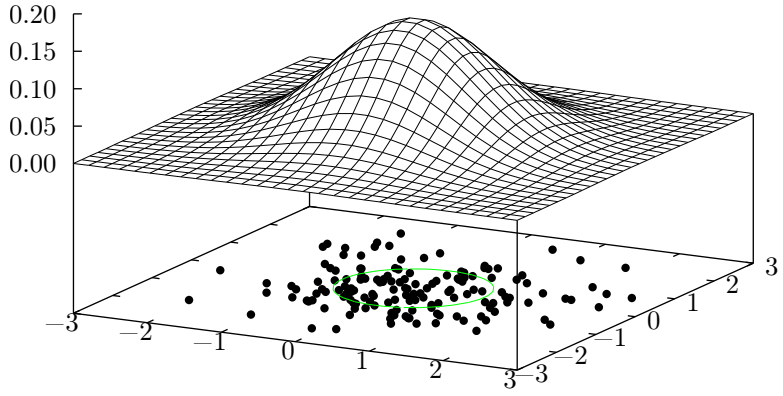

Figure 3.2: Example of a sampled representation of a posterior distribution in two dimensions. A set of samples is constructed in such a way that at any point, the posterior probability is proportional to the local density of samples in parameter space.

For typical problems in cosmology, the exploration of a posterior density meets practical challenges, depending on the dimension D of the parameter space. Due to the computational time requirements, direct integration and mapping of the posterior density is almost never a smart idea, except for $D < 4$. Besides, computing statistical summaries by marginalization means integrating out the other parameters. This is rarely possible analytically (Gaussian random fields being one notable exception), and even numerical direct integration is basically hopeless for $D > 5$.

In this thesis, we will be looking at cases where D is of the order of 10^7: the density in each voxel of the map to infer is a parameter of the analysis. This means that direct evaluation of the posterior is impossible and one has to rely on a numerical approximation: sampling the posterior distribution.

The idea is to approximate the posterior by a set of samples drawn from the real posterior distribution. In this fashion, one replaces the real posterior distribution, $\mathcal{P}(\theta|d)$, by the sum of N Dirac delta distributions, $\mathcal{P}_N(\theta|d)$:

$$\mathcal{P}(\theta|d) \approx \mathcal{P}_N(\theta|d) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathrm{D}}(\theta - \theta_i). \tag{3.5}$$

A sampled representation of the posterior is constructed in such a way that at any point, the posterior probability is proportional to the local density of samples in parameter space (see figure 3.2).

An intuitive way to think about these samples is to consider each of them as a "possible version of the truth". The variation between different samples quantifies the uncertainty. At this point, it is worth stressing again that an advantage of Bayesian approach is that it deals with uncertainty independently of its origin, i.e. there is no fundamental distinction between "statistical uncertainty" coming from the stochastic nature of the experiment and "systematic uncertainty", deriving from deterministic effects that are only partially known.

The advantage of a sampling approach is that marginalization over some parameters becomes trivial: one

just has to histogram. Specifically, it is sufficient to count the number of samples falling within different bins of some subset of parameters, simply ignoring the values of the others parameters. Integration to get means and variances is also much simpler, since the problem is limited to the computation of discrete sums. More generally, the expectation value of any function of the parameters, $f(\theta)$ is

$$\langle f(\theta)\rangle = \int f(\theta)\mathcal{P}(\theta)\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i). \tag{3.6}$$

We make an extensive use of this last property in part IV of this thesis, when exploiting the BORG SDSS analysis for cosmic web classification.

How can one get a sampled representation of the posterior? The ideal case would be to have an infinitely powerful computer. Then, a naïve but straightforward sampling algorithm would be the following: simulate data from the generative model (draw θ from the prior, then data from the likelihood knowing θ) and check that the real data agree with the simulated data. If it is the case, keep θ as one sample, otherwise try again. This is correct in principle, but hugely inefficient, particularly in high dimensions where it can become prohibitively expensive to evaluate the posterior pdf. Fortunately, a battery of powerful methods exists for approximating and sampling from probability distributions. Interestingly, sampling algorithms exist that do not evaluate the posterior pdf (except perhaps occasionally, to maintain high numerical precision).

One class of approaches is Approximate Bayesian Computation (ABC) sometimes also known as "likelihood-free" methods (see Marin *et al.*, 2012 for an overview, or Cameron & Pettitt, 2012; Weyant, Schafer & Wood-Vasey, 2013; Lin & Kilbinger, 2015 for applications to astrophysics). The general principle is similar to the naïve approach described above, but ABC makes it practical by using an approximate forward model, the outcomes $\tilde{d}$ of which are compared with the observed data d. The candidate sample $\tilde{d}$ is accepted with tolerance $\varepsilon > 0$ if $\rho(\tilde{d}, d) \leq \varepsilon$, where the distance measure ρ determines the allowed level of discrepancy between $\tilde{d}$ and d based on a given metric.

Another important class of standard techniques to sample the posterior is to use Markov Chain Monte Carlo, which is the subject of section 3.4.

3.3.3 Second level analysis: Bayesian model comparison

In the case where there are several competing theoretical models, second level inference (or Bayesian model comparison) provides a systematic way of evaluating their relative probability in light of the data and any prior information available. It does not replace parameter inference, but rather extends the assessment of hypotheses to the space of theoretical models.

This allows quantitatively to address everyday questions in cosmology – Is the Universe flat or should one allow a non-zero curvature parameter? Are the primordial perturbations Gaussian or non-Gaussian? Are there isocurvature modes? Are the perturbations strictly scale-invariant ($n_\mathrm{s} = 1$) or should the spectrum be allowed to deviate from scale-invariance? Is there evidence for a deviation from general relativity? Is the equation of state of dark energy equal to -1?

In many of the situations above, Bayesian model comparison offers a way of balancing complexity and goodness of fit: it is obvious that a model with more free parameters will always fit the data better, but it should also be "penalized" for being more complex and hence, less predictive. The notion of predictiveness really is central to Bayesian model comparison in a very specific way: the evidence is actually the prior predictive pdf, the pdf over all data sets predicted for the experiment before data are taken. Since predictiveness is a criterion for good science everyone can agree on, it is only natural to compare models based on how well they predicted the data set before it was obtained. This criterion arises automatically in the Bayesian framework.

The guiding scientific principle is known as Occam's razor: the simplest model compatible with the available information ought to be preferred. We now understand this principle as a consequence of using predictiveness as the criterion. A model that is so vague (e.g. has so many parameters) that it can predict a large range of possible outcomes will predict any data set with smaller probability than a model that is highly specific and therefore has to commit to predicting only a small range of possible data sets. It is clear that the specific model should be preferred if the data falls within the narrow range of its prediction. Conversely we default to the broader more general model only if the data are incompatible with the specific model. Therefore, Bayesian model comparison offers formal statistical grounds for selecting models based on an evaluation whether the data truly favor the extra complexity of one model compared to another.

Contrary to frequentists goodness-of-fit tests, second level inference always requires an alternative explanation for comparison (finding that the data are unlikely within a theory does not mean that the theory itself is improbable, unless compared with an alternative). The prior specification is crucial for model selection issues: since it is the range of values that parameters can take that controls the sharpness of Occam's razor, the prior should exactly reflect the available parameter space under the model before obtaining the data.

The evaluation of model $\mathcal{M}$'s performance given the data is quantified by $\mathcal{P}(\mathcal{M}|d)$. Using Bayes' theorem to invert the order of conditioning, we see that it is proportional to the product of the prior probability for the model itself, $\mathcal{P}(\mathcal{M})$, and of the Bayesian evidence already encountered in first level inference, $\mathcal{P}(d|\mathcal{M})$:

$$\mathcal{P}(\mathcal{M}|d) \propto \mathcal{P}(\mathcal{M})\,\mathcal{P}(d|\mathcal{M}). \tag{3.7}$$

Usually, prior probabilities for the models are taken as all equal to $1/N_\mathrm{m}$ if one considers N_m different models (this choice is said to be *non-committal*). When comparing two competing models denoted by $\mathcal{M}_1$ and $\mathcal{M}_2$, one is interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\mathcal{P}_{12} \equiv \frac{\mathcal{P}(\mathcal{M}_1|d)}{\mathcal{P}(\mathcal{M}_2|d)} = \frac{\mathcal{P}(\mathcal{M}_1)}{\mathcal{P}(\mathcal{M}_2)}\frac{\mathcal{P}(d|\mathcal{M}_1)}{\mathcal{P}(d|\mathcal{M}_2)}. \tag{3.8}$$

With non-committal priors on the models, $\mathcal{P}(\mathcal{M}_1) = \mathcal{P}(\mathcal{M}_2)$, the ratio simplifies to the ratio of evidences, called the *Bayes factor*,

$$\mathcal{B}_{12} \equiv \frac{\mathcal{P}(d|\mathcal{M}_1)}{\mathcal{P}(d|\mathcal{M}_2)}. \tag{3.9}$$

The Bayes factor is the relevant quantity to update our state of belief in two competing models in light of the data, regardless of the relative prior probabilities we assign to them: a value of $\mathcal{B}_{12}$ greater than one means that the data support model $\mathcal{M}_1$ over model $\mathcal{M}_2$. Note that, generally, the Bayes factor is very different from the ratio of likelihoods: a more complicated model will always yield higher likelihood values, whereas the evidence will favor a simpler model if the fit is nearly as good, through the smaller prior volume.

Posterior odds (or directly the Bayes factor in case of non-committal priors) are usually interpreted against the Jeffreys' scale for the strength of evidence. For two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ with non-committal priors ($\mathcal{P}(\mathcal{M}_1) = \mathcal{P}(\mathcal{M}_2) = 1/2$) and exhausting the model space ($\mathcal{P}(\mathcal{M}_1|d) + \mathcal{P}(\mathcal{M}_2|d) = 1$), the relevant quantity is the logarithm or the Bayes factor, $\ln \mathcal{B}_{12}$ for which thresholds at values of 1.0, 2.5 and 5.0 are set (corresponding to odds of about 3:1, 12:1 and 150:1, representing weak, moderate and strong evidence, respectively). The use of a logarithm in this empirical scale quantifies the principle that the evidence for a model only accumulates slowly with new informative data: rising up one level in the evidence strength requires about one order of magnitude more support.

The computation of the Bayesian evidence is generally technically challenging. For this reason, simplifying assumptions often have to be introduced (see Heavens, Kitching & Verde, 2007, for the Gaussian likelihood approximation within a model selection context). Another important particular situation is when $\mathcal{M}_2$ is a simpler model, described by fewer ($n' < n$) parameters than $\mathcal{M}_1$. $\mathcal{M}_2$ is said to be *nested* in model $\mathcal{M}_1$ if the n' parameters of $\mathcal{M}_2$ are also parameters of $\mathcal{M}_1$. $\mathcal{M}_1$ has $p \equiv n - n'$ extra parameters that are fixed to fiducial values in $\mathcal{M}_2$. For simplicity, let us assume that there is only one extra parameter ζ in model $\mathcal{M}_1$, fixed to 0 in $\mathcal{M}_2$ (ζ describes the continuous deformation from one model to the other). Let us denote the set of other parameters by θ. Under these hypotheses, the evidence for $\mathcal{M}_1$ is $\mathcal{P}(d|\mathcal{M}_1) \equiv \mathcal{P}(d|\mathcal{M}_{\theta,\zeta})$ and the evidence for $\mathcal{M}_2$ is $\mathcal{P}(d|\mathcal{M}_2) \equiv \mathcal{P}(d|\mathcal{M}_{\theta,\zeta=0}) = \mathcal{P}(d|\zeta = 0, \mathcal{M}_{\theta,\zeta})$. We also assume non-committal priors for $\mathcal{M}_1$ and $\mathcal{M}_2$.

If the prior for the additional parameter ζ is independent of the other parameters (which makes the joint prior separable: $\mathcal{P}(\zeta, \theta|\mathcal{M}_{\theta,\zeta}) = \mathcal{P}(\zeta|\mathcal{M}_{\theta,\zeta})\mathcal{P}(\theta|\mathcal{M}_{\theta,\zeta=0})$), it can be shown that the Bayes factor takes a simple form, the Savage-Dickey ratio (Dickey, 1971; Verdinelli & Wasserman, 1995)

$$\mathcal{B}_{12} = \frac{\mathcal{P}(d|\mathcal{M}_{\theta,\zeta})}{\mathcal{P}(d|\mathcal{M}_{\theta,\zeta=0})} = \frac{\mathcal{P}(\zeta = 0|\mathcal{M}_{\theta,\zeta})}{\mathcal{P}(\zeta = 0|d, \mathcal{M}_{\theta,\zeta})}, \tag{3.10}$$

that is, the ratio of the marginal prior and the marginal posterior of the larger model $\mathcal{M}_1$, where the additional parameter ζ is held at its fiducial value. The Bayes factor favors the "larger" model only if the data decreases the posterior pdf at the fiducial value compared to the prior. Operationally, if $n - n'$ is small, one can easily compute the Savage-Dickey ratio given samples from the posterior and prior of $\mathcal{M}_1$ by simply estimating the marginal densities at the fiducial value.

3.4 Markov Chain Monte Carlo techniques for parameter inference

This section draws from section 3 in Leclercq, Pisani & Wandelt (2014).

3.4.1 Markov Chains

The purpose of Markov Chain Monte Carlo (MCMC) sampling is to construct a sequence of points in parameter space (a so-called "chain"), whose density is proportional to the pdf that we want to sample.

A sequence $\{\theta_0, \theta_1, \theta_2, ..., \theta_n, ...\}$ of random elements of some set (the "state space") is called a *Markov Chain* if the conditional distribution of θ_{n+1} given all the previous elements θ_1, ... θ_n depends only on θ_n (the *Markov property*). It is said to have *stationary transition probability* if, additionally, this distribution does not depend on n. This is the main kind of Markov chains of interest for MCMC.

Such stationary chains are completely characterized by the marginal distribution for the first element θ_0 (the *initial distribution*) and the conditional distribution of θ_{n+1} given θ_n, called the *transition probability distribution*.

Let us denote by $\mathcal{P}(\theta)$ the target pdf and by $\mathcal{T}(\theta'|\theta)$ the transition pdf. When designing a MCMC method, we want to construct a chain with the following properties.

1. The desired distribution $\mathcal{P}(\theta)$ should be an *invariant distribution* of the chain, namely the probability of the next state being θ must satisfy the general balance property,

$$\mathcal{P}(\theta) = \int \mathcal{T}(\theta|\theta')\,\mathcal{P}(\theta')\,\mathrm{d}\theta'. \tag{3.11}$$

   Formally, an invariant distribution is a fixed point of the transition probability operator, i.e. an eigenvector with eigenvalue 1.

2. The chain should be *ergodic* (or *irreducible*) which means that it is possible to go from every state to every state (not necessarily in one move).

Property 1 ensures the existence of an invariant distribution, and property 2 its uniqueness: it is the target pdf $\mathcal{P}(\theta)$. Therefore, the crucial property of such Markov chains is that, after some steps depending on the initial position (the so-called "burn-in" phase), they reach a state where successive elements of the chain are drawn from the high-density regions of the target distribution, in our case the posterior of a Bayesian parameter inference: the probability to draw θ as the n-th element of the chain, $\mathcal{P}^{(n)}(\theta)$, satisfies

$$\mathcal{P}^{(n)}(\theta) \to \mathcal{P}(\theta) \text{ as } n \to \infty, \text{ for any } \theta_0. \tag{3.12}$$

Exploiting this property, MCMC algorithms use Markovian processes to move from one state to another in parameter space; then, given a set of random samples, they reconstruct the probability heuristically. Several MCMC algorithms exist and the relevant choice is highly dependent on the problem addressed and on the posterior distribution to be explored (see the discussion of the "no-free lunch" theorem in section 3.2.2), but the basic principle is always similar to that of the popular CosmoMC code (Lewis & Bridle, 2002): perform a random walk in parameter space, constrained by the posterior probability distribution.

Many useful transition probabilities satisfy the detailed balance property,

$$\mathcal{T}(\theta|\theta')\,\mathcal{P}(\theta') = \mathcal{T}(\theta'|\theta)\,\mathcal{P}(\theta). \tag{3.13}$$

While general balance expresses the "balance of flow" into and out of any state θ, detailed balance expresses the "balance of flow" between every pair of states: the flow from θ to θ' is the flow from θ' to θ. Markov chains that satisfy detailed balance are also called *reversible Markov chains*. The reason why the detailed balance property is of interest is that it is a sufficient (but not necessary) condition for the invariance of the distribution $\mathcal{P}$ under the transition pdf $\mathcal{T}$ (equation (3.11)), which can be easily checked:

$$\int \mathcal{T}(\theta|\theta')\,\mathcal{P}(\theta')\,\mathrm{d}\theta' = \int \mathcal{T}(\theta'|\theta)\,\mathcal{P}(\theta)\,\mathrm{d}\theta' = \mathcal{P}(\theta)\int \mathcal{T}(\theta'|\theta)\,\mathrm{d}\theta' = \mathcal{P}(\theta). \tag{3.14}$$

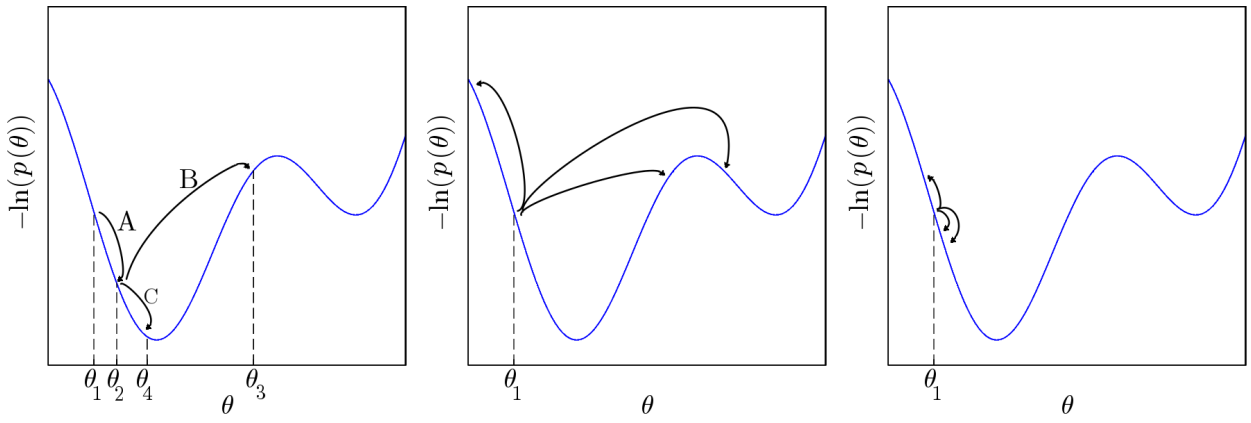

Figure 3.3: *Left panel.* An example of Markov chain constructed by the Metropolis-Hastings algorithm: starting at θ_1, θ_2 is proposed and accepted (step A), θ_3 is proposed and refused (step B), θ_4 is proposed and accepted (step C). The resulting chain is $\{\theta_1, \theta_2, \theta_2, \theta_4, ...\}$. *Central panel.* An example of what happens with too broad a jump size: the chain lacks mobility because all the proposals are unlikely. *Right panel.* An example of what happens with too narrow a jump size: the chain samples the parameter space very slowly.

3.4.2 The Metropolis-Hastings algorithm

A popular version of MCMC is called the Metropolis-Hastings (MH) algorithm, which works as follows. Initially, one chooses an arbitrary point θ_0 to be the first sample, and specifies a distribution $\mathcal{Q}(\theta'|\theta)$ which proposes a candidate θ' for the next sample value, given the previous sample value θ ($\mathcal{Q}$ is called the proposal density or jumping distribution). At each step, one draws a realization θ' from $\mathcal{Q}(\theta'|\theta)$ and calculates the *Hastings ratio*:

$$r(\theta, \theta') \equiv \frac{\mathcal{P}(\theta')}{\mathcal{P}(\theta)} \frac{\mathcal{Q}(\theta|\theta')}{\mathcal{Q}(\theta'|\theta)}. \tag{3.15}$$

The proposed move to θ' is accepted with probability $a(\theta, \theta') \equiv \min[1; r(\theta, \theta')] = \mathcal{T}(\theta'|\theta)$. In case it is accepted, θ' becomes the new state of the chain, otherwise the chain stays at θ. A graphical illustration of the MH algorithm is shown in figure 3.3. Note that each step only depends on the previous one and is also independent of the number of previous steps, therefore the ensemble of samples of the target distribution, constructed by the algorithm, is a stationary Markov chain.

The probability that the next state is θ' is the sum of the probability that the current state is θ' and the update leads to rejection – which happens that a probability that we note $\mathcal{R}(\theta')$ – and of the probability that the current state is some θ and a move from θ to θ' is proposed and accepted. This is formally written

$$\mathcal{P}(\theta') = \int \mathcal{P}(\theta)\mathcal{T}(\theta'|\theta)\,\mathrm{d}\theta = \mathcal{P}(\theta')\,\mathcal{R}(\theta') + \int \mathcal{P}(\theta)\,\mathcal{Q}(\theta'|\theta)\,\mathrm{d}\theta. \tag{3.16}$$

The probability to depart from θ' to any θ is $\int \mathcal{Q}(\theta|\theta')\,\mathrm{d}\theta = 1 - \mathcal{R}(\theta')$.

The special case of a symmetric proposal distribution, i.e. $\mathcal{Q}(\theta|\theta') = \mathcal{Q}(\theta'|\theta)$ for all θ and θ' is called the *Metropolis update*. Then the Hastings ratio simplifies to

$$r(\theta, \theta') = \frac{\mathcal{P}(\theta')}{\mathcal{P}(\theta)} \tag{3.17}$$

and is called the *Metropolis ratio*. Given this result, the detailed balance condition, equation (3.13) reads

$$\mathcal{P}(\theta') \min\left[1\,; \frac{\mathcal{P}(\theta)}{\mathcal{P}(\theta')}\right] = \mathcal{P}(\theta) \min\left[1\,; \frac{\mathcal{P}(\theta')}{\mathcal{P}(\theta)}\right], \tag{3.18}$$

which is easily seen to be true.

In many cases, the MH algorithm will be inefficient if the proposal distribution is sub-optimal. It is often hard to find good proposal distributions if the parameter space has high dimension (e.g. larger than 10). Typically, the chain moves very slowly, either due to a tiny step size, either because only a tiny fraction of proposals are

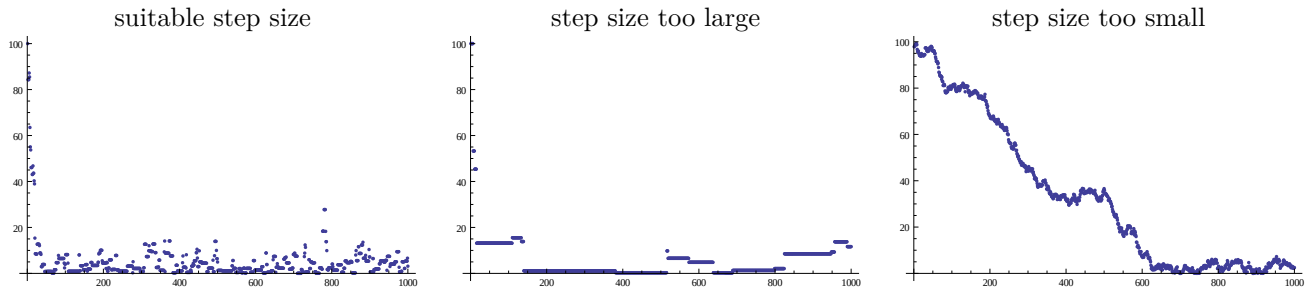

Figure 3.4: Example of Markov chains constructed by the Metropolis-Hastings algorithm, sampling the same target distribution but with varying proposal distribution (step size). The plots show the value of the sampled parameter as a function of the position in the chain. The ideal behavior with a suitable step size is shown in the left panel. On the central panel, the step size is too large: the maximum likelihood region is not well sampled. On the right panel, the step size is too small: the burn-in phase is very long and the sampling is slow. Note that this phenomena are easily diagnosed using the auto-correlation function of the chain, equation (3.19).

accepted. The initial burn-in phase can be very long, i.e. the chain takes some time to reach high likelihood regions, where the initial position chosen has no influence on the statistics of the chain. Even in the stationary state, sufficient sampling of the likelihood surface can take a very large number of steps. In the central and left panels of figure 3.3, we illustrate what happens with too broad a jump size (the chain lacks mobility and all proposals are unlikely) or too narrow (the chain moves slowly to sample all the parameter space). Note that the step-size issues can be diagnosed using the lagged auto-correlation function of the chain,

$$\xi(\Delta) = \int \theta(t)\theta(t+\Delta)\,\mathrm{d}t. \tag{3.19}$$

A convergence criterion using different chains or sections of chains is proposed in Gelman & Rubin (1992). Possible solutions to the issues mentioned involve an adaptive step size or refinements of the standard Metropolis-Hastings procedure.

In some particular cases, the proposal density itself satisfies the detailed balance property,

$$\mathcal{Q}(\theta|\theta')\,\mathcal{P}(\theta') = \mathcal{Q}(\theta'|\theta)\,\mathcal{P}(\theta), \tag{3.20}$$

which implies that the Hastings ratio is always unity, i.e. that proposed states are always accepted ($\mathcal{Q}$ is $\mathcal{T}$ and $\mathcal{R}$ is zero). For example, Gibbs sampling is a particular case of a generalized MH algorithm, alternating between different proposals (see e.g. Wandelt, Larson & Lakshminarayanan, 2004 for a cosmological example). It is particularly helpful when the joint probability distribution is difficult to sample directly, but the conditional distribution of some parameters given the others is known. It uses a block scheme of individual *Gibbs updates* to sample an instance from the distribution of each variable in turn, conditional on the current values of the other variables. Formally, the proposal for a single Gibbs update is from a conditional distribution of the target pdf: $\mathcal{Q}(\theta'|\theta) \equiv \mathcal{P}(\theta'|f(\theta))$ where $f(\theta)$ is θ with some components omitted. θ' is an update of these missing components, keeping the others at the values they had in θ. Therefore, $f(\theta') = f(\theta)$, and we have

$$\mathcal{Q}(\theta'|\theta) \equiv \mathcal{P}(\theta'|f(\theta)) = \mathcal{P}(\theta'|f(\theta')) = \mathcal{P}(\theta'), \tag{3.21}$$

which trivially implies the detailed balance property (equation (3.20)) and ensures an acceptance rate of unity.

3.4.3 Hamiltonian Monte Carlo

A very efficient MCMC algorithm for high-dimensional problems such as those encountered in cosmology is Hamiltonian Monte Carlo (HMC, originally introduced under the name of hybrid Monte Carlo, Duane *et al.*, 1987). A detailed overview is provided by Neal (2011).

The general idea of HMC is to use concepts borrowed from classical mechanics to solve statistical problems. As it is a core ingredient in the BORG code, we now discuss the most important features of HMC. We start by reviewing physical properties of Hamiltonian dynamics. The system is described by the Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p})$,

a function of the D-dimensional position vector $\boldsymbol{\theta}$ and of the D-dimensional momentum vector $\mathbf{p}$.[4] Its time evolution is described by Hamilton's equations,

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \frac{\partial H}{\partial \mathbf{p}}, \tag{3.22}$$

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{\theta}}. \tag{3.23}$$

For any time interval of duration s, these equations define a mapping T_s from the state at any time t to the state at time $t+s$. The first important property of Hamiltonian dynamics is time reversibility, which means for any s, that the mapping T_s has an inverse. It is easy to check that this inverse is T_{-s}.

A second property of the dynamics is that it conserves the Hamiltonian during the evolution, which can be checked explicitly:

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{\theta}}\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} + \frac{\partial H}{\partial \mathbf{p}}\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{\theta}}\frac{\partial H}{\partial \mathbf{p}} - \frac{\partial H}{\partial \mathbf{p}}\frac{\partial H}{\partial \boldsymbol{\theta}} = 0. \tag{3.24}$$

In $2D$ dimensions, using $\mathbf{z} = (\boldsymbol{\theta}, \mathbf{p})$ and the matrix

$$\mathbf{J} = \begin{pmatrix} \mathbf{0}_D & \mathbf{I}_D \\ -\mathbf{I}_D & \mathbf{0} \end{pmatrix}, \tag{3.25}$$

one can rewrite Hamilton's equations as

$$\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = \mathbf{J} \cdot \nabla H. \tag{3.26}$$

The third important property is that Hamiltonian dynamics is *symplectic*, which means that the Jacobian matrix $\mathbf{B}_s$ of the mapping T_s satisfies

$$\mathbf{B}_s^\intercal \, \mathbf{J}^{-1} \, \mathbf{B}_s = \mathbf{J}^{-1}. \tag{3.27}$$

This property implies volume conservation in $(\boldsymbol{\theta}, \mathbf{p})$ phase space (a result also known as Liouville's theorem), since $\det(\mathbf{B}_s)^2$ must be one.

Crucially, reversibility and symplecticity are properties that can be maintained exactly, even when Hamiltonian dynamics is approximated by numerical integrators (see section 4.3.4).

The link between probabilities and Hamiltonian dynamics is established via the concept of *canonical distribution* from statistical mechanics. Given the energy distribution $E(\mathbf{x})$ for possibles states $\mathbf{x}$ of the physical system, the canonical distribution over states $\mathbf{x}$ has pdf

$$\mathcal{P}(\mathbf{x}) = \frac{1}{Z} \exp\left(\frac{-E(\mathbf{x})}{k_\mathrm{B} T}\right) \tag{3.28}$$

where k_B is the Boltzmann constant, T the temperature of the system, and the *partition function* Z is the normalization constant needed to ensure $\int \mathcal{P}(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$. In Hamiltonian dynamics, H is an energy function for the joint state of positions $\boldsymbol{\theta}$ and momenta $\mathbf{p}$, and hence defines a joint pdf as

$$\mathcal{P}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{Z} \exp\left(\frac{-H(\boldsymbol{\theta}, \mathbf{p})}{k_\mathrm{B} T}\right) \tag{3.29}$$

Viewing this the opposite way, if we are interested in some joint distribution with probability $\mathcal{P}(\boldsymbol{\theta}, \mathbf{p})$, we can obtain it as a canonical distribution with temperature $k_\mathrm{B}T = 1$, by setting $H(\boldsymbol{\theta}, \mathbf{p}) = -\ln \mathcal{P}(\boldsymbol{\theta}, \mathbf{p}) - \ln Z$, where Z is any convenient positive constant (we choose $Z = 1$ in the following for simplicity).

We are now ready to discuss the Hamiltonian Monte Carlo algorithm. HMC interprets the negative logarithm of the pdf to sample as a physical potential, $\psi(\boldsymbol{\theta}) = -\ln \mathcal{P}(\boldsymbol{\theta})$ and introduces auxiliary variables: "conjugate momenta" p_i for all the different parameters. Using these new variables as nuisance parameters, one can formulate a Hamiltonian describing the dynamics in the multi-dimensional phase space. Such a Hamiltonian is given as:

$$H(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^\intercal \, \mathbf{M}^{-1} \, \mathbf{p} + \psi(\boldsymbol{\theta}) = -\ln \mathcal{P}(\boldsymbol{\theta}, \mathbf{p}), \tag{3.30}$$

where the kinetic term, $K(\mathbf{p}) \equiv \frac{1}{2}\mathbf{p}^\intercal \, \mathbf{M}^{-1} \, \mathbf{p}$ involves $\mathbf{M}$, a symmetric positive definite "mass matrix" whose choice can strongly impact the performance of the sampler. Masses characterize the inertia of parameters when

[4] In this section we use boldface notations for all vectors, to strengthen the link between physics and statistics.

moving through the parameter space. Consequently, too large masses will result in slow exploration efficiency, while too light masses will result in large rejection rates (see also figure 3.4).

Each iteration of the HMC algorithm works as follows. One draws a realization of the momenta from the distribution defined by the kinetic energy term, i.e. a multi-dimensional Gaussian with a covariance matrix $\mathbf{M}$, then moves the positions $\boldsymbol{\theta}$ using a Hamiltonian integrator in parameter space, respecting symplectic symmetry. In other words, we first "kick the system" then follow its deterministic dynamical evolution in phase space according to Hamilton's equations, which read

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \mathbf{M}^{-1}\,\mathbf{p}, \tag{3.31}$$

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \tag{3.32}$$

If the integrator is reversible, then the proposal is symmetric, and the acceptance probability for the new point $(\boldsymbol{\theta}', \mathbf{p}')$ follows the Metropolis rule (see equation (3.17)):

$$a(\boldsymbol{\theta}', \mathbf{p}', \boldsymbol{\theta}, \mathbf{p}) = \min\left[1\,; \frac{\mathcal{P}(\boldsymbol{\theta}', \mathbf{p}')}{\mathcal{P}(\boldsymbol{\theta}, \mathbf{p})}\right] = \min\left[1\,; \exp(-H(\boldsymbol{\theta}', \mathbf{p}') + H(\boldsymbol{\theta}, \mathbf{p}))\right]. \tag{3.33}$$

Using the results of sections 3.4.1 and 3.4.2, this proves that detailed balance is verified and that HMC leaves the canonical distribution invariant.

In exact Hamiltonian dynamics, the energy is conserved, and therefore, ideally, this procedure always provides an acceptance rate of unity. In practice, numerical errors can lead to a somewhat lower acceptance rate but HMC remains computationally much cheaper than standard MH techniques in which proposals are often refused. In the end, we discard the momenta and yield the target parameters by marginalization:

$$\mathcal{P}(\boldsymbol{\theta}) = \int \mathcal{P}(\boldsymbol{\theta}, \mathbf{p})\,\mathrm{d}\mathbf{p}. \tag{3.34}$$

Applications of HMC in cosmology include: the determination of cosmological parameters (Hajian, 2007; in combination with Pico, Fendt & Wandelt, 2007), CMB power spectrum inference (Taylor, Ashdown & Hobson, 2008) and Bayesian approach to non-Gaussianity analysis (Elsner & Wandelt, 2010), log-normal density reconstruction (Jasche & Kitaura, 2010; including from photometric redshift surveys, Jasche & Wandelt, 2012), dynamical, non-linear reconstruction of the initial conditions from galaxy surveys (Jasche & Wandelt, 2013a), joint power spectrum and bias model inference (Jasche & Wandelt, 2013b), inference of CMB lensing (Anderes, Wandelt & Lavaux, 2015).