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.

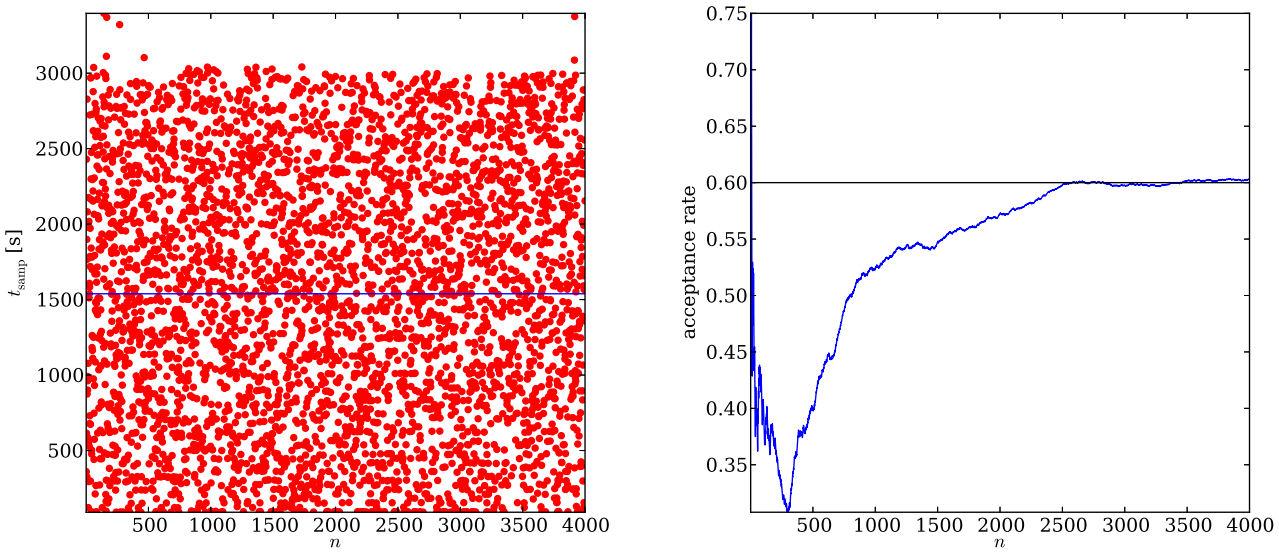

Figure 5.1: Diagnotics of the Markov chain: scatter plot of sample generation times (left panel) and Markov acceptance rates during the initial burn-in phase (right panel). As shown by the left panel, times to generate individual samples range from zero to about 3000 seconds. The average execution time 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.

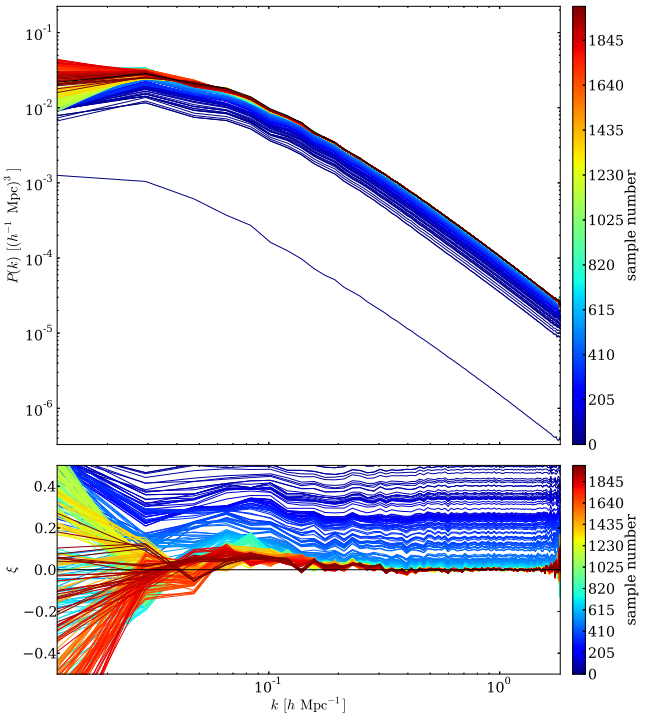

Figure 5.2: Burn-in power spectra measured from the first 2000 samples of the Markov chain colored corresponding to their sample number as indicated by the colorbar. The black line represents a fiducial reference power spectrum for the cosmology assumed in this work. Subsequent power spectra approach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

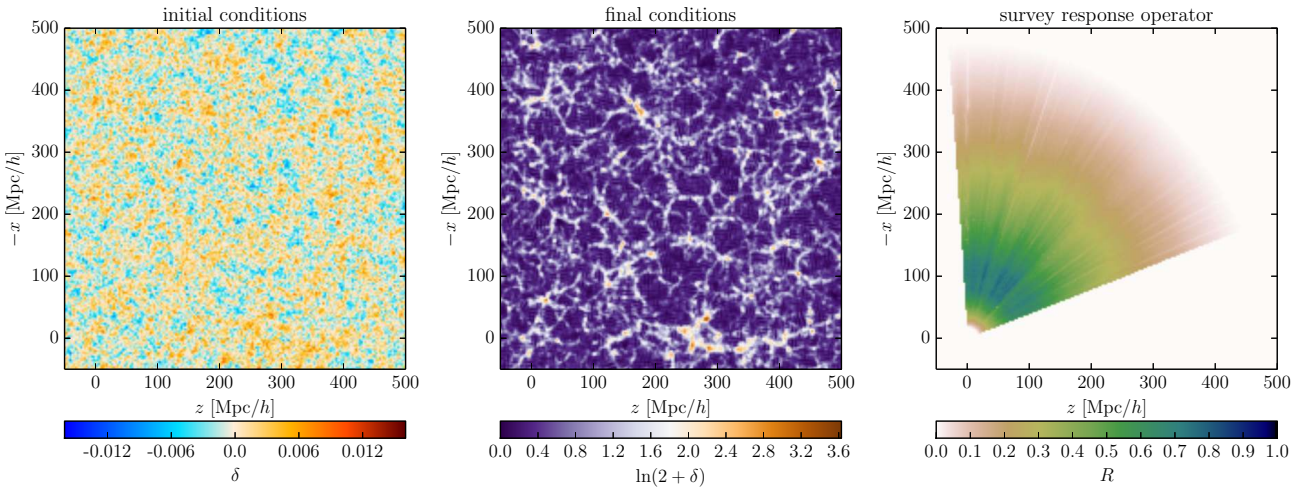

Figure 5.3: Slices through the initial (left panel) and corresponding final (middle panel) density fields of the 5000th sample. The right panel shows a corresponding slice through the combined survey response operator R for the six absolute magnitude bins considered in this work. As can be seen, unobserved 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

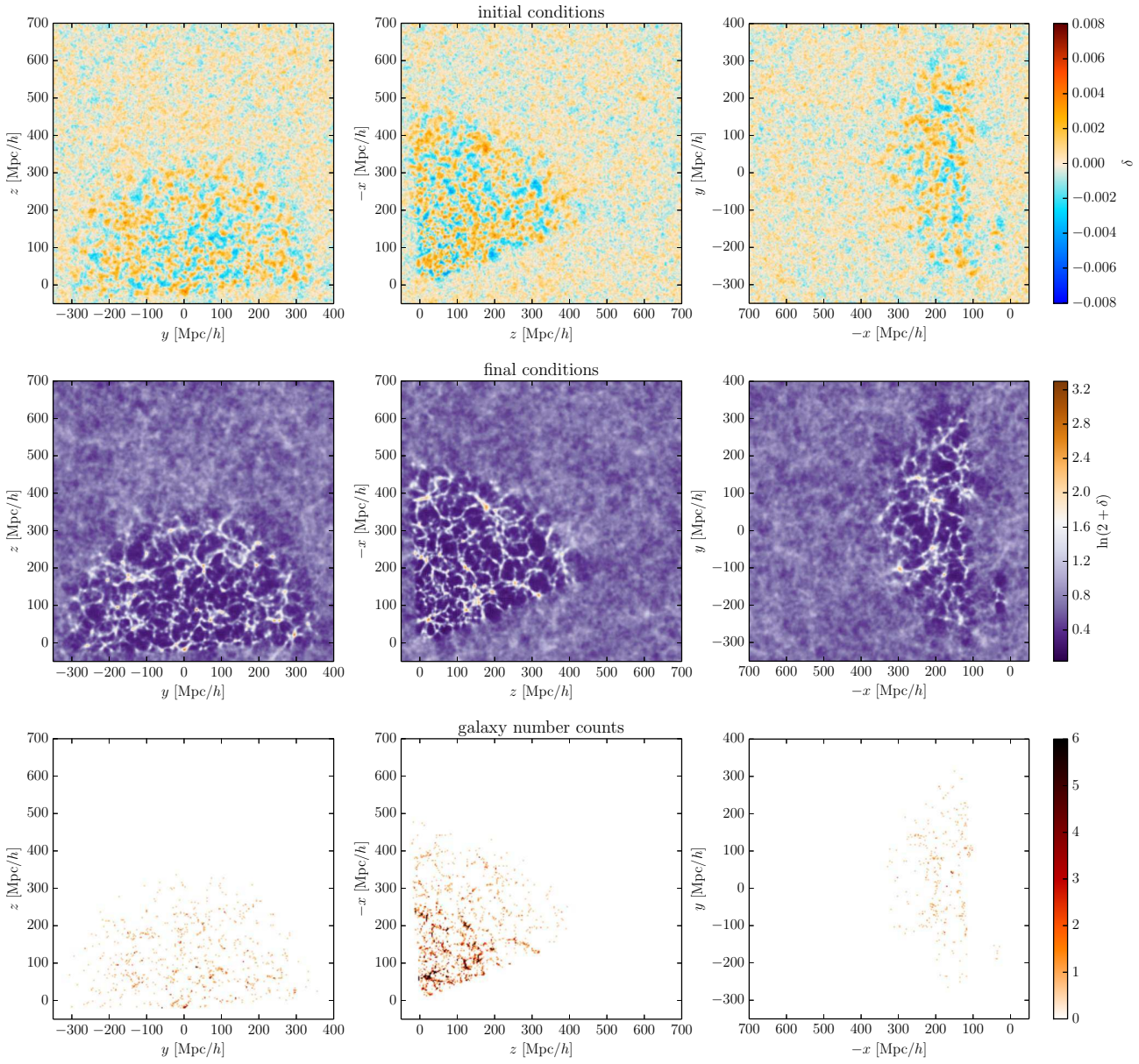

Figure 5.4: Three slices from different directions through the three dimensional ensemble posterior means for the initial (upper panels) and final density fields (middle panels) estimated from 12,000 samples. The lower panels depict corresponding slices through the galaxy number counts field of the SDSS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

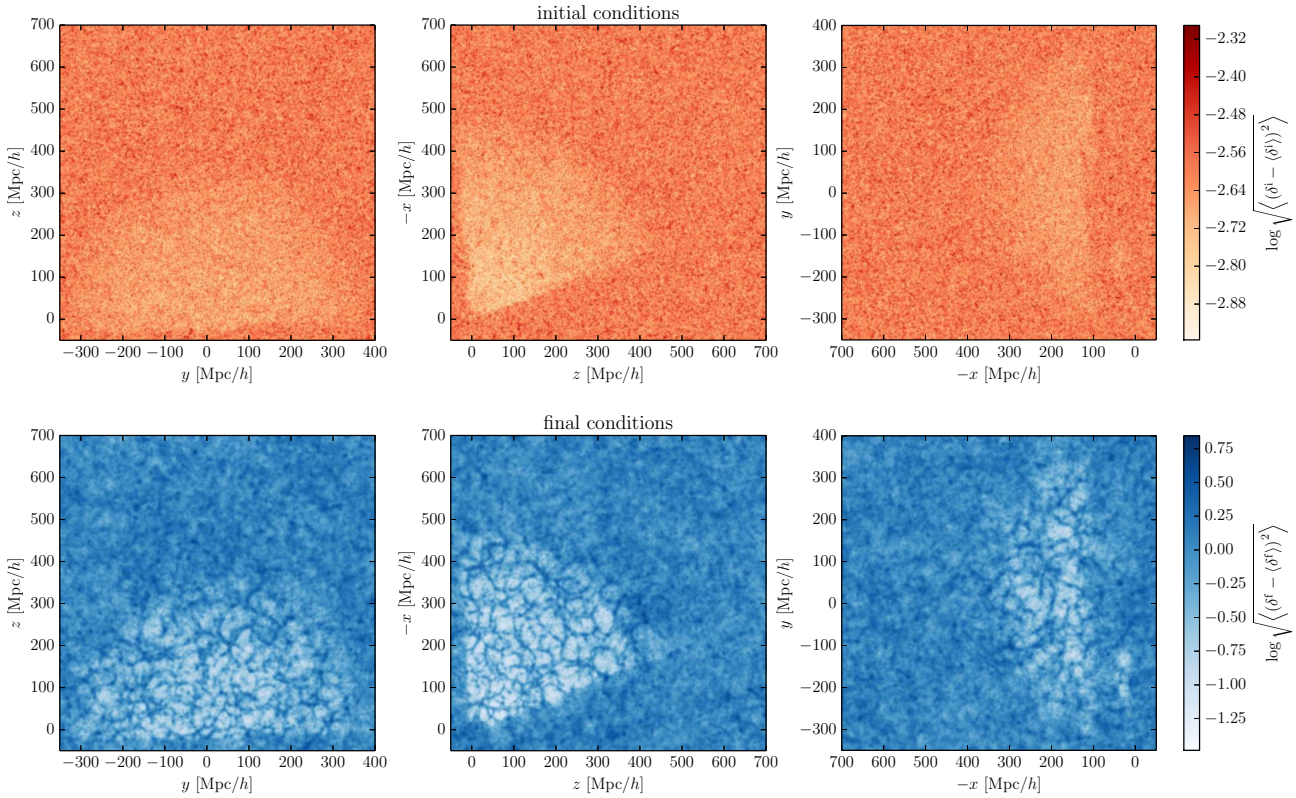

Figure 5.5: Three slices from different directions through the three dimensional voxel-wise posterior standard deviation for the initial (upper panels) and final density fields (lower panels) estimated from 12,000 samples. It can be seen that regions covered by observations show on average lower variance 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,

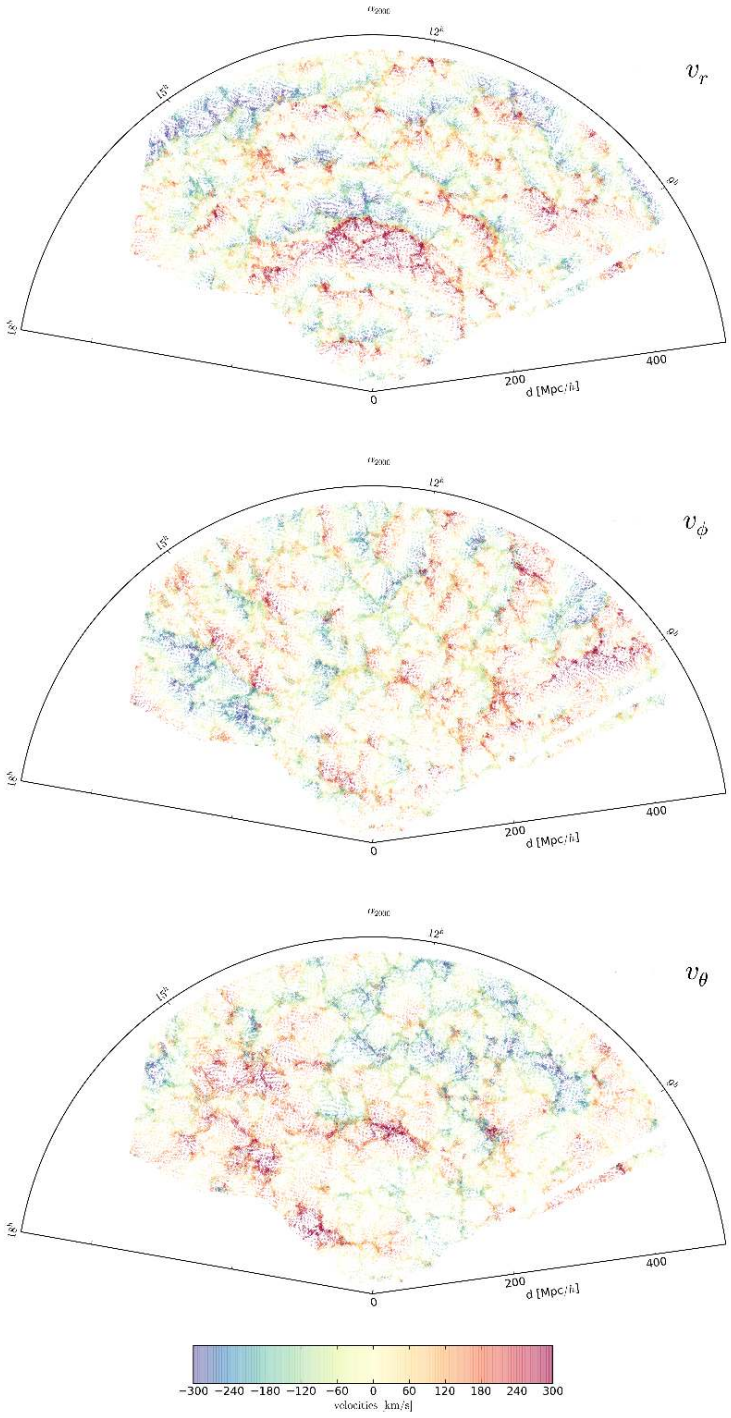

Figure 5.6: Slices through the 3D velocity fields, derived from the 5000th sample, for the radial (upper panel), polar (middle panel) and the azimuthal (lower panel) velocity components. The plot shows 2LPT particles in a 4 Mpc/h thick slice around the celestial equator for the observed domain, colored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

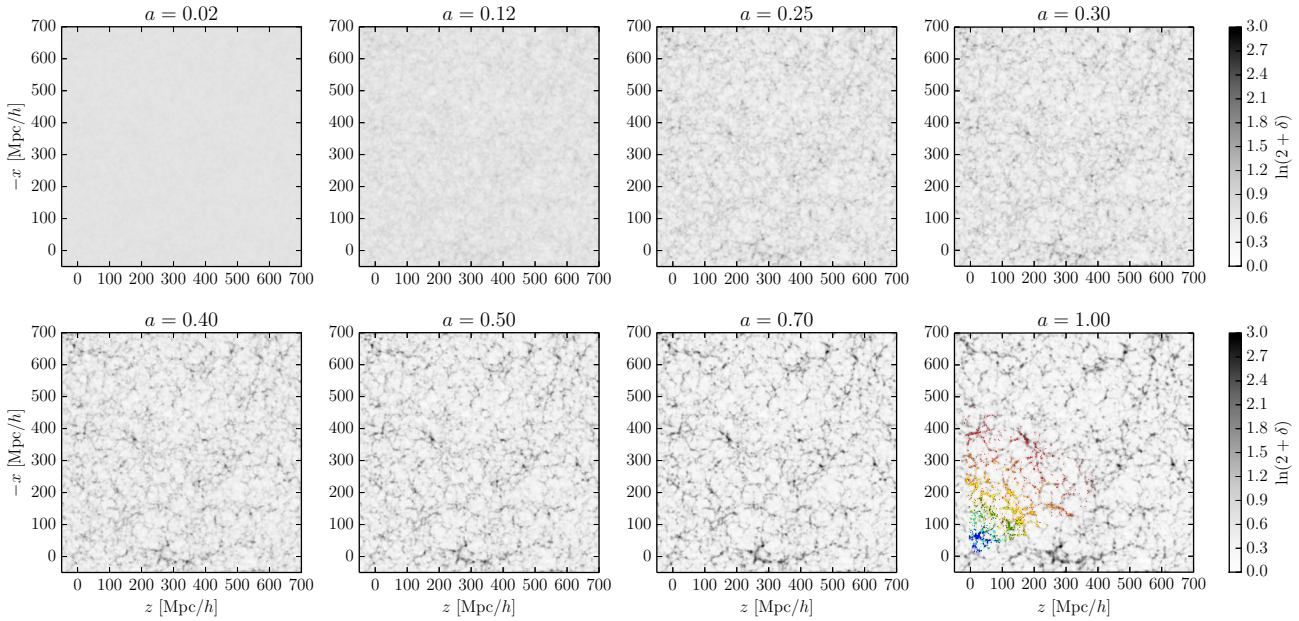

Figure 5.7: Slices through the inferred three dimensional density field of the 5000th sample at different stages of its evolution, as indicated by the cosmic scale factor in the respective panels. The plot describes a possible formation scenario for the LSS in the observed domain starting at a scale 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

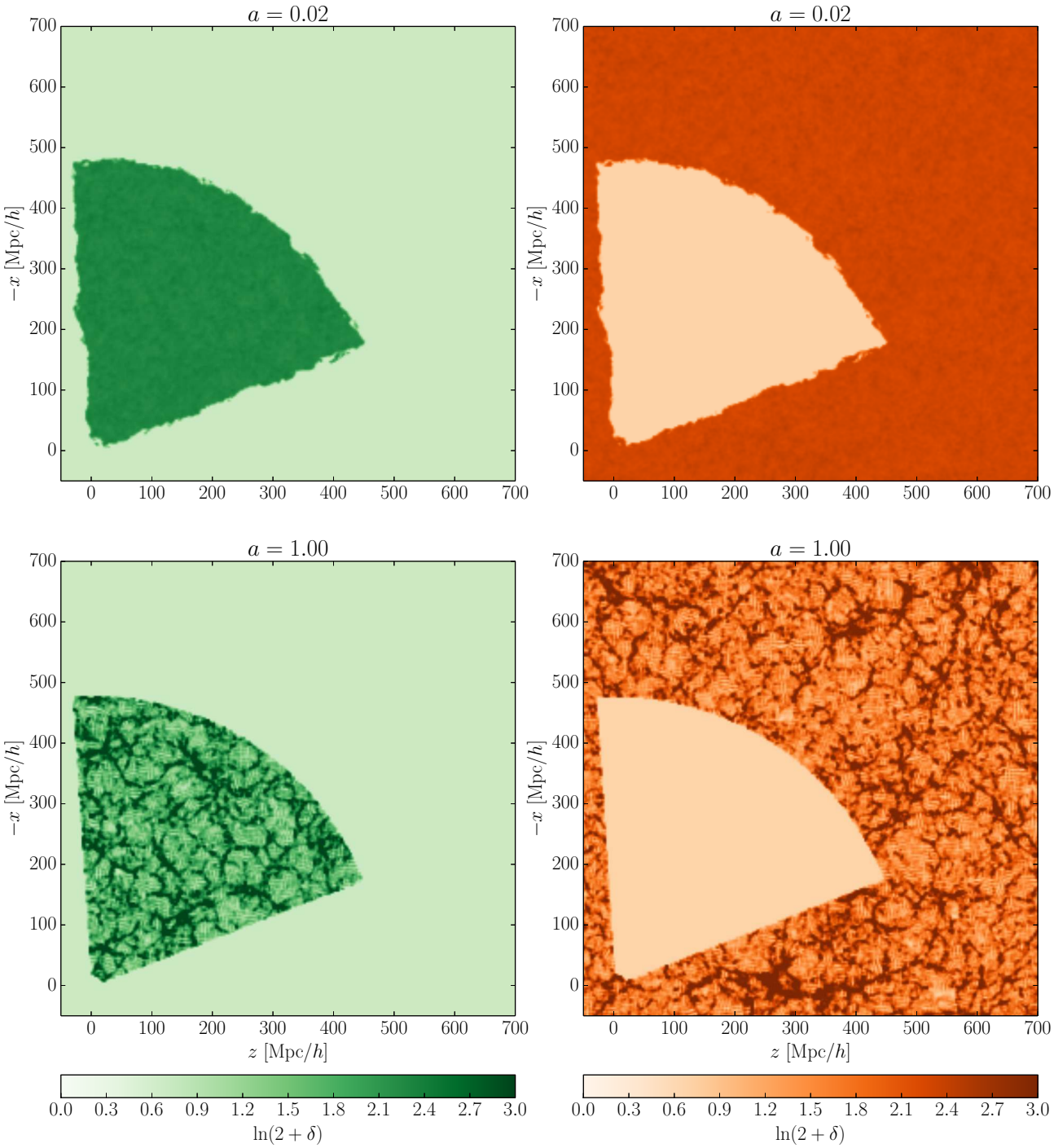

Figure 5.8: Slices through the distribution of particles in the 5000th sample, which are located inside (left panels) and outside (right panels) the observed domain at the time of observation, at two time snapshots as indicated above the panels. It can be seen that particles located within the observed 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.