Part II

Bayesian large-scale structure inference

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.

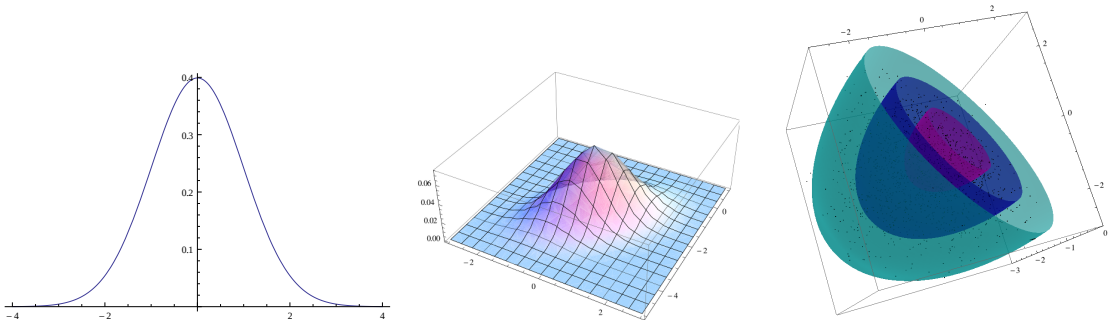

Figure 3.1: Example visualizations of posterior densities in low-dimensional parameter spaces (from left to right: one, two and three).

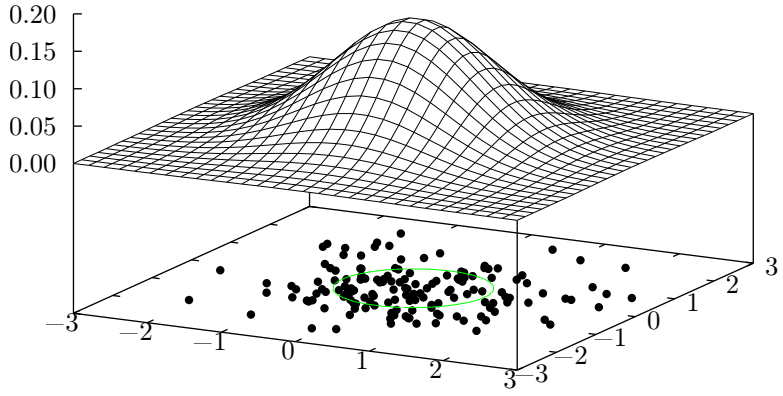

Figure 3.2: Example of a sampled representation of a posterior distribution in two dimensions. A set of samples is constructed in such a way that at any point, the posterior probability is proportional to the local density of samples in parameter space.

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:

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

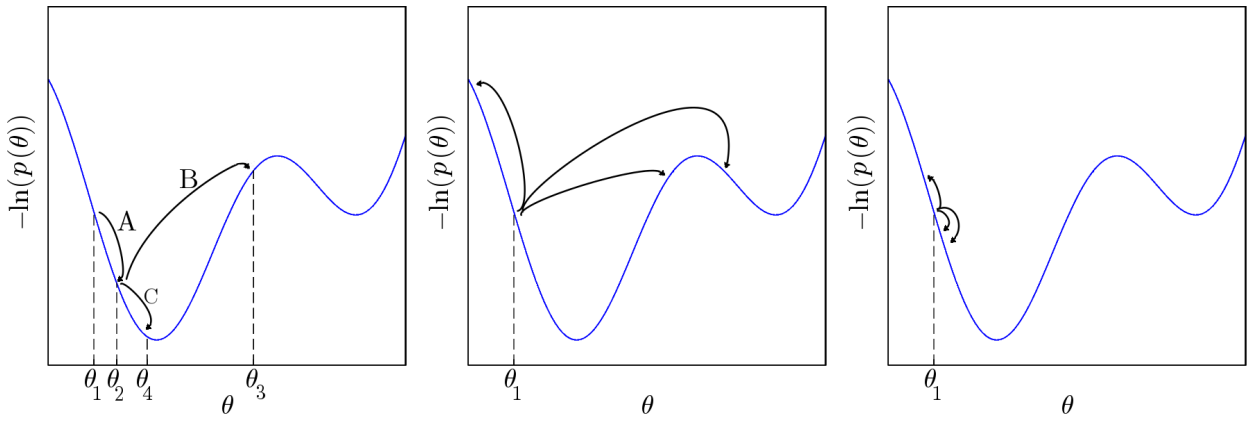

Figure 3.3: *Left panel.* An example of Markov chain constructed by the Metropolis-Hastings algorithm: starting at θ_1, θ_2 is proposed and accepted (step A), θ_3 is proposed and refused (step B), θ_4 is proposed and accepted (step C). The resulting chain is $\{\theta_1, \theta_2, \theta_2, \theta_4, ...\}$. *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

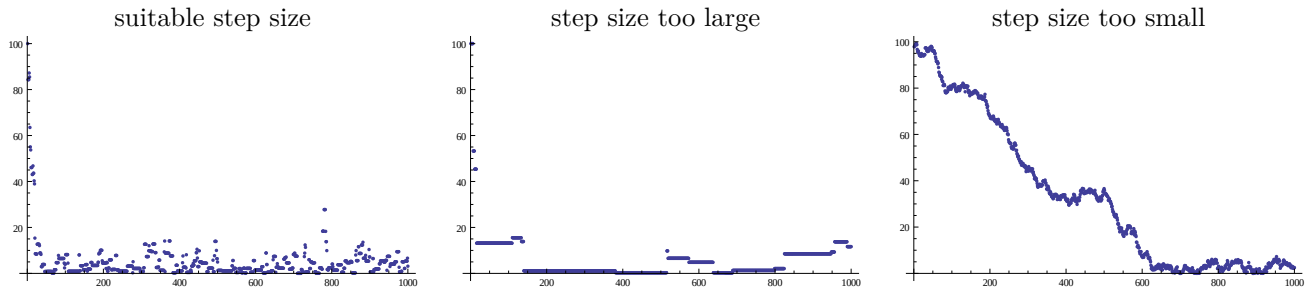

Figure 3.4: Example of Markov chains constructed by the Metropolis-Hastings algorithm, sampling the same target distribution but with varying proposal distribution (step size). The plots show the value of the sampled parameter as a function of the position in the chain. The ideal behavior with a suitable 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).

Chapter 4

Physical large-scale structure inference with the BORG algorithm

Contents

> "We are the Borg. Lower your shields and surrender your ships. We will add your biological and technological distinctiveness to our own. Your culture will adapt to service us. Resistance is futile."
> — Star Trek: First Contact (1996)

Abstract

This chapter describes the development and implementation of the BORG algorithm, which aims at physical large-scale structure inference in the linear and mildly non-linear regime. It describes the data model, which jointly accounts for the shape of three-dimensional matter field and its formation history. Based on an efficient implementation of the Hamiltonian Monte Carlo algorithm, BORG samples the joint posterior of the several millions parameters involved, which allows for thorough uncertainty quantification.

This chapter presents BORG (Bayesian Origin Reconstruction from Galaxies), a data assimilation method for probabilistic, physical large-scale structure inference. In section 4.1, the main challenge faced, namely the curse of dimensionality, is discussed. In section 4.2, we describe the latest formulation of BORG data model, initially introduced by Jasche & Wandelt (2013a) and updated by Jasche, Leclercq & Wandelt (2015). Section 4.3 gives considerations about the sampling procedure and the numerical implementation of the algorithm. Finally, in section 4.4, we report on a test of the BORG algorithm using a synthetic catalog of tracers.

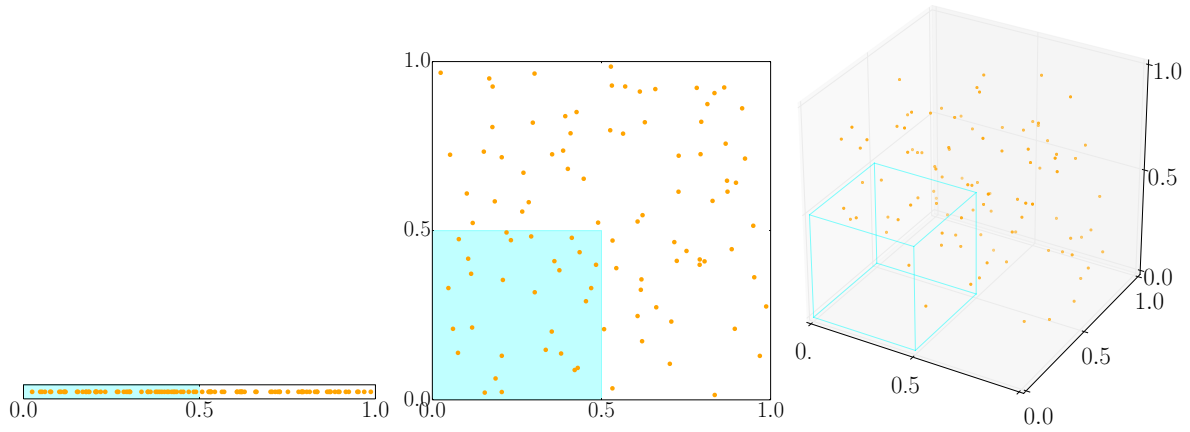

Figure 4.1: Illustration of the curse of dimensionality in one, two and three dimensions. We draw an original sample of 100 random points uniformly distributed in $[0;1]$, then progressively add a second and third coordinate, also uniformly drawn in $[0;1]$. The sparsity of the data (here illustrated by the number of samples in the $\left[0;\frac{1}{2}\right]$ hypercube, in cyan) increases exponentially with the number of dimensions.

Dimension D	$\mathcal{P}_D = 2^{-D}$	Numerical representation
1	2^{-1}	0.5
10	2^{-10}	9.77×10^{-4}
100	2^{-100}	7.89×10^{-31}
1000	2^{-1000}	9.33×10^{-302}
10000	2^{-10000}	0.

Table 4.1: Probability for a sample uniformly drawn in $[0;1]^D$ to be in $\left[0;\frac{1}{2}\right]^D$, as a function of the dimension D. The mathematical result, 2^{-D} (second column) is compared to its double-precision computer representation (third column). For $D \geq 1075$, $\mathcal{P}_D$ is below the minimum positive subnormal double representable.

4.1 The challenge: the curse of dimensionality

Statistical analyses of large-scale structure surveys require to go from the few parameters describing the homogeneous Universe to a point-by-point characterization of the inhomogeneous Universe. The latter description typically involves tens of millions of parameters: the density in each voxel of the discretized survey volume.

"Curse of dimensionality" phenomena (Bellman, 1961) are the significant obstacle in this high-dimensional data analysis problem. They refer to the difficulties caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space. In the following, we discuss the basic aspects of the high-dimensional situation. In particular, we outline three aspects of the curse of dimensionality phenomena.

4.1.1 Sparse sampling

The first and most obvious aspect is the exponential increase of sparsity given a fixed amount of sampling points. Reciprocally, the number of points drawn from a uniform distribution, needed for sampling at a constant density a region in parameter space, increases exponentially with its dimension.

We illustrate this phenomenon in figure 4.1 with 100 points randomly drawn in $[0;1]^D$ for $D = 1, 2, 3$. The number of samples that fall in some fixed region in parameter space exponentially decreases with the dimensionality of the problem. For example, the probability $\mathcal{P}_D$ for a random point to be in the $\left[0;\frac{1}{2}\right]^D$ hyperquadrant (shown in cyan in figure 4.1) is 2^{-D}. Difficulties to represent such probabilities numerically (table 4.1) arise well before $D = 10^7$, as we now discuss.

In standard double-precision binary floating-point format (the IEEE 754 "`binary64`" norm), numbers are represented in base $b = 2$. The bits are laid out as follows (figure 4.2): 1 sign bit, 11 bits for the exponent width, and $p = 52$ bits for the significand precision. The real value assigned by the machine to a set of `binary64` digits

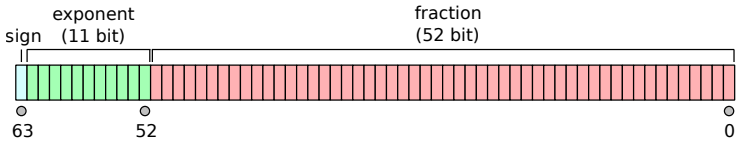

Figure 4.2: Computer representation of double-precision binary floating-point numbers. One bit is used to store the sign, 11 to store the exponent, and 52 bits to store the fractional part. This representation on a finite number of bits implies the existence of both a minimal and a maximal positive representable number.

is

$$(-1)^{\text{sign}} \left(1 + \sum_{i=1}^{52} b_{52-i} 2^{-i}\right) \times 2^{e-1023}, \tag{4.1}$$

where $1 \leq e \leq 2046$ is the "biased exponent" encoded in the 11 exponent bits and b_i are the values of the significand bits.

This representation implies that the maximum relative rounding error when rounding a number to the nearest representable one (the "machine epsilon") is $b^{-(p-1)} = 2^{-52}$. Therefore, the maximum positive double is `max_double` $\equiv (1+(1-2^{-52})) \times 2^{1023} \approx 1.798 \times 10^{308}$ and the minimum positive double is `min_normal_double` $\equiv 2^{-1022} \approx 2.225 \times 10^{-308}$.

In a normal floating-point value, there are no leading zeros in the significand; instead leading zeros are moved to the exponent. By using leading zeros in the significand, it is possible to represent "subnormal numbers", i.e. numbers where this representation would result in an exponent that is too small for the allowed number of bits. The smallest subnormal number representable with the `binary64` norm is `min_subnormal_double` $\equiv 2^{-52} \times 2^{-1022} \approx 4.941 \times 10^{-324}$.

Coming back to the representation of $\mathcal{P}_D$ is a large number of dimensions, the discussion above implies that $\mathcal{P}_D$ is exactly zero, at computer precision, for $D \geq 1075$. More generally, typical probabilities are often below `min_subnormal_double` for $D \gtrsim 1000$, which means that their computer representations as doubles is impossible. Representing such numbers requires more than 64 bits. This number of dimensions is well below that of the problem that we want to tackle, $D \approx 10^7$.

4.1.2 Shape of high-dimensional pdfs

Generally, high-dimensional functions can have more complex features than low-dimensional functions (there is more "space" for that), and hence can be harder to characterize.

Since it is not possible to store arbitrarily small positive numbers, numerical representations of high-dimensional pdfs will tend to have narrow support and very peaked features. This can also cause difficulties, as pdfs have to be normalized to unity: if the support is sufficiently small, the value of the pdf at its peaks can easily be above the maximum double `max_double`, which will cause computer crashes.

4.1.3 Algorithms in high dimensions

It is important to note that curse of dimensionality phenomena are generally not an intrinsic problem of high-dimensional problems, but a joint problem of the data set and the algorithm used. In particular, a dramatic increase of computational time (both to get one sample and to reach the required number of samples) is common. The curse of dimensionality often means that the number of samples available is small compared to the dimension of the space, which can lead to issues such as overfitting the data or getting poor classification or clustering when searching for specific patterns (Verleysen & François, 2005).

For most MCMC algorithms, the slow convergence, due a high rejection rate, is the most significant obstacle. In particular, for many interesting problems (typically non-linear and where components are not independently distributed), traditional sampling techniques that perform a random walk in parameter space, like the Metropolis-Hastings algorithm (see section 3.4.2) will unequivocally fail in $D \approx 10^7$.[1] However, gradients

[1] At least, unless the proposal distribution approximates extremely well the target distribution – which would imply to have already solved the problem!

Code	Density field model	Response operator	Multi-survey	$P(k)$	Photo-z	Galaxy bias model	b	$\tilde{N}$	RSD
ARES	Gaussian (J+10a)	(J+10a)	(JW13b)	(J+10a)		linear (J+10a); M-dep., linear (JW13b)	sampled (JW13b)	(JW13b)	(J+in prep.)
HADES	Log-normal (JK10)	(JK10)	(J+in prep.)	(J+in prep.)	(JW12)	linear (JK10)		(J+in prep.)	
BORG	2LPT (JW13a)	(JW13a)	(JLW15)			linear (JW13a); M-dep., power-law (JLW15)	calibrated with ARES (LJ16); sampled (J+in prep.)	(JLW15)	

Table 4.2: Current status of Bayesian large-scale structure analysis codes ARES, HADES and BORG. Green cells correspond to features implemented in the data model and tested, as reported in the corresponding papers. Blue cells correspond to features which will be described in upcoming publications. The column correspond respectively to: the model used to describe the prior density field; treatment of the survey response operator (survey mask and selection effects); treatment of multiple, independent surveys (or sub-samples of the same survey); power spectrum sampling; photometric redshifts sampling; galaxy bias model (M-dep. stands for luminosity-dependent bias); treatment of bias parameters; sampling of noise levels; treatment of peculiar velocities and redshift-space distortions. The references are J+10a = Jasche *et al.* (2010a); JK10 = Jasche & Kitaura (2010); JW12 = Jasche & Wandelt (2012); JW13a = Jasche & Wandelt (2013a); JW13b = Jasche & Wandelt (2013b); JLW15 = Jasche, Leclercq & Wandelt (2015); LJ16 = Lavaux & Jasche (2016).

of pdfs carry capital information, as they indicate the direction to high-density regions, permitting fast travel through a large volume in parameter space.

One way forward is to reduce the dimensionality of the problem, which is actually an entire research field. For example, principal component analysis converts a set of correlated variables to a set of linearly uncorrelated "principal components". Unfortunately, due to the highly non-linear and complex physics involved in structure formation (see chapter 1), no obvious reduction of the problem size exists in our case. Under the assumption of an initial grf with independent density amplitudes in Fourier space, we cannot make any further dimension reduction, and we have to deal with all $D \approx 10^7$ dimensions. Dimensionality can only be reduced by coarsening the resolution and discarding information.

As we will demonstrate in the rest of this chapter, Hamiltonian Monte Carlo (see section 3.4.3) beats the curse of dimensionality for the problem of physical large-scale structure inference. In particular, the approximate conservation of the Hamiltonian enables us to keep a high acceptance rate, and the use of gradients of the posterior pdf ($\partial\psi(\mathbf{\theta})/\partial\mathbf{\theta}$ in Hamilton's equations) allows efficient search for high density of probability regions.

4.2 The BORG data model

In this section, we discuss the BORG data model, i.e. the set of assumptions concerning the generation of observed large-scale structure data. In other words, we write down a probabilistic data-generating process.

This model was initially introduced by Jasche & Wandelt (2013a). In Jasche, Leclercq & Wandelt (2015), we updated the data model and modified to the original formulation of the BORG sampling scheme to introduce the improvements presented in Jasche & Wandelt (2013b). These improvements permit to account for luminosity-dependent galaxy bias and to perform automatic noise level calibration.

BORG is the successor of ARES (Algorithm for REconstruction and Sampling, Jasche *et al.*, 2010a; Jasche & Wandelt, 2013b) and HADES (HAmiltonian Density Estimation and Sampling Jasche & Kitaura, 2010; Jasche & Wandelt, 2012). In table 4.2, we summarize the different aspects covered by the ARES, HADES, and BORG data models. Contrary to ARES and HADES, which use phenomenological models to describe the density field, BORG involves a physical structure formation model (see table 4.2). LSS observations are merged with actual dynamics. Therefore, even if it is the least advanced algorithm in terms of the aspects covered by the data

model, its physical modeling is the most sophisticated.

In the following, x labels one of the D voxels of the discretized domain, δ^{i} and δ^{f} are realizations of the initial (at $a = 10^{-3}$) and final (at $a = 1$) density contrast, respectively, expressed as D-dimensional vectors. For improved clarity, we use colors in equations to distinguish the different quantities that are involved in the data model.

4.2.1 The physical density prior

In contrast to earlier algorithms (see table 4.2) BORG includes a physical density prior i.e. involves a model for structure formation. This makes the prior (expressed in terms of the final density contrast) highly non-Gaussian and non-linear. Writing down this prior is the subject of the present section.

4.2.1.1 The initial Gaussian prior

As discussed in the introduction and in chapter 1, it is commonly admitted that the density contrast early in the matter era obeys Gaussian statistics. Consistently with the discussion of section 3.2.2, this is the prior that we adopt.

Explicitly, in Fourier space, the prior for the initial density contrast is a multivariate Gaussian process with zero mean and diagonal covariance matrix $\hat{S}$ (see equation (1.14)):

$$\mathcal{P}(\hat{\delta}^{\mathrm{i}}|\hat{S}) = \frac{1}{\sqrt{\left|2\pi\hat{S}\right|}} \exp\left(-\frac{1}{2}\sum_{k,k'} \hat{\delta}^{\mathrm{i}}_{k} \hat{S}^{-1}_{kk'} \hat{\delta}^{\mathrm{i}}_{k'}\right). \tag{4.2}$$

where we explicitly noted by a hat the Fourier-space quantities.

The elements in matrix $\hat{S}$ are fixed parameters in BORG. They characterize the variance of the initial density field and therefore contain a cosmological dependence. We further assume that the covariance matrix $\hat{S}$ is diagonal in Fourier space (this is assuming statistical homogeneity of the initial density contrast, as seen in section 1.2.4.1). The diagonal coefficients are $\sqrt{P(k)/(2\pi)^{3/2}}$, where $P(k)$ are the initial power spectra coefficients for the adopted fiducial cosmological parameters. They are chosen to follow the prescription of Eisenstein & Hu (1998, 1999), including baryonic wiggles.

Alternatively, using the configuration space representation yields

$$\mathcal{P}(\delta^{\mathrm{i}}|S) = \frac{1}{\sqrt{|2\pi S|}} \exp\left(-\frac{1}{2}\sum_{x,x'} \delta^{\mathrm{i}}_{x} S^{-1}_{xx'} \delta^{\mathrm{i}}_{x'}\right). \tag{4.3}$$

4.2.1.2 Translating to the final density field

Following Jasche & Wandelt (2013a), we now show that the problem of physical inference of final density fields can be recast into the problem of inferring the corresponding initial conditions, given the structure formation model.

As seen before, it is straightforward to express a prior in the initial conditions, $\mathcal{P}(\delta^{\mathrm{i}})$. Given this, we can obtain a prior distribution for the final density contrast at scale factor a by using the standard formula for conditional probabilities:

$$\mathcal{P}(\delta^{\mathrm{f}}) = \int \mathcal{P}(\delta^{\mathrm{f}}, \delta^{\mathrm{i}})\, \mathrm{d}\delta^{\mathrm{i}} \tag{4.4}$$

$$= \int \mathcal{P}(\delta^{\mathrm{f}}|\delta^{\mathrm{i}})\, \mathcal{P}(\delta^{\mathrm{i}})\, \mathrm{d}\delta^{\mathrm{i}}. \tag{4.5}$$

For a deterministic model of structure formation $\delta^{\mathrm{i}} \mapsto \mathcal{G}\left(\delta^{\mathrm{i}}, a\right)$, the conditional probability is given by Dirac delta distributions:

$$\mathcal{P}(\delta^{\mathrm{f}}|\delta^{\mathrm{i}}) = \prod_{x} \delta_{\mathrm{D}}\left(\delta^{\mathrm{f}}_{x} - \left[\mathcal{G}(\delta^{\mathrm{i}}, a)\right]_{x}\right). \tag{4.6}$$

Therefore, given a model $\mathcal{G}$ for structure formation, a prior distribution for the late-time density field can be obtained by a two-step sampling process:

1. drawing an initial condition realization from the prior $\mathcal{P}(\delta^{\rm i})$;
2. propagating the initial state forward in time with $\mathcal{G}$ (this step is entirely deterministic).

This process amounts to drawing samples from the joint prior distribution of initial and final conditions:

$$\mathcal{P}(\delta^{\rm f}, \delta^{\rm i}) = \mathcal{P}(\delta^{\rm i}) \prod_x \delta_{\rm D}\left(\delta^{\rm f}_x - \left[\mathcal{G}(\delta^{\rm i}, a)\right]_x\right). \tag{4.7}$$

Marginalization over initial density realizations then yields samples of the non-Gaussian prior for final density fields. In practice, as initial conditions are also interesting for a variety of cosmological applications, we do not discard them and we always store them, whenever we draw a sample from the prior.

4.2.1.3 The structure formation model

Ideally, the structure formation model should be fully non-linear gravity. For reasons of computational feasibility, in BORG, $\mathcal{G}$ is obtained from second-order Lagrangian perturbation theory and the cloud-in-cell scheme. More specifically, the initial density field is populated by dark matter particles that are evolved according to the equations for 2LPT displacements given in section 1.5.3. In the final state, these particles are assigned to the grid using a CiC scheme, yielding the final density contrast $\delta^{\rm f}$. The reader is referred to appendix B for details on the numerical implementation of 2LPT and CiC.

Using equations (4.3) and (4.7), the joint physical prior for initial and late-time density fields is found to be

$$\mathcal{P}(\delta^{\rm f}, \delta^{\rm i}|S) = \frac{1}{\sqrt{|2\pi S|}} \exp\left(-\frac{1}{2}\sum_{x,x'} \delta^{\rm i}_x S^{-1}_{xx'} \delta^{\rm i}_{x'}\right) \prod_x \delta_{\rm D}\left(\delta^{\rm f}_x - \left[\mathcal{G}(\delta^{\rm i}, a)\right]_x\right). \tag{4.8}$$

Note that the first part (corresponding to the initial conditions) is more easily handled in Fourier space, while the second part (corresponding to the propagation from initial to final conditions) involves density fields in configuration space.

4.2.2 The large-scale structure likelihood

This section discusses the BORG likelihood, $\mathcal{P}(d|\delta^{\rm i})$. The data d used by BORG are galaxy (or matter tracer) number counts in each voxel of the discretized domain. To compute it, the position of galaxies is translated from spherical to Cartesian coordinates using the following coordinate transform:

$$x = d_{\rm com}(z)\cos(\lambda)\cos(\eta), \tag{4.9}$$
$$y = d_{\rm com}(z)\cos(\lambda)\sin(\eta), \tag{4.10}$$
$$z = d_{\rm com}(z)\sin(\lambda), \tag{4.11}$$

with λ being the declination, η the right ascension and $d_{\rm com}(z)$ the radial comoving distance to redshift z for the fiducial cosmology. Galaxies are then binned using the Nearest Grid Point (NGP) assignment scheme to get voxel-wise galaxy number counts.

4.2.2.1 Splitting the galaxy distribution

In order to account for the luminosity-dependence of selection effects and galaxy biases, we split the data into several bins of absolute magnitude. In the following, ℓ labels one of these bins, and N^ℓ is the data set containing the number counts of galaxies in the luminosity bin ℓ and in voxel x, N^ℓ_x.

BORG treats different magnitude bins as independent data sets. Each of them is assigned a likelihood function, $\mathcal{P}(N^\ell|\delta^{\rm i})$. Since it is fair to assume that galaxies in different luminosity bins are independent and identically distributed, once the density field is given, the final likelihood of the total data set $d = \{N^\ell\}$ is obtained by multiplying these likelihood functions,

$$\mathcal{P}(d|\delta^{\rm i}) = \prod_\ell \mathcal{P}(N^\ell|\delta^{\rm i}). \tag{4.12}$$

4.2.2.2 The galaxy distribution as an inhomogeneous Poisson process

Galaxies are tracers of the mass distribution. The statistical uncertainty due to the discrete nature of their distribution is often modeled as a Poisson process (Layzer, 1956; Peebles, 1980; Martínez & Saar, 2002). Before BORG, Poissonian likelihoods have been successfully applied to perform reconstructions of the matter density by Kitaura, Jasche & Metcalf (2010); Jasche & Kitaura (2010); Jasche *et al.* (2010a). Adopting this picture, we write

$$\mathcal{P}(N^\ell|\lambda(\delta^{\rm i})) = \prod_x \frac{\exp\left(-\lambda^\ell_x(\delta^{\rm i})\right)\left(\lambda^\ell_x(\delta^{\rm i})\right)^{N^\ell_x}}{N^\ell_x!}. \tag{4.13}$$

The Poisson intensity field, $\lambda^\ell(\delta^{\rm i})$, characterizes the expected number of galaxies in voxel x given the initial density contrast $\delta^{\rm i}$. As it depends on the position, it is an *inhomogeneous* Poisson process.

Real galaxy samples can have a sub- or super-Poissonian behavior (i.e. be under- or over-dispersed), depending on local and non-local properties (Mo & White, 1996; Somerville *et al.*, 2001; Casas-Miranda *et al.*, 2002). These effects are neglected here, but in the context of large-scale structure reconstructions, deviations from Poissonity have been introduced in the likelihood by Kitaura (2012); Ata, Kitaura & Müller (2015).

4.2.2.3 The Poisson intensity field

The expected number of galaxies in a voxel depends – of course – on the underlying large-scale structure, but also on galaxy bias, redshift-space distortions, dynamical processes along the observer's backwards lightcone, selection effects, and instrumental noise. All these effects should in principle be taken into account in the Poisson intensity field. In the following, we detail, step by step, how to go from $\delta^{\rm i}$ to $\lambda(\delta^{\rm i})$ in the BORG likelihood.

1. **Structure formation.** The first step is to translate initial to evolved dark matter overdensity:

$$\delta^{\rm i} \mapsto \mathcal{G}(\delta^{\rm i}, a). \tag{4.14}$$

   As discussed before, for this step BORG relies on 2LPT instead of fully non-linear gravitational dynamics, meaning that there exists some degree of approximation in the inference process. Accurate quantification this level of approximation is unfortunately not currently possible, as it would require the fully non-linear inference process for reference, which so far is not computationally tractable.

2. **Lightcone effects.** Along with step 1, we could account for lightcone effects so that the distant structures are less evolved than the closest ones. This is exploiting the dependence of $\mathcal{G}$ on a to build the dark matter density on the lightcone. For simplicity, this is not currently implemented in BORG; rather, we run 2LPT up to $a = 1$ everywhere. In the following we simplify the notations and we write $\delta^{\rm f} \equiv \mathcal{G}(\delta^{\rm i}) \equiv \mathcal{G}(\delta^{\rm i}, a = 1)$.

3. **Redshift-space distortions.** At this point, the data model could also include a treatment of redshift-space distortions (see Heavens & Taylor, 1995; Tadros *et al.*, 1999; Percival, Verde & Peacock, 2004; Percival, 2005a; Percival & White, 2009). Though not explicitly included in the present BORG data model, we find empirically that redshift-space distortions are mitigated by the prior preference for homogeneity and isotropy (see chapter 5): BORG interprets deviations from isotropy as noise, and fits an isotropic distribution to the data.

4. **Galaxy bias.** The following step is to get the galaxy density $\rho_{\rm g}$ given the dark matter density ρ. This is making assumptions for physical biasing in galaxy formation. Various LSS inference algorithms assume a linear bias model. In order to be well defined, a Poisson likelihood requires intensities of the inhomogeneous Poisson process to be strictly positive. Since a linear bias model does not guarantee a positive density field and corresponding Poisson intensity, it is not applicable to the present case. For this reason, we assume a phenomenological power-law to account for galaxy biasing:

$$\rho_g \propto \beta\rho^\alpha. \tag{4.15}$$

   In luminosity bin ℓ and in terms of the dark matter overdensity, this is step written

$$\delta^{\rm f} \mapsto \beta^\ell(1+\delta^{\rm f})^{\alpha^\ell} \propto \rho^\ell_g. \tag{4.16}$$

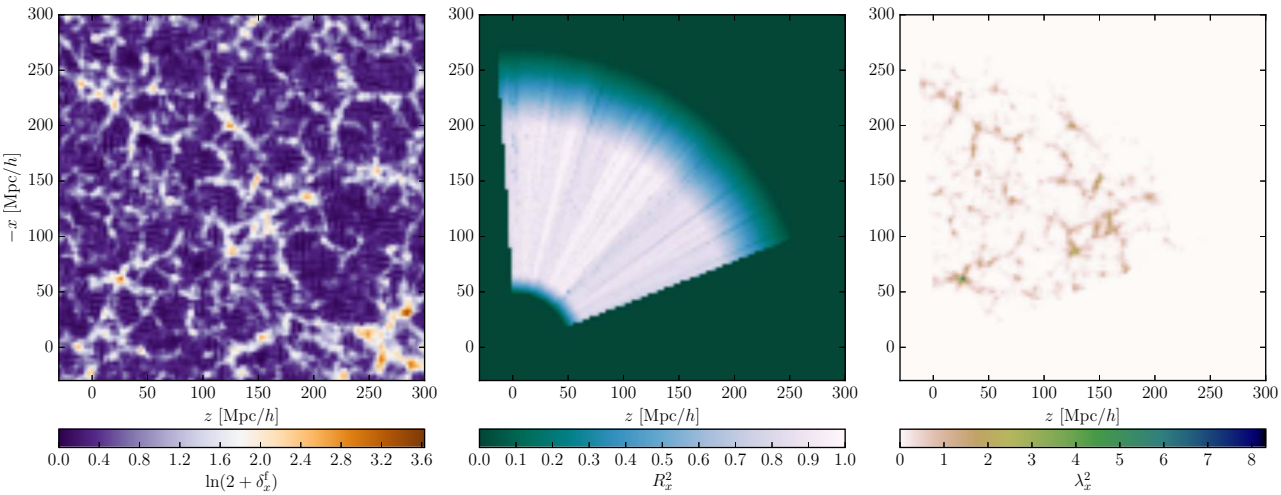

Figure 4.3: Slices through the box used in the BORG SDSS analysis (see chapter 5). *Left panel.* Density in one sample (for clarity, the quantity shown is $\ln(2+\delta_x^{\mathrm{f}})$). *Middle panel.* Survey response operator R_x^2 in the $\ell = 2$ luminosity bin, corresponding to absolute r-band magnitudes in the range $-19.67 < M_{0.1r}^2 < -19.00$. *Right panel.* Poisson intensity field λ_x^2 for this sample and luminosity bin, computed with equation (4.20). The bias and noise parameters are respectively $\alpha^2 = 1.30822$ and $\widetilde{N}^2 = 1.39989$ (see table 5.1).

Note that coefficients α^ℓ and β^ℓ depend on ℓ, which means that the data model accounts for *luminosity-dependent* galaxy biases. Parameters β^ℓ are automatically calibrated during the generation of the Markov Chain (see section 4.3.1). For simplicity, parameters α^ℓ are kept at fixed, fiducial values. In the BORG analysis of the SDSS (chapter 5), these values are determined using a standard model for luminosity-dependent galaxy bias. In their analysis of the 2M++ catalog (Lavaux & Hudson, 2011), Lavaux & Jasche (2016) show that it is possible to calibrate these values with a preliminary ARES inference, for subsequent use in BORG.

5. **Mean number of galaxies.** To get the expected number of galaxies from the unnormalized galaxy density, the quantity $\beta^\ell(1+\delta^{\mathrm{f}})^{\alpha^\ell}$ has to be multiplied by the mean number of galaxies in bin ℓ, $\bar{N}^\ell$. This step is therefore simply:

$$\beta^\ell(1+\delta^{\mathrm{f}})^{\alpha^\ell} \mapsto \bar{N}^\ell\beta^\ell(1+\delta^{\mathrm{f}})^{\alpha^\ell}. \tag{4.17}$$

6. **Observational effects.** The last step is to put in the luminosity-dependent selection effects and the survey mask. For this, we multiply with the linear survey response operator R_x^ℓ, a voxel-wise three-dimensional function that incorporates survey geometries and selection effects:

$$\bar{N}^\ell\beta^\ell(1+\delta_x^{\mathrm{f}})^{\alpha^\ell} \mapsto R_x^\ell\bar{N}^\ell\beta^\ell(1+\delta_x^{\mathrm{f}})^{\alpha^\ell}. \tag{4.18}$$

Eventually, the Poisson intensity field is given by

$$\lambda_x^\ell(\delta^{\mathrm{i}}) = R_x^\ell\bar{N}^\ell\beta^\ell\left(1+\left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x\right)^{\alpha^\ell}. \tag{4.19}$$

We note that $\bar{N}^\ell$ and β^ℓ are degenerate, in the sense that only the product $\bar{N}^\ell\beta^\ell$ matters. We define $\widetilde{N}^\ell \equiv \bar{N}^\ell\beta^\ell$, so that

$$\lambda_x^\ell(\delta^{\mathrm{i}}) = R_x^\ell\widetilde{N}^\ell\left(1+\left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x\right)^{\alpha^\ell}. \tag{4.20}$$

$\widetilde{N}^\ell$ represents the overall noise level in bin ℓ. With the improved BORG data model (Jasche, Leclercq & Wandelt, 2015), we automatically calibrate this parameter (see section 4.3.1). In figure 4.3, we illustrate the construction of the Poisson intensity field for the $\ell = 2$ bin of the SDSS analysis. We show the dark matter density, δ_x^{f}, the survey response operator R_x^2 and the Poisson intensity λ_x^2.

4.2.2.4 The comprehensive large-scale structure likelihood

Noting $d \equiv \{N^\ell\}$ the total data set, i.e. all available galaxy number counts, and $\widetilde{N} \equiv \{\widetilde{N}^\ell\}$ the set of noise parameters in each bin, we obtain the final expression for the LSS likelihood using equations (4.12), (4.13) and (4.20). It reads

$$\mathcal{P}(d|\delta^{\mathrm{i}}, \widetilde{N}) = \prod_{x,\ell} \frac{\exp\left(-R_x^\ell \widetilde{N}^\ell (1 + \left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x)^{\alpha^\ell}\right) \left(R_x^\ell \widetilde{N}^\ell (1 + \left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x)^{\alpha^\ell}\right)^{N_x^\ell}}{N_x^\ell !} \tag{4.21}$$

In this equation, we omitted on the right side of the conditioning bar the sets $\{R_x^\ell\}$ and $\{\alpha^\ell\}$ (one can consider that all probabilities inferred by BORG are conditional on these). However, we now write explicitly $\widetilde{N}$, as this will be of importance later.

4.2.3 The posterior distribution

As usual in Bayesian statistics, the posterior distribution is obtained, up to a normalization constant, by the use of Bayes' formula,

$$\mathcal{P}(\delta^{\mathrm{i}}|d, S, \widetilde{N}) \propto \mathcal{P}(\delta^{\mathrm{i}}|S, \widetilde{N})\, \mathcal{P}(d|\delta^{\mathrm{i}}, S, \widetilde{N}) = \mathcal{P}(\delta^{\mathrm{i}}|S)\, \mathcal{P}(d|\delta^{\mathrm{i}}, \widetilde{N}). \tag{4.22}$$

Substituting equations (4.3) and (4.21) allows to write down the full problem solved by BORG for the density distribution:

$$\mathcal{P}(\delta^{\mathrm{i}}|d, S, \widetilde{N}) \propto \frac{1}{\sqrt{|2\pi S|}} \exp\left(-\frac{1}{2} \sum_{x,x'} \delta_x^{\mathrm{i}} S^{-1}_{xx'} \delta^{\mathrm{i}}_{x'}\right) \prod_{x,\ell} \frac{\exp\left(-R_x^\ell \widetilde{N}^\ell (1 + \left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x)^{\alpha^\ell}\right) \left(R_x^\ell \widetilde{N}^\ell (1 + \left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x)^{\alpha^\ell}\right)^{N_x^\ell}}{N_x^\ell !}. \tag{4.23}$$

It is simpler to express the BORG posterior in terms of the initial conditions, but recall that one gets the final conditions (and in fact the entire LSS history, as demonstrated in chapter 5) automatically and entirely deterministically via the structure formation model $\mathcal{G}$ (see section 4.2.1.2):

$$\mathcal{P}(\delta^{\mathrm{f}}, \delta^{\mathrm{i}}|d, S, \widetilde{N}) = \mathcal{P}(\delta^{\mathrm{i}}|d, S, \widetilde{N}) \prod_x \delta_{\mathrm{D}}\left(\delta_x^{\mathrm{f}} - \left[\mathcal{G}(\delta^{\mathrm{i}})\right]_x\right). \tag{4.24}$$

4.2.4 The Γ-distribution for noise sampling

This section draws from appendix A of Jasche, Leclercq & Wandelt (2015).

We aim at automatically calibrating, during the sampling procedure, the noise level of each luminosity bin, given the data and the current density sample. This requires to write down the conditional probability $\mathcal{P}(\widetilde{N}^\ell|N^\ell, \delta^{\mathrm{f}})$, which we do in this section.

According to Bayes' formula, we can write

$$\mathcal{P}(\widetilde{N}^\ell|N^\ell, \delta^{\mathrm{f}}) \propto \mathcal{P}(\widetilde{N}^\ell)\, \mathcal{P}(N^\ell|\widetilde{N}^\ell, \delta^{\mathrm{f}}), \tag{4.25}$$

where we have assumed the conditional independence $\mathcal{P}(\widetilde{N}^\ell|\delta^{\mathrm{f}}) = \mathcal{P}(\widetilde{N}^\ell)$. In the absence of any further information on the parameter $\widetilde{N}^\ell$, we follow the maximum agnostic approach pursued by Jasche & Wandelt (2013b) by setting the prior distribution $\widetilde{N}^\ell$ constant. By using the Poisson likelihood for $\mathcal{P}(N^\ell|\widetilde{N}^\ell, \delta^{\mathrm{f}})$ (equations (4.13) and (4.20)) into equation (4.25), we obtain the conditional posterior for the noise parameter $\widetilde{N}^\ell$ as:

$$\mathcal{P}(\widetilde{N}^\ell|N^\ell, \delta^{\mathrm{f}}) \propto \exp\left(-\widetilde{N}^\ell A_\ell\right) \times \left(\widetilde{N}^\ell\right)^{B_\ell}, \tag{4.26}$$

where $A_\ell \equiv \sum_x R_x^\ell (1 + \delta_x^{\mathrm{f}})^{\alpha^\ell}$ and $B_\ell \equiv \sum_x N_x^\ell$. By choosing $k_\ell \equiv B_\ell + 1$ and $\theta_\ell \equiv 1/A_\ell$, we yield a properly normalized Γ-distribution for the noise parameter $\widetilde{N}^\ell$, given as:

$$\mathcal{P}(\widetilde{N}^\ell|N^\ell, \delta^{\mathrm{f}}) = \Gamma[k_\ell, \theta_\ell]\left(\widetilde{N}^\ell\right) = \frac{\left(\widetilde{N}^\ell\right)^{k_\ell - 1} \exp\left(-\frac{\widetilde{N}^\ell}{\theta_\ell}\right)}{\theta_\ell^{k_\ell} \Gamma(k_\ell)}. \tag{4.27}$$

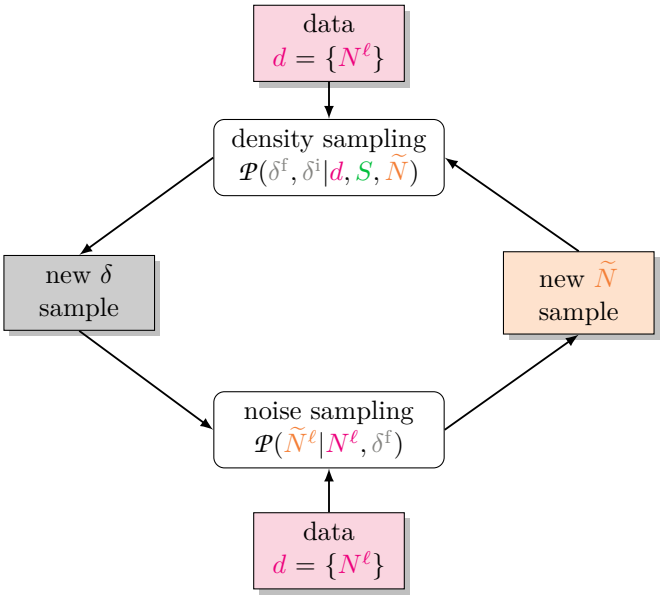

Figure 4.4: Flow chart depicting the multi-step iterative block sampling procedure. In the first step, BORG generates random realizations of the initial and final density fields conditional on the galaxy samples d and on the noise levels $\{\widetilde{N}^\ell\}$. In a subsequent step, the noise parameters $\widetilde{N}^\ell$ are sampled conditional on the previous density realizations.

with shape parameter

$$k_\ell \equiv 1 + \sum_x N^\ell_x, \tag{4.28}$$

and scale parameter

$$\theta_\ell \equiv \frac{1}{\sum_x R^\ell_x (1+\delta^{\rm f}_x)^{\alpha^\ell}}. \tag{4.29}$$

4.3 Sampling procedure and numerical implementation

4.3.1 Calibration of the noise level

This section draws from section 3.2. in Jasche, Leclercq & Wandelt (2015).

Following the approach described in Jasche & Wandelt (2013b), density fields and noise level parameters can be jointly inferred by introducing an additional sampling block to the original implementation of the BORG algorithm. The additional sampling block is designed to provide random samples of the noise parameters $\widetilde{N}^\ell$ given the galaxy data set N^ℓ and the current final density sample $\delta^{\rm f}$.

As indicated by figure 4.4, in a first step, the algorithm infers density fields, then conditionally samples the noise parameters. Iteration of this procedure yields Markovian samples from the joint target distribution.

As demonstrated in section 4.2.4, the posterior distributions of noise parameters $\widetilde{N}^\ell$ are Γ-distributions. In the new sampling block, random variates of the Γ-distribution are generated by standard routines provided by the GNU scientific library (Galassi *et al.*, 2003).

4.3.2 Hamiltonian Monte Carlo and equations of motion for the LSS density

Sampling of the posterior distribution for density fields is achieved via Hamiltonian Monte Carlo. As described in section 3.4.3, HMC permits to explore the non-linear posterior by following Hamiltonian dynamics in the high-dimensional parameter space. Omitting normalization constants, the Hamiltonian potential $\psi(\delta^{\rm i})$ can be written as:

$$\psi(\delta^{\rm i}) = -\ln \mathcal{P}(\delta^{\rm i}|d, S, \widetilde{N}) - \ln Z \tag{4.30}$$

$$= \psi_{\rm prior}(\delta^{\rm i}) + \psi_{\rm likelihood}(\delta^{\rm i}), \tag{4.31}$$

with the "prior potential" $\psi_{\rm prior}(\delta^{\rm i})$ given as

$$\psi_{\rm prior}(\delta^{\rm i}) = \frac{1}{2}\sum_{x,x'} \delta^{\rm i}_x S^{-1}_{xx'} \delta^{\rm i}_{x'}, \tag{4.32}$$

and the "likelihood potential" $\psi_{\rm likelihood}(\delta^{\rm i})$ given as

$$\psi_{\rm likelihood}(\delta^{\rm i}) = \sum_{x,\ell} R^\ell_x \widetilde{N}^\ell \left(1 + \left[\mathcal{G}(\delta^{\rm i})\right]_x\right)^{\alpha^\ell} - N^\ell_x \ln\left(R^\ell_x \widetilde{N}^\ell \left(1 + \left[\mathcal{G}(\delta^{\rm i})\right]_x\right)^{\alpha^\ell}\right). \tag{4.33}$$

Given the above definitions of the potential $\psi(\delta^{\rm i})$, one can obtain the required Hamiltonian force (see equation (3.32)) by differentiating with respect to $\delta^{\rm i}_x$:

$$\frac{\partial\psi(\delta^{\rm i})}{\partial\delta^{\rm i}_x} = \frac{\partial\psi_{\rm prior}(\delta^{\rm i})}{\partial\delta^{\rm i}_x} + \frac{\partial\psi_{\rm likelihood}(\delta^{\rm i})}{\partial\delta^{\rm i}_x}. \tag{4.34}$$

The prior term is given by

$$\frac{\partial\psi_{\rm prior}(\delta^{\rm i})}{\partial\delta^{\rm i}_x} = \sum_{x'} S^{-1}_{xx'}\delta^{\rm i}_{x'} \tag{4.35}$$

The likelihood term cannot be obtained trivially. However, the choice of 2LPT and a CiC kernel to model $\mathcal{G}(\delta^{\rm i})$ makes possible to derive this term analytically. This is of crucial importance, because a numerical estimation of gradients is very expensive. A detailed computation can be found in appendix D of Jasche & Wandelt (2013a). The result is

$$\frac{\partial\psi_{\rm likelihood}(\delta^{\rm i})}{\partial\delta^{\rm i}_x} = -D_1 J_x + D_2 \sum_{a>b}\left(\tau^{aabb}_x + \tau^{bbaa}_x - 2\tau^{abab}_x\right), \tag{4.36}$$

where D_1 and D_2 are the first and second-order growth factors at the desired time ($a = 1$), and J_x and τ^{abcd}_x are a vector and a tensor that depend on R^ℓ_x, $\widetilde{N}^\ell$, α^ℓ, N^ℓ_x.

Finally, the equations of motion for the Hamiltonian system can be written as

$$\frac{{\rm d}\delta^{\rm i}_x}{{\rm d}t} = \sum_{x'} M^{-1}_{xx'}\, p_{x'}, \tag{4.37}$$

$$\frac{{\rm d}p_x}{{\rm d}t} = -\sum_{x'} S^{-1}_{xx'}\delta^{\rm i}_{x'} + D_1 J_x(\delta^{\rm i}) - D_2 \sum_{a>b}\left(\tau^{aabb}_x(\delta^{\rm i}) + \tau^{bbaa}_x(\delta^{\rm i}) - 2\tau^{abab}_x(\delta^{\rm i})\right) \tag{4.38}$$

4.3.3 The mass matrix

As mentioned in section 3.4.3, the HMC algorithm possesses a large number of tunable parameters contained in the mass matrix M, whose choice can strongly impact the efficiency of the sampler. As shown in Jasche & Wandelt (2013a, section 5.2 and appendix F), a good approach to obtain suitable masses is to perform a stability analysis of the numerical leapfrog scheme (see section 4.3.4) implemented as integrator. This results in the following prescription:

$$M_{xx'} \equiv S^{-1}_{xx'} - \delta_{\rm K}^{xx'} D_1 \frac{\partial J_x(\delta^{\rm i})}{\partial\delta^{\rm i}_x}(\xi_x), \tag{4.39}$$

where $\delta_{\rm K}$ is a Kronecker delta symbol and ξ_x is assumed to be the mean initial density contrast in high probability regions, i.e. once the sampler has moved beyond the burn-in phase.

Due to the high-dimensionality of the problem, inverting M and storing M^{-1} is computationally impractical. Therefore, a diagonal mass matrix is constructed from equation (4.39).

4.3.4 The leapfrog scheme integrator

For computer implementation, Hamilton's equations, (4.37) and (4.38), must be approximated by discretizing time, using some small stepsize, ε. Several choices of integrator, such as the popular Euler's method, are possible (see section B.5.1).

As discussed in section 3.4.3, it is essential that the adopted scheme respect reversibility and symplecticity, to ensure incompressibility in phase space. Additionally, achieving high acceptance rates require the numerical integration scheme to be very accurate in order to conserve the Hamiltonian. For these reasons, the integrator adopted for implementing BORG is the leapfrog scheme (e.g. Birdsall & Langdon, 1985), which relies on a sequence of "kick–drift–kick" operations that work as follows (see also figure B.3):

$$p_x\left(t+\frac{\varepsilon}{2}\right) = p_x(t) - \frac{\varepsilon}{2}\frac{\partial\psi(\delta^{\mathrm{i}})}{\partial\delta^{\mathrm{i}}_x}\left(\delta^{\mathrm{i}}_x(t)\right), \tag{4.40}$$

$$\delta^{\mathrm{i}}_x(t+\varepsilon) = \delta^{\mathrm{i}}_x(t) + \varepsilon\frac{p_x\left(t+\frac{\varepsilon}{2}\right)}{m_x}, \tag{4.41}$$

$$p_x(t+\varepsilon) = p_x\left(t+\frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2}\frac{\partial\psi(\delta^{\mathrm{i}})}{\partial\delta^{\mathrm{i}}_x}\left(\delta^{\mathrm{i}}_x(t+\varepsilon)\right), \tag{4.42}$$

where m_x is the element of the diagonal mass matrix at position x.

The equations of motion are integrated by making n such steps with a finite step size ε. In order to prevent resonant trajectories, time steps are slightly randomized (ε is randomly drawn from a uniform distribution).

4.4 Testing BORG

Demonstrating of the performance of the BORG algorithm is the subject of sections 6 and 7 in Jasche & Wandelt (2013a). As these results are relevant to set the BORG SDSS analysis on firm statistical grounds, in the following, we briefly report on the original test using mock observations.

4.4.1 Generating mock observations

The first step is to generate an initial Gaussian random field (see section B.3). This was done on a three-dimensional Cartesian grid of 128^3 voxels covering a comoving cubic box of length 750 Mpc/h with periodic boundary conditions. The Fourier-space covariance matrix includes an Eisenstein & Hu (1998, 1999) cosmological power spectrum with baryonic wiggles. The cosmological parameters are fixed at fiducial values,

$$\Omega_\Lambda = 0.78, \Omega_{\mathrm{m}} = 0.22, \Omega_{\mathrm{b}} = 0.04, \sigma_8 = 0.807, h = 0.702, n_{\mathrm{s}} = 0.961. \tag{4.43}$$

The Gaussian initial conditions are populated by a Lagrangian lattice of 256^3 particles, that are propagated forward in time using the same implementation of second-order Lagrangian perturbation theory as used in BORG. The final density field is constructed from the resultant particle distribution using the cloud-in-cell scheme. Note that it is crucial to use the 2LPT model for structure formation at this point, instead of, for example, a full N-body simulation, in order to demonstrate that BORG correctly infers the input field. Only in this fashion can we demonstrate that the BORG complicated statistical machinery works, and compare the input and output without differences due to additional physics.

An artificial tracer catalog is then generated by simulating an inhomogeneous Poisson process characterized by equations (4.13) and (4.20) (see also figure 4.3 for an illustration). For the purpose of the test run, the problem is simplified to only one luminosity bin ($\ell = 0$), the mean number of galaxies $\bar{N}^0$ is fixed, and the tracers are supposed to be unbiased (which amounts to fixing $\alpha^0 = 1$, $\beta^0 = 1$). However, the survey response operator R^0_x involves a highly-structured survey mask (mimicking the geometry of the Sloan Digital Sky Survey data release 7) and realistic selection functions (based on standard Schechter luminosity functions), in order to demonstrate the possibility of doing large-scale structure inference from real data sets.

4.4.2 Convergence and correlations of the Markov Chain

As mentioned in section 3.4.3, HMC is designed to have the target distribution as its stationary distribution. Therefore, the sampling process provides samples of the posterior distribution (equation (4.23)) after an initial burn-in phase. Jasche & Wandelt (2013a) showed that during this phase, of the order of 600 samples, the power spectrum converges at all scales towards the true power in the initial density field. The absence of any power excess or deficiency demonstrates the correct treatment of the response operator. The analysis also showed that burn-in also manifests itself in the acceptance rate, which has a dip around after 100 samples, then increases and asymptotes at a constant value of around 84%.

Generally, successive samples of the chain will be correlated to previous samples. The correlation length of the chain determines the amount of independent samples that can be drawn from the total chain. Jasche & Wandelt (2013a) estimated the correlation length to about 200 samples and obtained a total of 15,000 samples; which amounts to around 72 independent samples after burn-in.

These statistical tests demonstrate that exploring the large-scale structure posterior is numerically feasible despite the high dimensionality of the problem.

4.4.3 Large-scale structure inference

This section discusses the large-scale structure inferred via the application of BORG to the synthetic data set. Figure 4.5 shows slices through various three-dimensional quantities: the true initial density field, one sample of initial conditions, the posterior mean for the initial density field; the same quantities for final density fields; the posterior standard deviation in the initial and final conditions; and the mock data set.

Comparison of initial and final density fields permits to check the correspondence between structures with growing statistical complexity. Furthermore, comparison of final density fields to the data demonstrates the accuracy of the inference of the underlying dark matter density field. In particular, one can see that the algorithm extrapolates unobserved filaments between clusters, based on the physical picture of structure formation provided by 2LPT. At high redshift or near the survey boundaries, complex structures appear continuous, which proves that the algorithm augments unobserved or poorly constrained regions with statistically correct information, consistently with the structure formation model. Therefore, each individual sample is a physical dark matter realization, to the level of accuracy of 2LPT.

The variation between samples quantifies joint and correlated uncertainties. This is illustrated in figure 4.5 by unobserved regions in the posterior means, where the values in different samples average to cosmic mean density, and by the posterior standard deviations. Therefore, contrary to other reconstruction approaches found in the literature, BORG possesses a demonstrated capability of quantifying uncertainty of inferred maps, locally and globally. These uncertainties can then be propagated to any derived quantity, as we demonstrate for example with cosmic web types in chapter 9.

Finally, Jasche & Wandelt (2013a) demonstrated that the inferred initial density contrast follows Gaussian one-point statistics, that inferred density fields cross-correlate with the true solution as expected (i.e. $R(k) \equiv P_{\delta_{\rm inferred}\times\delta_{\rm true}}/\sqrt{P_{\delta_{\rm inferred}}P_{\delta_{\rm true}}} \to 1$ as $k \to 0$), and that BORG also infers the underlying velocity field in detail.

4.5 Future extensions of BORG

The method described in this chapter forms the basis of a sophisticated, but also extensible, physical large-scale structure inference framework. In particular, natural extensions of the BORG algorithm would enable automatic calibration of bias parameters (the exponents α^ℓ in previous sections) and of the covariance matrix of initial fluctuations (the matrix S). This would allow precise inference of the early-time matter power spectrum from biased catalogs of tracers. As noted in the introduction, this endeavor could yield a vast gain of information for the determination of cosmological parameters, in comparison to state-of-the-art techniques.

Let us consider a set of comoving wavenumbers $\{k_n\}$ and let us denote by $P \equiv \{P(k_n)\}$ the set of corresponding power spectrum coefficients. Since direct sampling from $\mathcal{P}(P|d)$ is impossible, or at least difficult, Jasche *et al.* (2010a) proposed to explore the full multi-dimensional joint posterior of power spectra coefficients and density fluctuations, $\mathcal{P}(\delta^{\rm f}, P|d)$. They employ a two-steps Gibbs sampling scheme, a method previously applied to CMB data analysis (Wandelt, Larson & Lakshminarayanan, 2004; Eriksen *et al.*, 2004; Jewell, Levin & Anderson, 2004):

$$\delta^{\rm f} \curvearrowleft \mathcal{P}(\delta^{\rm f}|P, d), \tag{4.44}$$

$$P \curvearrowleft \mathcal{P}(P|\delta^{\rm f}, d), \tag{4.45}$$

where the arrow denotes a random draw from the pdf on its right. The ARES code is an implementation of this scheme. It assumes the conditional independence $\mathcal{P}(P|\delta^{\rm f}, d) = \mathcal{P}(P|\delta^{\rm f})$, which yields an inverse-Gamma distribution for power spectrum coefficients, and a Gaussian prior for $\delta^{\rm f}$ (i.e. a Wiener posterior for $\mathcal{P}(\delta^{\rm f}|P, d)$; see Jasche *et al.*, 2010a). In Jasche & Wandelt (2013b), updates and improvements of ARES are introduced, in order to account for uncertainties arising from galaxy biases and normalizations of the galaxy density (i.e. noise levels).

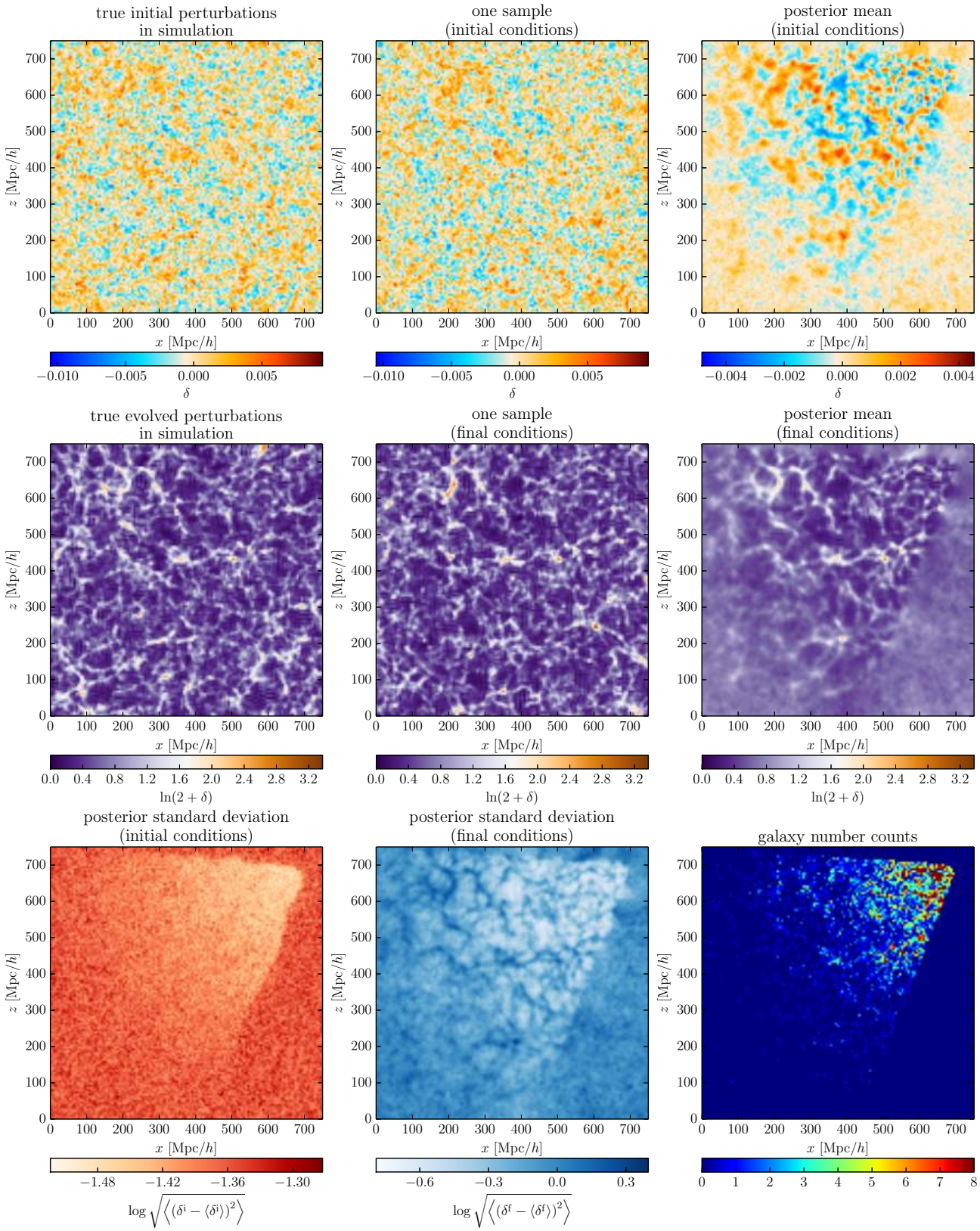

Figure 4.5: Slices through the box used for testing BORG on a synthetic data set. Various quantities (indicated above the panels) are shown. The comparison between panels illustrates the performance of BORG at inferring density fields and demonstrates its capability of quantifying uncertainties. This figure shows results originally obtained by Jasche & Wandelt (2013a), courtesy of Jens Jasche.

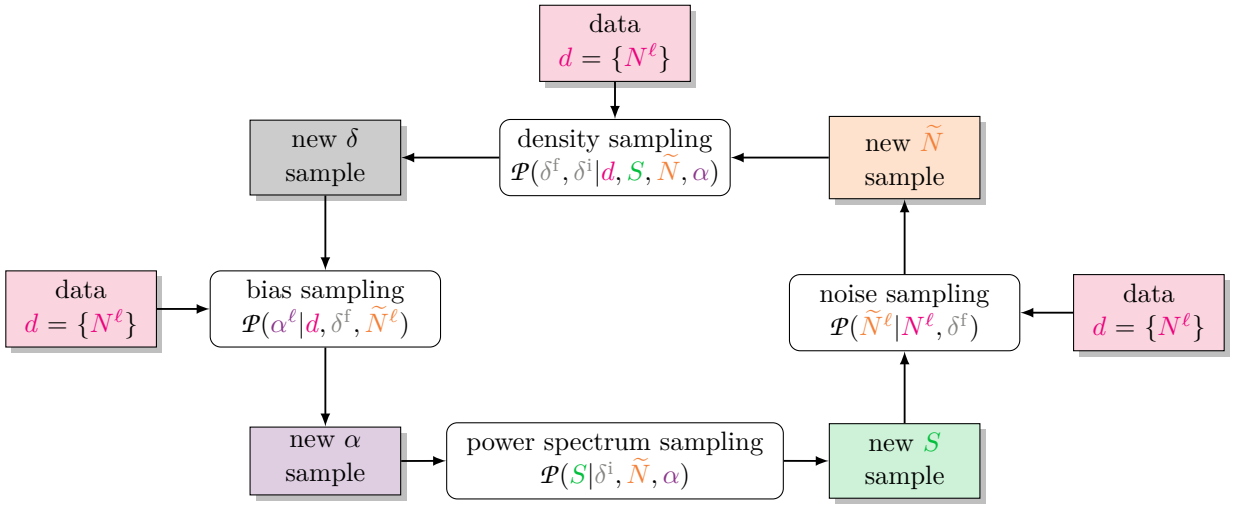

Figure 4.6: Flow chart depicting the multi-step iterative block sampling procedure for a natural extension of the BORG algorithm. In the first step, BORG generates random realizations of initial and final density fields conditional on the galaxy samples d, on the covariance matrix of initial fluctuations, S, on the noise levels $\{\widetilde{N}^\ell\}$ and on the bias parameters $\{\alpha^\ell\}$. In subsequent steps, the bias parameters, the covariance matrix and the noise parameters are sampled conditional on respective previous samples and on the data when necessary. Iterations of this procedure yield samples from the full joint posterior distribution, $\mathcal{P}(\delta^{\rm f}, \delta^{\rm i}, S, \widetilde{N}, \alpha|d)$.

Following these ideas, an extended BORG algorithm should perform iterative block sampling according to the scheme given in figure 4.6 (for reference, see also figure 4.4 for the current BORG algorithm, and figure 1 in Jasche & Wandelt, 2013b, for the ARES algorithm). In comparison to the conditional posterior expressions written down by Jasche *et al.* (2010a) and Jasche & Wandelt (2013b), this procedure would involve the expression of $\mathcal{P}(\alpha^\ell|d, \delta^{\rm f}, \widetilde{N}^\ell)$ in terms of the BORG power-law bias model (instead of the linear bias model of ARES) and of $\mathcal{P}(S|\delta^{\rm i}, \widetilde{N}, \alpha)$ in terms of initial (instead of final) density fields.[2] In ARES, density sampling is by far the most expensive step. It can be done by constructing the Wiener-filtered map (which requires inversions of large matrices, see equations (1.27) and (1.28)) and augmenting missing fluctuations from the prior (Jasche *et al.*, 2010a), by means of HMC (Jasche & Wandelt, 2013b), or by using an auxiliary messenger field, which removes the need for matrix inversion (Jasche & Lavaux, 2015; see also Elsner & Wandelt, 2013). For the BORG data model, involving a structure formation model instead of a Gaussian prior for the galaxy density, HMC is the state-of-the-art technique.

An upcoming improvement of BORG will involve the joint sampling of density $\delta^{\rm i}$, noise levels $\widetilde{N}^\ell$ and bias parameters α^ℓ. Unfortunately, computational time issues mean that joint, physical inference of density and power spectra is still out of reach. Correlation lengths are of the order of 200 samples for BORG density fields (Jasche & Wandelt, 2013a) and 100 samples for ARES power spectrum coefficients (Jasche & Wandelt, 2013b).[3] Preliminary tests indicate that the correlation length for the joint inference process is of the order of a few hundred samples. However, even with a correlation length of 100 samples, accurate characterization of power spectra and corresponding uncertainties require, at least, about 40,000 samples. With the current performance of the BORG sampler (discussed in sections 4.4.2 and 5.2), such a run would take several years on a typical computer. For this reason, this thesis focuses on sampling the matter density field for a fixed power spectrum of primordial fluctuations, rather than sampling this as well. Algorithmic and methodological innovations that would render such a run possible are currently being discussed but will require a considerable additional implementation effort and are outside the scope of this thesis.

[2] As noted in section 4.2.1.1, the Fourier-space representation of S is a diagonal matrix containing the coefficients $\sqrt{P(k)/(2\pi)^{3/2}}$.
[3] See Jasche & Wandelt, 2013b; Jewell *et al.*, 2009, for the discussion of a method designed to reduce the otherwise prohibitively long correlation length of ARES chains.

Chapter 5

Past and present cosmic structure in the Sloan Digital Sky Survey

Contents

> "Map-making had never been a precise art on the Discworld. People tended to start off with good intentions and then get so carried away with the spouting whales, monsters, waves, and other twiddly bits of cartographic furniture that they often forgot to put the boring mountains and rivers in at all."
> — Terry Pratchett (1990), *Moving Pictures*

Abstract

We present a chrono-cosmography project, aiming at the inference of the four dimensional formation history of the observed large-scale structure from its origin to the present epoch. To do so, we perform a full-scale Bayesian analysis of the northern galactic cap of the Sloan Digital Sky Survey (SDSS) Data Release 7 main galaxy sample, relying on a fully probabilistic, physical model of the non-linearly evolved density field. Besides inferring initial conditions from observations, our methodology naturally and accurately reconstructs non-linear features at the present epoch, such as walls and filaments, corresponding to high-order correlation functions generated by late-time structure formation. Our inference framework self-consistently accounts for typical observational systematic and statistical uncertainties such as noise, survey geometry and selection effects. We further account for luminosity dependent galaxy biases and automatic noise calibration within a fully Bayesian approach. As a result, this analysis provides highly-detailed and accurate reconstructions of the present density field on scales larger than $\sim$ 3 Mpc/h, constrained by SDSS observations. This approach also leads to the first quantitative inference of plausible formation histories of the dynamic large scale structure underlying the observed galaxy distribution. The results described in this chapter constitute the first full Bayesian non-linear analysis of the cosmic large scale structure with the demonstrated capability of uncertainty quantification. Some of these results have been made publicly available along with the corresponding paper. The level of detail of inferred results and the high degree of control on observational uncertainties pave the path towards high precision chrono-cosmography, the subject of simultaneously studying the dynamics and the morphology of the inhomogeneous Universe.

This chapter is adapted from its corresponding publication, Jasche, Leclercq & Wandelt (2015).

This chapter describes the BORG analysis of the Sloan Digital Sky Survey Data Release 7 main galaxy sample. It is structured as follows. In section 5.1, we give a brief overview about the SDSS data set used in the analysis. In section 5.2, we demonstrate the application of the BORG inference algorithm to observations and discuss

the general performance of the Hamiltonian Monte Carlo sampler. Section 5.3 describes the inference results obtained in the course of this work. In particular, we present results on inferred 3D initial and final density as well as velocity fields and show the ability of our method to provide accurate uncertainty quantification for any finally inferred quantity. Further, we also demonstrate the ability of our methodology to perform chrono-cosmography, by accurately inferring plausible 4D formation histories for the observed LSS from its origins to the present epoch. In section 5.4, we conclude by summarizing and discussing the results obtained in the course of this project.

5.1 The SDSS galaxy sample

In this work, we follow a similar procedure as described in Jasche *et al.* (2010b), by applying the BORG algorithm to the SDSS main galaxy sample. Specifically, we employ the `Sample dr72` of the New York University Value Added Catalogue[1] (NYU-VAGC). This is an updated version of the catalogue originally constructed by Blanton *et al.* (2005) and is based on the final data release (DR7; Abazajian *et al.*, 2009) of the Sloan Digital Sky Survey (SDSS; York *et al.*, 2000). Based on `Sample dr72`, we construct a flux-limited galaxy sample with spectroscopically measured redshifts in the range $0.001 < z < 0.4$, r-band Petrosian apparent magnitude $r \leq 17.6$ after correction for Galactic extinction, and r-band absolute magnitude $-21 < M_{0.1r} < -17$. Absolute r-band magnitudes are corrected to their $z = 0.1$ values using the K-correction code of Blanton *et al.* (2003a); Blanton & Roweis (2007) and the luminosity evolution model described in Blanton *et al.* (2003b). We also restrict our analysis to the main contiguous region of the SDSS in the northern Galactic cap, excluding the three survey strips in the southern cap (about 10 per cent of the full survey area). The NYU-VAGC provides required information on the incompleteness in our spectroscopic sample. This includes a mask, indicating which areas of the sky have been targeted and which not. The mask defines the effective area of the survey on the sky, which is 6437 deg^2 for the sample we use here. This survey area is divided into a large number of smaller subareas, called *polygons*, for each of which the NYU-VAGC lists a spectroscopic completeness, defined as the fraction of photometrically identified target galaxies in the polygon for which usable spectra were obtained. Throughout our sample the average completeness is 0.92. To account for radial selection functions, defined as the fraction of galaxies in the absolute magnitude range considered here, that are within the apparent magnitude range of the sample at a given redshift, we use a standard luminosity function proposed by Schechter (1976) with r-band parameters $\alpha = -1.05$, $M_* - 5\log_{10}(h) = -20.44$ (Blanton *et al.*, 2003c).

Our analysis accounts for luminosity dependent galaxy biases, by following the approach described in section 4.2. In order to do so, we subdivide our galaxy sample into six equidistant bins in absolute r-band magnitude in the range $-21 < M_{0.1r} < -17$, resulting in a total of 372, 198 main sample galaxies to be used in the analysis. As described in section 4.2, splitting the galaxy sample permits us to treat each of these sub-samples as an individual data set, with its respective selection effects, biases and noise levels.

5.2 The BORG SDSS analysis

We performed the analysis of the SDSS main galaxy sample on a cubic Cartesian domain with a side length of 750 Mpc/h consisting of 256^3 equidistant grid nodes, resulting in $\sim 1.6\times 10^7$ inference parameters. Thus, the inference procedure provides data-constrained realizations for initial and final density fields at a grid resolution of about ~ 3 Mpc/h. For the analysis, we assume a standard ΛCDM cosmology with the set of cosmological parameters

$$\Omega_\Lambda = 0.728, \Omega_\mathrm{m} = 0.272, \Omega_\mathrm{b} = 0.045, \sigma_8 = 0.807, h = 0.702, n_\mathrm{s} = 0.961. \tag{5.1}$$

The cosmological power spectrum for initial density fields is calculated according to the prescription provided by Eisenstein & Hu (1998, 1999). In order to sufficiently resolve the final density field, the 2LPT model is evaluated with 512^3 particles, by oversampling initial conditions by a factor of eight.

We adjusted the parameters α^ℓ of the assumed power-law bias model during the initial 1000 sampling steps, but kept them fixed afterwards. For the purpose of this work, the power-law indices α^ℓ of the bias relations are determined by requiring them to resemble the linear luminosity dependent bias when expanded in a Taylor series to linear order as:

$$(1+\delta^\mathrm{f})^{\alpha^\ell} = 1 + \alpha^\ell \delta^\mathrm{f} + \mathcal{O}\left(\left(\delta^\mathrm{f}\right)^2\right). \tag{5.2}$$

[1] http://sdss.physics.nyu.edu/vagc/

$M^{\ell}_{0.1_r}$	α^{ℓ}	$\widetilde{N}^{\ell}$
$-21.00 < M^{0}_{0.1_r} < -20.33$	1.58029	$4.67438 \times 10^{-2} \pm 3.51298 \times 10^{-4}$
$-20.33 < M^{1}_{0.1_r} < -19.67$	1.41519	$9.54428 \times 10^{-2} \pm 5.77786 \times 10^{-4}$
$-19.67 < M^{2}_{0.1_r} < -19.00$	1.30822	$1.39989 \times 10^{-1} \pm 1.21087 \times 10^{-3}$
$-19.00 < M^{3}_{0.1_r} < -18.33$	1.23272	$1.74284 \times 10^{-1} \pm 1.89168 \times 10^{-3}$
$-18.33 < M^{4}_{0.1_r} < -17.67$	1.17424	$2.19634 \times 10^{-1} \pm 3.42586 \times 10^{-3}$
$-17.67 < M^{5}_{0.1_r} < -17.00$	1.12497	$2.86236 \times 10^{-1} \pm 5.57014 \times 10^{-3}$

Table 5.1: Bias and noise parameters, as described in the text, for six galaxy sub-samples, subdivided by their absolute r-band magnitudes.

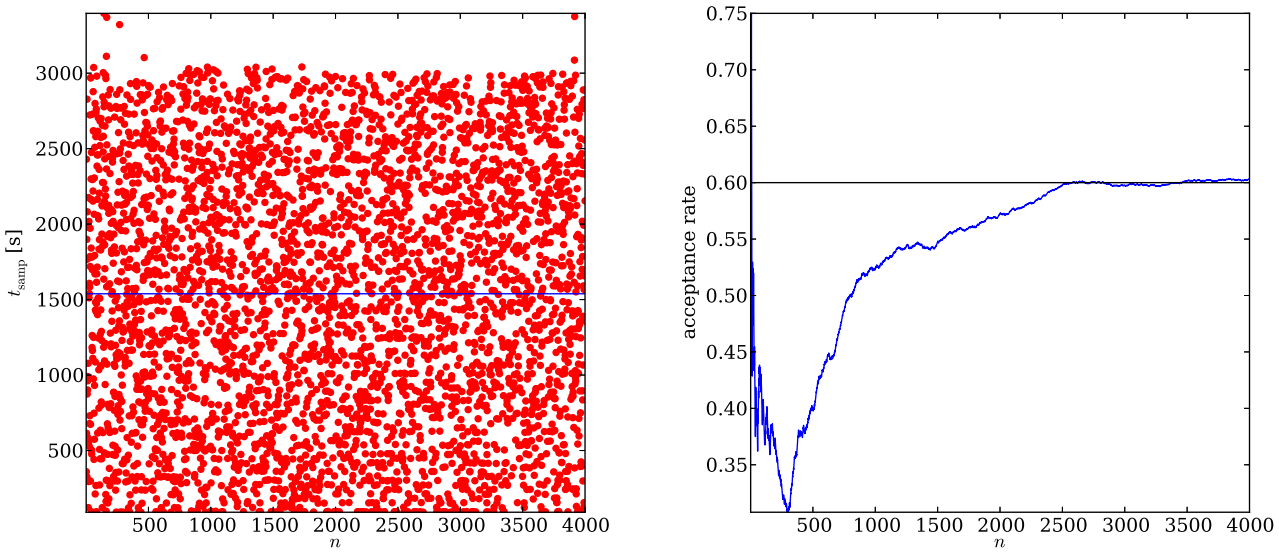

Figure 5.1: Diagnotics of the Markov chain: scatter plot of sample generation times (left panel) and Markov acceptance rates during the initial burn-in phase (right panel). As shown by the left panel, times to generate individual samples range from zero to about 3000 seconds. The average execution time per sample generation is about 1500 seconds on 16 cores. Initially, acceptance rates drop during burn-in but rise again to reach an asymptotic value of about 60 percent.

In particular, we assume the functional shape of the luminosity dependent bias parameter α^{ℓ} to follow a standard model for the linear luminosity dependent bias in terms of absolute r-band magnitudes $M_{0.1_r}$, as given by:

$$\alpha^{\ell} = b(M^{\ell}_{0.1_r}) = b_* \left(a + b \times 10^{0.4\left(M_* - M^{\ell}_{0.1_r}\right)} + c \times \left(M^{\ell}_{0.1_r} - M_*\right) \right) , \tag{5.3}$$

with the fitting parameters $a = 0.895$, $b = 0.150$, $c = -0.040$ and $M_* = -20.40$ (see e.g. Norberg *et al.*, 2001; Tegmark *et al.*, 2004, for details). The parameter b_* was adjusted during the initial burn-in phase and was finally set to a fixed value of $b_* = 1.44$, such that the sampler recovers the correct shape of the assumed initial power spectrum.

As described in sections 4.2.4 and 4.3.1, contrary to bias exponents, corresponding noise parameters $\widetilde{N}^{\ell}$ are sampled and explored throughout the entire Markov chain. Inferred ensemble means and standard deviations for the $\widetilde{N}^{\ell}$ along with chosen power-law parameters α^{ℓ} are provided in table 5.1.

The entire analysis yielded 12,000 realizations for initial and final density fields. The generation of a single Markov sample requires an operation count equivalent to about $\sim$ 200 2LPT model evaluations. Typical generation times for data-constrained realizations are shown in the left panel of figure 5.1. On average the sampler requires about 1500 seconds to generate a single density field realization on 16 cores. The total analysis consumed several months of computing time and produced on the order of $\sim$ 3 TB of information represented by the set of Markov samples.

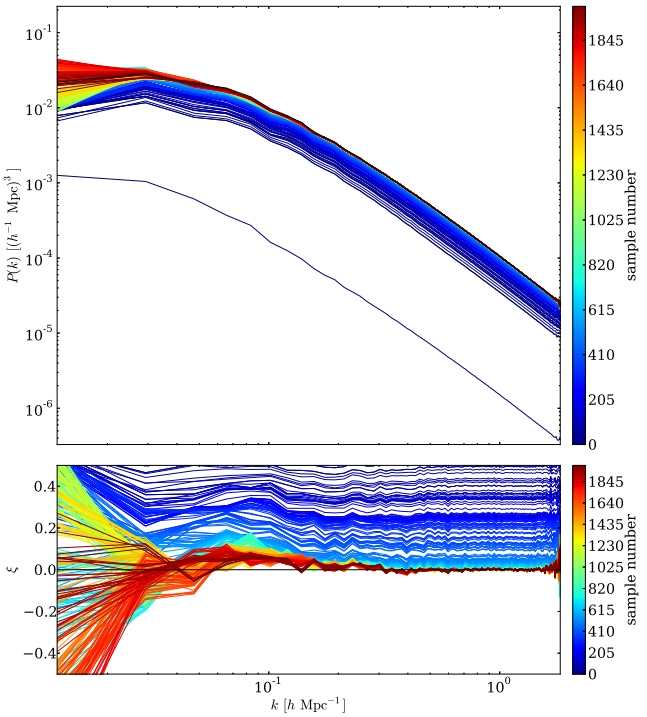

Figure 5.2: Burn-in power spectra measured from the first 2000 samples of the Markov chain colored corresponding to their sample number as indicated by the colorbar. The black line represents a fiducial reference power spectrum for the cosmology assumed in this work. Subsequent power spectra approach the fiducial cosmological power spectrum homogeneously throughout all scales in Fourier space.

The numerical efficiency of any Markov Chain Monte Carlo algorithm, particularly in high dimensions, is crucially determined by the average acceptance rate. As demonstrated by the right panel of figure 5.1, after an initial burn-in period, the acceptance rate asymptotes at a value of about 60 percent, rendering our analysis numerically feasible. As a simple consistency check, we follow a standard procedure to determine the initial burn-in behavior of the sampler via a simple experiment (see e.g. Eriksen *et al.*, 2004; Jasche & Kitaura, 2010; Jasche & Wandelt, 2013a, for more details). The sampler is initialized with an overdispersed state, far remote from the target region in parameter space, by scaling normal random amplitudes of the initial density field at a cosmic scale factor of $a = 10^{-3}$ by a constant factor of 0.01. In the course of the initial burn-in phase, the Markov chain should then drift towards preferred regions in parameter space. As demonstrated by figure 5.2, this drift is manifested by a sequence of posterior power spectra measured from subsequent initial density field realizations. It can be clearly seen that the chain approaches the target region within the first 2000 sampling steps. The sequence of power spectra shows a homogeneous drift of all modes with no indication of any particular hysteresis or bias across different scales in Fourier space. As improper treatment of survey systematics, uncertainties and galaxy bias typically result in obvious erroneous features in power spectra, figure 5.2 clearly demonstrates that these effects have been accurately accounted for by the algorithm.

5.3 Inference results

This section describes inference results obtained by our Bayesian analysis of the SDSS main galaxy sample.

5.3.1 Inferred 3D density fields

A major goal of this work is to provide inferred 3D initial and final density fields along with corresponding uncertainty quantification in a $\sim 1.6 \times 10^7$ dimensional parameter space. To do this, the BORG algorithm provides a sampled LSS posterior distribution in terms of an ensemble of data-constrained samples, via an efficient implementation of a Markov Chain Monte Carlo algorithm. It should be remarked that, past the initial

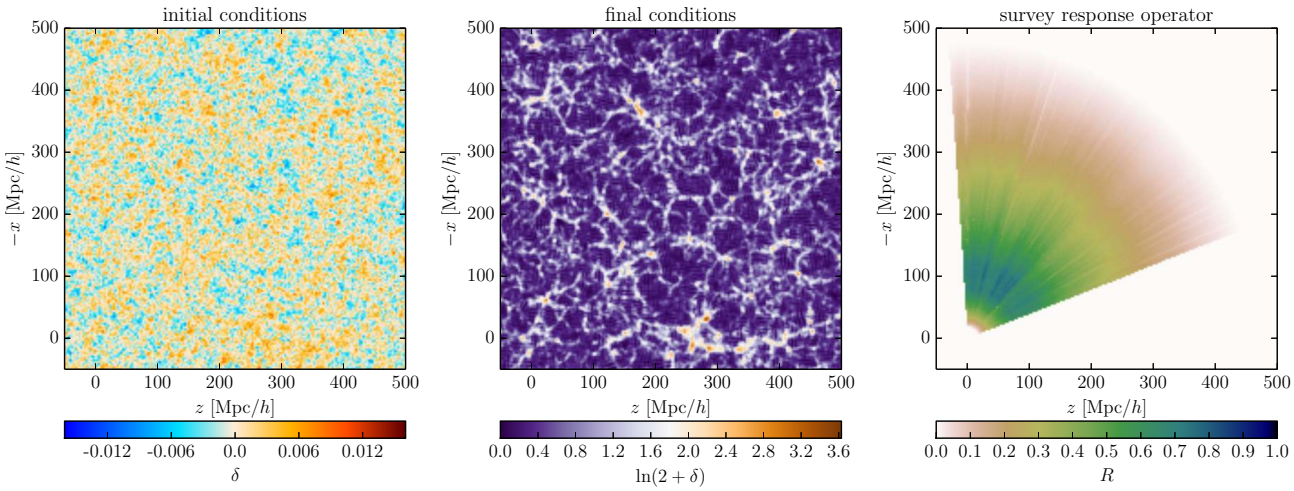

Figure 5.3: Slices through the initial (left panel) and corresponding final (middle panel) density fields of the 5000th sample. The right panel shows a corresponding slice through the combined survey response operator R for the six absolute magnitude bins considered in this work. As can be seen, unobserved and observed regions in the inferred initial and final density fields do not appear visually distinct, demonstrating the fact that individual data-constrained realizations constitute physically meaningful density fields. It also shows that the sampler naturally extends observed large scale structures beyond the survey boundaries in a physically and statistically fully consistent fashion.

burn-in phase, all individual samples reflect physically meaningful density fields, limited only by the validity of the employed 2LPT model. In particular, the present analysis correctly accounts for selection effects, survey geometries, luminosity dependent galaxy biases and automatically calibrates the noise levels of the six luminosity bins as described above. As can be seen in figure 5.2, past the initial burn-in phase, individual samples possess physically correct power throughout all ranges in Fourier space, and do not show any sign of attenuation due to survey characteristics such as survey geometry, selection effects or galaxy biases.

To further illustrate that individual samples qualify for physically meaningful density fields, in figure 5.3 we show slices through data-constrained realizations of the initial and final density fields of the 5000th sample as well as the corresponding slice through the combined survey response operator R, averaged over the six luminosity bins. It can be seen that the algorithm correctly augments unobserved regions with statistically correct information. Note that unobserved and observed regions in the inferred final density fields do not appear visually distinct, a consequence of the excellent approximation of 2LPT not just to the first but also higher-order moments (Moutarde *et al.*, 1991; Buchert, Melott & Weiß, 1994; Bouchet *et al.*, 1995; Scoccimarro, 2000; Scoccimarro & Sheth, 2002). Figure 5.3 therefore clearly reflects the fact that our sampler naturally extends observed large scale structures beyond the survey boundaries in a physically and statistically fully consistent fashion. This is a great advantage over previous methods relying on Gaussian or log-normal models specifying the statistics of the density field correctly only to two-point statistics by assuming a cosmological power spectrum. The interested reader may want to qualitatively compare with figure 2 in Jasche *et al.* (2010b), where a log-normal model, unable to represent filamentary structures, was employed.

The ensemble of the 12,000 inferred data-constrained initial and final density fields permits us to provide any desired statistical summary, such as mean and variance, for full 3D fields. In figure 5.4, we show slices through the ensemble mean initial and final density fields, to be used in subsequent analyses. The plot shows the correct anticipated behavior for inferred posterior mean final density fields, since observed regions represent data constraints, while unobserved regions approach cosmic mean density. This behavior is also present in corresponding initial density fields. In particular, the ensemble mean final density field shows a highly detailed LSS in regions where data constraints are available, and approaches cosmic mean density in regions where data are uninformative on average (see also Jasche *et al.*, 2010b, for comparison). Analogously, these results translate to the ensemble mean initial density field. Comparing the ensemble mean final density field to the galaxy number densities, depicted in the lower panels of figure 5.4, demonstrates the performance of the method in regions only poorly sampled by galaxies. In particular, comparing the right middle and right lower panel of figure 5.4 reveals the capability of our algorithm to recover highly detailed structures even in noise dominated regions (for a discussion see chapter 4 and Jasche & Wandelt, 2013a). By comparing ensemble mean initial and

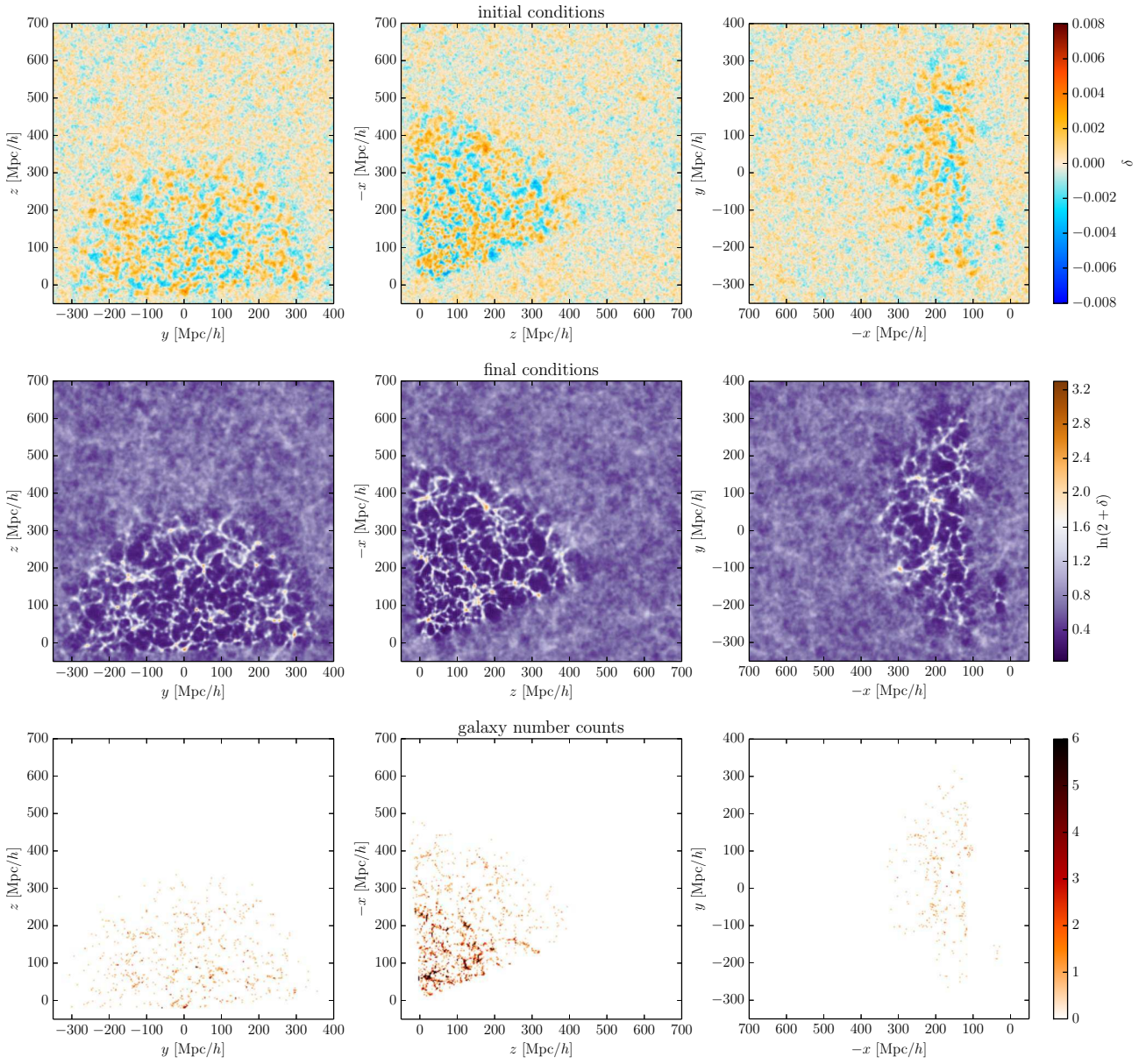

Figure 5.4: Three slices from different directions through the three dimensional ensemble posterior means for the initial (upper panels) and final density fields (middle panels) estimated from 12,000 samples. The lower panels depict corresponding slices through the galaxy number counts field of the SDSS main sample.

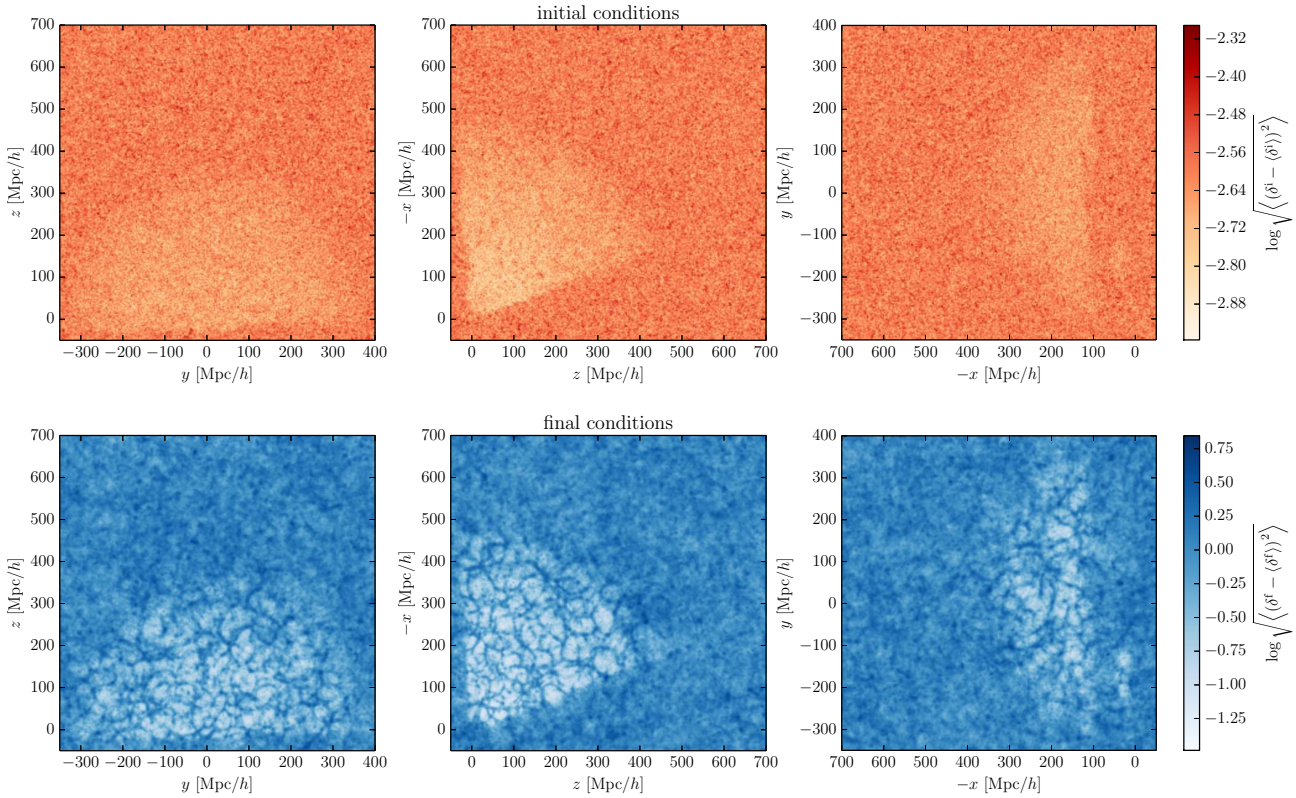

Figure 5.5: Three slices from different directions through the three dimensional voxel-wise posterior standard deviation for the initial (upper panels) and final density fields (lower panels) estimated from 12,000 samples. It can be seen that regions covered by observations show on average lower variance than unobserved regions. Also note, that voxel-wise standard deviations for the final density fields are highly structured, reflecting the signal-dependence of the inhomogeneous shot noise of the galaxy distribution. In contrast, voxel-wise standard deviations in the initial conditions are more homogeneously distributed, manifesting the flow of information between data and initial conditions as discussed in the text.

final density fields, upper and middle panels in figure 5.4, one can also see correspondences between structures in the present Universe and their origins at a scale factor of $a = 10^{-3}$.

The ensemble of data-constrained realizations also permits to provide corresponding uncertainty quantification. In figure 5.5 we plot voxel-wise standard deviations for initial and final density fields estimated from 12,000 samples. It can be seen that regions covered by data exhibit on average lower variances than unobserved regions, as expected. Note that for non-linear inference problems, signal and noise are typically correlated. This is particularly true for inhomogeneous point processes, such as discrete galaxy distributions tracing an underlying density field. In figure 5.5, the correlation between signal and noise is clearly visible for standard deviation estimates of final density fields. In particular high density regions also correspond to high variance regions, as is expected for Poissonian likelihoods since signal-to-noise ratios scale as the square root of the number of observed galaxies (also see Jasche *et al.*, 2010b, for a similar discussion). Also note that voxel-wise standard deviations for final density fields are highly structured, while standard deviations of initial conditions appear to be more homogeneous. This is related to the fact that our algorithm naturally and correctly translates information of the observations non-locally to the initial conditions via Lagrangian transport, as discussed below in section 5.3.3.

As mentioned in the introduction, results for the ensemble mean final density field and corresponding voxel-wise standard deviations have been published as as supplementary material to the article (Jasche, Leclercq & Wandelt, 2015).[2]

[2] These data can be accessed at http://iopscience.iop.org/1475-7516/2015/01/036.

5.3.2 Inference of 3D velocity fields

In addition to initial and final density fields, the analysis further provides information on the dynamics of the large scale structure as mediated by the employed 2LPT model. Indeed, the BORG algorithm shows excellent performance in recovering large scale modes, typically poorly constrained by masked galaxy observations (Jasche & Wandelt, 2013a).

This a crucial feature when deriving 3D velocity fields, which are predominantly governed by the largest scales. In this fashion, we can derive 3D velocity fields from our inference results. Note that these velocity fields are derived *a posteriori* and are only predictions of the 2LPT model given inferred initial density fields, since currently the algorithm does not exploit velocity information contained in the data. However, since inferred 2LPT displacement vectors are constrained by observations, and since 2LPT displacement vectors and velocities differ only by constant prefactors given a fixed cosmology, inferred velocities are considered to be accurate. For this reason, exploitation of velocity information contained in the data itself, being the subject of a future publication, is not expected to crucially change present results. To demonstrate the capability of recovering 3D velocity fields, in figure 5.6 we show the three components of the velocity field for the 5000th sample in spherical coordinates. More precisely, figure 5.6 shows the corresponding 2LPT particle distribution evolved to redshift $z = 0$ in a 4 Mpc/h slice around the celestial equator. Particles are colored by their radial (upper panel), polar (middle panel) and azimuthal (lower panel) velocity components. To translate between Cartesian and spherical coordinates we used the standard coordinate transform,

$$x = d_{\rm com}\cos(\lambda)\cos(\eta) \tag{5.4}$$
$$y = d_{\rm com}\cos(\lambda)\sin(\eta) \tag{5.5}$$
$$z = d_{\rm com}\sin(\lambda), \tag{5.6}$$

where λ is the declination, η is the right ascension and $d_{\rm com}$ is the radial comoving distance.

5.3.3 Inference of LSS formation histories

As described in chapter 4, the BORG algorithm employs a 2LPT model to connect initial conditions to present SDSS observations in a fully probabilistic approach. Besides inferred 3D initial and final density fields, our algorithm therefore also provides full four dimensional formation histories for the observed LSS as mediated by the 2LPT model. As an example, in figure 5.7 we depict the LSS formation history for the 5000th Markov sample ranging from a scale factor of $a = 0.02$ to the present epoch at $a = 1.00$. Initially, the density field seems to obey close to Gaussian statistics and corresponding amplitudes are low. In the course of cosmic history, amplitudes grow and higher-order statistics such as three-point statistics are generated, as indicated by the appearance of filamentary structures. The final panel of figure 5.7, at a cosmic scale factor of $a = 1.00$, shows the inferred final density field overplotted by SDSS galaxies for the six bins in absolute magnitude, as described previously. Observed galaxies nicely trace the underlying density field. This clearly demonstrates that our algorithm infers plausible formation histories for large scale structures observed by the SDSS survey. By exploring the corresponding LSS posterior distribution, the BORG algorithm naturally generates an ensemble of such data-constrained LSS formation histories, permitting to accurately quantify the 4D dynamical state of our Universe and corresponding observational uncertainties inherent to galaxy surveys. Detailed and quantitative analysis of these cosmic formation histories will be the subject of forthcoming publications (see also chapter 9).

The BORG algorithm also provides a statistically valid framework for propagating observational systematics and uncertainties from observations to any finally inferred result. This is of particular importance, since detailed treatment of survey geometries and selection effects is a crucial issue if inferred results are to be used for thorough scientific analyses. These effects generally vary greatly across the observed domain and will result in erroneous artifacts if not accounted for properly. Since large scale structure formation is a non-local process, exact information propagation is complex, as it requires to translate uncertainties and systematics from observations to the inferred initial conditions. Consequently, the information content of observed data has to be distributed differently in initial and final density fields, even though the total amount of information is conserved. Following 2LPT particles from high density regions, and corresponding high signal-to-noise regions in the data, backward in time, demonstrates that the same amount of information contained in the data will be distributed over a larger region in the initial conditions. Analogously, for underdense regions, such as voids, the information content of the data will amass in a smaller volume at the initial state. This means that the signal-to-noise ratio for a given comoving Eulerian volume is a function of time along inferred cosmic histories (Jasche & Wandelt,

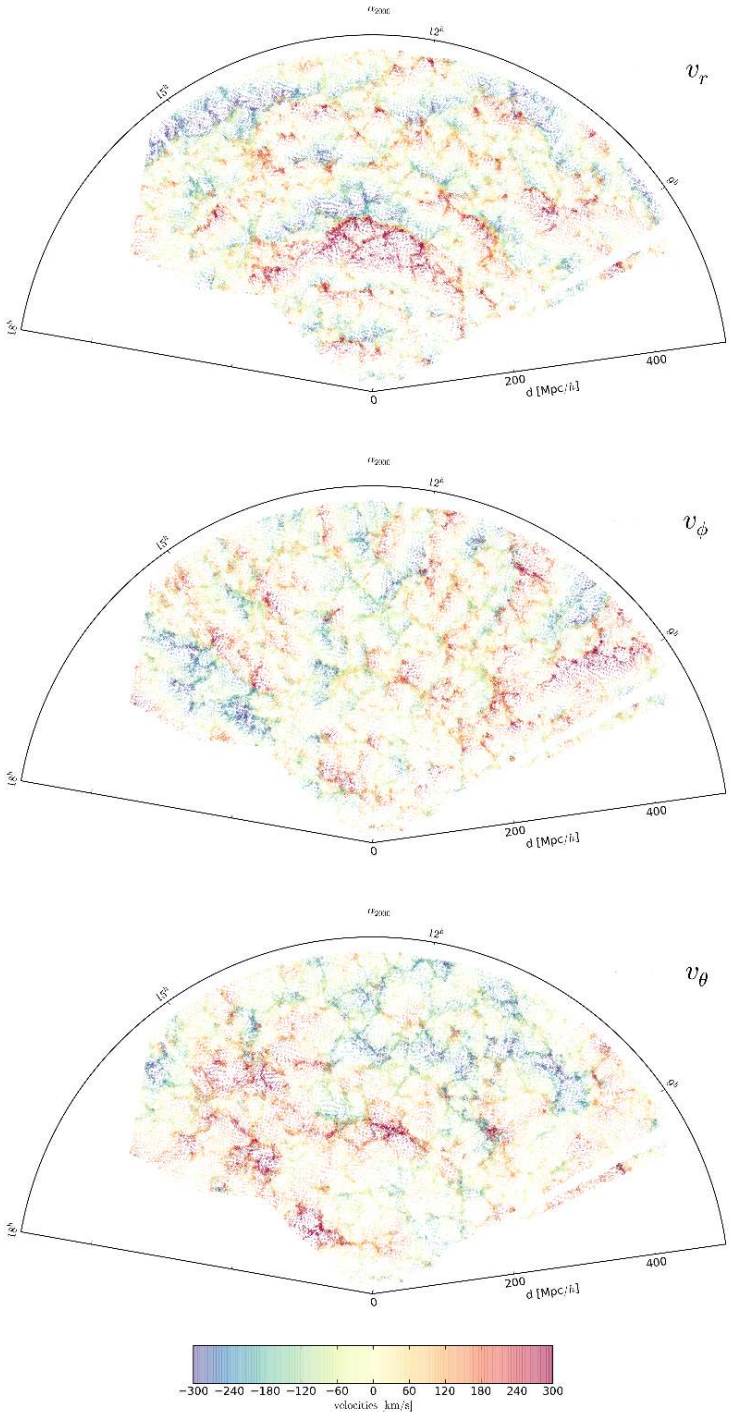

Figure 5.6: Slices through the 3D velocity fields, derived from the 5000th sample, for the radial (upper panel), polar (middle panel) and the azimuthal (lower panel) velocity components. The plot shows 2LPT particles in a 4 Mpc/h thick slice around the celestial equator for the observed domain, colored by their respective velocity components.

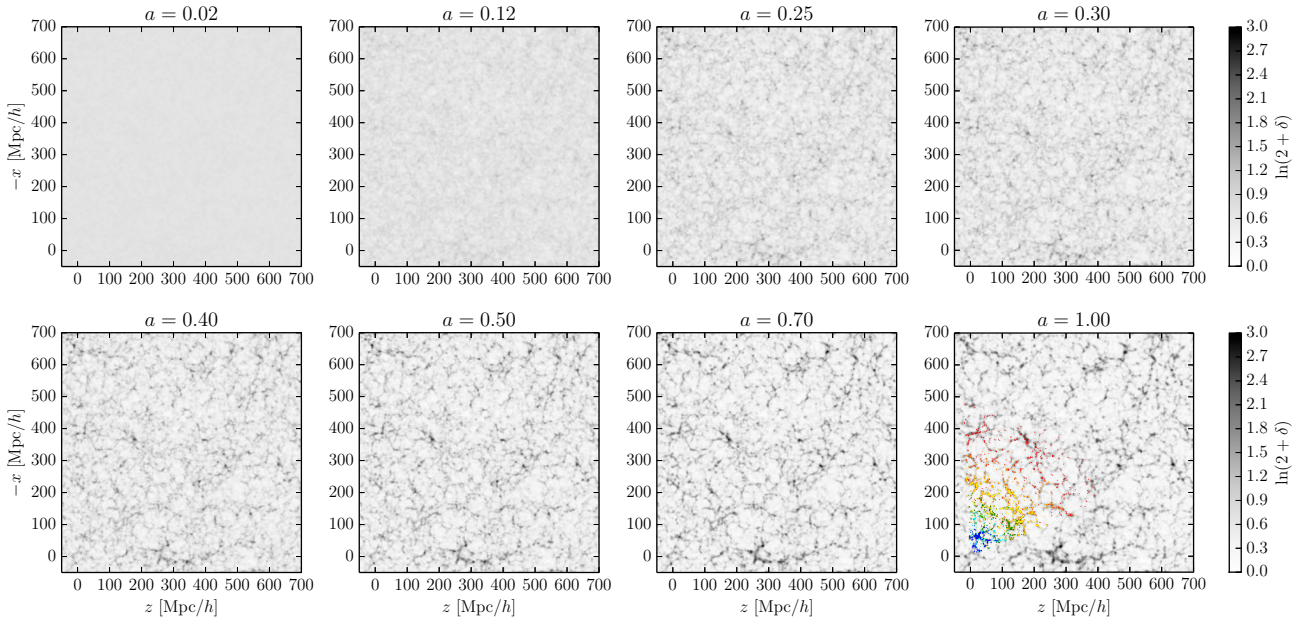

Figure 5.7: Slices through the inferred three dimensional density field of the 5000th sample at different stages of its evolution, as indicated by the cosmic scale factor in the respective panels. The plot describes a possible formation scenario for the LSS in the observed domain starting at a scale factor of $a = 0.02$ to the present epoch $a = 1.0$. In the lower right panel, we overplotted the inferred present density field with the observed galaxies in the respective six absolute magnitude ranges $-21.00 < M_{0.1_r} < -20.33$ (red dots), $-20.33 < M_{0.1_r} < -19.67$ (orange dots), $-19.67 < M_{0.1_r} < -19.00$ (yellow dots), $-19.00 < M_{0.1_r} < -18.33$ (green dots) , $-18.33 < M_{0.1_r} < -17.67$ (cyan dots) and , $-17.67 < M_{0.1_r} < -17.00$ (blue dots). As can be clearly seen, observed galaxies trace the recovered three dimensional density field. Besides measurements of three dimensional initial and final density fields, this plot demonstrates that our algorithm also provides plausible four dimensional formation histories, describing the evolution of the presently observed LSS.

2013a). This fact manifests itself in the different behaviour of voxel-wise standard deviations for final and initial conditions, as presented in figure 5.5. While the signal-to-noise ratio is highly clustered in final conditions, the same amount of observational information is distributed more evenly over the entire volume in corresponding initial conditions.

Non-local propagation of observational information across survey boundaries, together with cosmological correlations in the initial density field, is also the reason why our method is able to extrapolate the cosmic LSS beyond survey boundaries, as discussed in section 5.3.1 above and demonstrated by figure 5.4. To further demonstrate this fact, in figure 5.8, we show the density field of the 5000th sample traced by particles from inside and outside the observed domain at the present epoch. At the present epoch, the set of particles can be sub-divided into two sets for particles inside and outside the observed domain. The boundary between these two sets of particles is the sharp outline of the SDSS survey geometry. When tracing these particles back to an earlier epoch at a scale factor of $a = 0.02$, it can clearly be seen that this sharp boundary starts to frazzle. Particles within the observed domain at the final state may originate from regions outside the corresponding Eulerian volume at the initial state, and vice versa. Information from within the observed domain non-locally influences the large scale structure outside the observed domain, thus increasing the region influenced by data beyond the survey boundaries. Figure 5.8 therefore demonstrates the ability of our algorithm to correctly account for information propagation via Lagrangian transport within a fully probabilistic approach. The ability to provide 4D dynamic formation histories for SDSS data together with accurate uncertainty quantification paves the path towards high precision chrono-cosmography, permitting us to study the inhomogeneous evolution of our Universe. Detailed and quantitative analysis of the various aspects of the results obtained in this chapter are discussed in part IV of this thesis and will be the subject of future publications.

5.4 Summary and conclusions

This chapter discusses a fully Bayesian chrono-cosmographic analysis of the 3D cosmological large scale structure underlying the SDSS main galaxy sample (Abazajian *et al.*, 2009). We presented a data application of the recently proposed BORG algorithm (see chapter 4 and Jasche & Wandelt, 2013a), which permits to simultaneously infer initial and present non-linear 3D density fields from galaxy observations within a fully probabilistic approach. As discussed in chapter 4, the algorithm incorporates a second-order Lagrangian perturbation model to connect observations to initial conditions and to perform dynamical large-scale structure inference from galaxy redshift surveys.

Besides correctly accounting for usual statistical and systematic uncertainties, such as noise, survey geometries and selection effects, this methodology also physically treats gravitational structure formation in the linear and mildly non-linear regime and captures higher-order statistics present in non-linear density fields (see e.g. Moutarde *et al.*, 1991; Buchert, Melott & Weiß, 1994; Bouchet *et al.*, 1995; Scoccimarro, 2000; Scoccimarro & Sheth, 2002). The BORG algorithm explores a high-dimensional posterior distribution via an efficient implementation of a Hamiltonian Monte Carlo sampler and therefore provides naturally and fully self-consistently accurate uncertainty quantification for any finally inferred quantity.

In the paper corresponding to this work (Jasche, Leclercq & Wandelt, 2015), we upgraded the original sampling procedure described in Jasche & Wandelt (2013a) to account for automatic noise calibration and luminosity dependent galaxy biases (see sections 4.2.4 and 4.3.1). To do so, we followed the philosophy described in Jasche & Wandelt (2013b) and splitted the main galaxy sample into six absolute magnitude bins in the range $-21 < M_{0.1r} < -17$. The Bayesian analysis treats each of this six galaxy sub-samples as an individual data set with its individual statistical and systematic uncertainties. As described in sections 4.2.4 and 4.3.1, the original algorithm described in Jasche & Wandelt (2013a) has been augmented by a power-law bias model and an additional sampling procedure to jointly infer corresponding noise levels for the respective galaxy samples.

As discussed in section 5.2, we applied this modified version of the BORG algorithm to the SDSS DR7 main galaxy samples and generated about 12,000 full three dimensional data-constrained initial conditions in the course of this work. The initial density field, at a scale factor of $a = 10^{-3}$, has been inferred on a comoving Cartesian equidistant grid, of side length 750 Mpc/h and 256^3 grid nodes. This amounts to a target resolution of about $\sim$ 3 Mpc/h for respective volume elements. Density amplitudes at these Lagrangian grid nodes correspond to about $\sim 10^7$ parameters to be constrained by our inference procedure. Typically, the generation of individual data-constrained realizations involves an equivalent of $\sim$ 200 2LPT evaluations and requires on the order of 1500 seconds on 16 cores. Despite the complexity of the problem, we demonstrated that our sampler

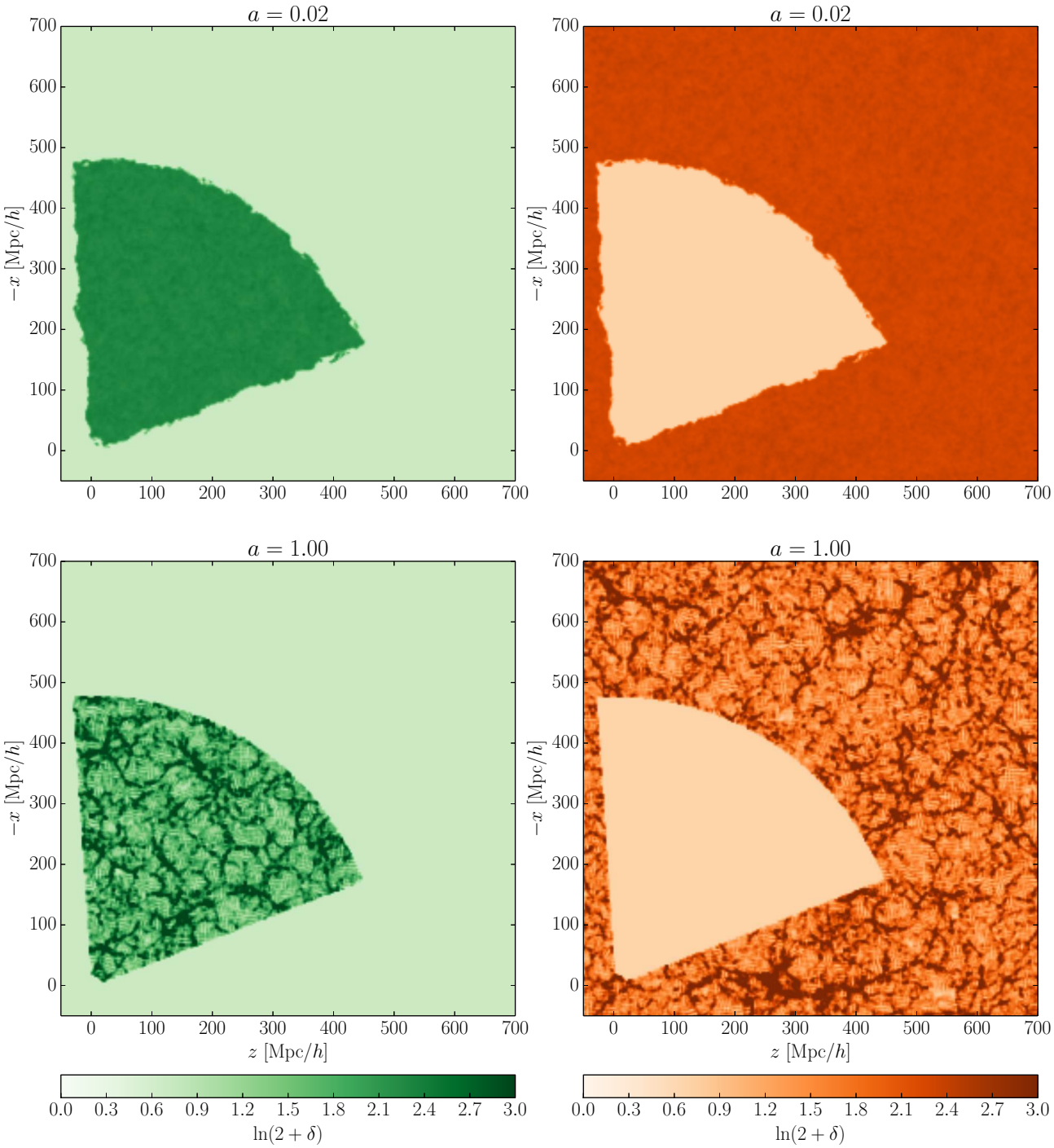

Figure 5.8: Slices through the distribution of particles in the 5000th sample, which are located inside (left panels) and outside (right panels) the observed domain at the time of observation, at two time snapshots as indicated above the panels. It can be seen that particles located within the observed region at the present time may originate from regions outside the corresponding comoving Eulerian volume at an earlier epoch and vice versa. As discussed in the text, this plot demonstrates the non-local transport of information, which provides accurate inference of the cosmic large scale structure beyond survey boundaries within a rigorous probabilistic approach.

can explore multi-million dimensional parameter spaces via efficient Markov Chain Monte Carlo algorithms with an asymptotic acceptance rate of about 60 percent, rendering our numerical inference framework numerically feasible.

To test the performance of the sampler, we followed a standard approach for testing the initial burn-in behavior via experiments (see e.g. Eriksen *et al.*, 2004; Jasche & Kitaura, 2010; Jasche & Wandelt, 2013a). We initialized the sampler with a Gaussian random field scaled by a factor of 0.01, to start from an over-dispersed state. During an initial burn-in period the sampler performed a systematic drift towards the target region in parameter space. We examined the initial burn-in behavior by following the sequence of *a posteriori* power spectra, measured from the first 2500 samples, and showed that subsequent samples homogeneously approach the target spectrum throughout all regions in Fourier space without any sign of hysteresis. This indicates the efficiency of the sampler to rapidly explore all scales of the inference problem. The absence of any particular bias or erroneous power throughout all scales in Fourier space, further demonstrates the fact that survey geometry, selection effects, galaxy biasing and observational noise have been accurately accounted for in this analysis. These *a posteriori* power spectra also indicate that individual data-constrained realizations possess the correct physical power in all regions in Fourier space, and can therefore be considered as physically meaningful density fields. This fact has been further demonstrated in section 5.3.1 by showing slices through an arbitrary data-constrained realization. These results clearly demonstrate the power of our Bayesian methodology to correctly treat the ill-posed inverse problem of inferring signals from incomplete observations, by augmenting unobserved regions with statistically and physically meaningful information. In particular, constrained and unconstrained regions in the samples are visually indistinguishable, demonstrating a major improvement over previous approaches, typically relying on Gaussian or log-normal statistics, incapable of representing the filamentary structure of the cosmic web (see e.g. Jasche & Kitaura, 2010). It should be remarked that this fact not only demonstrates the ability to access high-order statistics in finally inferred quantities such as 3D density maps, but also reflects the control of high-order statistics in uncertainty quantification far beyond standard normal statistics.

The ensemble of $12,000$ full 3D data-constrained samples permits us to estimate any desired statistical summary. In particular, in section 5.3.1, we showed ensemble mean density fields for final and initial conditions. A particularly interesting aspect is the fact that the algorithm manages to infer highly-detailed large scale structures even in regimes only poorly covered by observations (for further comments see chapter 4 and Jasche & Wandelt, 2013a). To demonstrate the possibility of uncertainty quantification, we also calculated the ensemble voxel-wise posterior standard deviation, which reflects the degree of statistical uncertainty at every volume element in the inference domain. As discussed in section 5.3.1, these results clearly reflect the signal-dependence of noise for any inhomogeneous point processes, such as discrete Poissonian galaxy distribution. As expected, high signal regions correspond to high variance regions. These results further demonstrate the ability to accurately translate uncertainties in the final conditions to initial density fields, as demonstrated by the plots of voxel-wise standard deviations for corresponding initial density fields. However, note that voxel-wise standard deviations are just an approximation to the full joint and correlated uncertainty that otherwise can by correctly quantified by considering the entire set of data-constrained realizations. Besides 3D initial and final density fields, the methodology also provides information on cosmic dynamics, as mediated by the 2LPT model. In section 5.3.2, we showed a velocity field realization in one sample. In particular, we showed the radial, polar and azimuthal velocity components in a 4 Mpc/h thick slice around the celestial equator for the observed domain. These velocities are not primarily constrained by observations, but are derived from the 2LPT model. However, since 2LPT displacement vectors are data-constrained, and since displacement vectors and velocities differ only by constant factors independent of the inference process, derived velocities are considered to be accurate.

As pointed out frequently, the BORG algorithm employs 2LPT as a dynamical model to connect initial conditions to present observations of SDSS galaxies. As a consequence, the algorithm not only provides 3D density and velocity fields but also infers plausible 4D formation histories for the observed LSS. In section 5.3.3, we illustrated this feature with an individual sample. We followed its cosmic evolution from a initial scale factor of $a = 0.02$ to the present epoch at $a = 1.00$. As could be seen, the initial density field appears homogeneous and obeys Gaussian statistics. In the course of structure formation clusters, filaments and voids are formed. To demonstrate that this formation history correctly recovers the observed large scale structure, we plotted the observed galaxies, for the six luminosity bins, on top of the final density field. These results clearly demonstrate the ability of our algorithm to infer plausible large scale structure formation histories compatible with observations. Additionally, since the BORG algorithm is a full Bayesian inference framework, it not only provides a single 4D history, but an ensemble of such data-constrained formation histories and thus accurate means to quantify corresponding observational uncertainties. In particular, our methodology correctly accounts

for the non-local transport of observational information between present observations and corresponding inferred initial conditions. As discussed in section 5.3.3, the information content in initial and final conditions has to be conserved but can be distributed differently. High-density regions in the final conditions, typically coinciding with high signal-to-noise regions in the data, form by clustering of matter which was originally distributed over a larger Eulerian volume in the initial conditions. For this reason, the observational information associated to a cluster in the final density field will be distributed over a larger volume in the corresponding initial density field. Conversely, the information content of voids in the final conditions will be confined to a smaller volume in the initial conditions. This fact is also reflected by the analysis of voxel-wise standard deviations presented in section 5.3.1. While the signal-to-noise ratio is highly clustered in the final conditions, the same amount of observational information is distributed more homogeneously over the entire volume in corresponding initial conditions. As discussed in section 5.3.3, particles within the observed domain at the final state may originate from regions outside the corresponding comoving Eulerian volume in the initial conditions and vice versa (also see chapter 4 and Jasche & Wandelt, 2013a). This non-local translation of information along Lagrangian trajectories is also the reason for the ability of our methodology to extrapolate beyond the survey boundaries of the SDSS and infer the LSS there within a fully probabilistic and rigorous approach. In particular, the high degree of control on statistical uncertainties permit us to perform accurate inferences on the nature of initial conditions and formation histories for the observed LSS in these regions. For these reasons we believe that inferred final ensemble mean fields and corresponding voxel-wise standard deviations as a means of uncertainty quantification, may be of interest to the scientific community. These data products have been published as supplementary material along with the article, and are accessible at http://iopscience.iop.org/1475-7516/2015/01/036.

In summary, this chapter describes an application of the previously proposed BORG algorithm to the SDSS DR7 main galaxy sample. As demonstrated, our methodology produces a rich variety of scientific results, various aspects of which are objects of detailed and quantitative analyses in subsequent chapters of this thesis and forthcoming publications. Besides pure three dimensional reconstructions of the present density field, the algorithm provides detailed information on corresponding initial conditions, large scale dynamics and formation histories for the observed LSS. Together with a thorough quantification of joint and correlated observational uncertainties, these results mark the first steps towards high precision chrono-cosmography, the subject of analyzing the four dimensional state of our Universe.