

Systematic-free inference of the cosmic matter density field from SDSS3-BOSS data

Guilhem Lavaux,^{1*} Jens Jasche² and Florent Leclercq³

¹ CNRS & Sorbonne Université, UMR7095, Institut d’Astrophysique de Paris, F-75014, Paris, France

² The Oskar Klein Centre, Department of Physics, Stockholm University, AlbaNova University Centre, SE 106 91 Stockholm, Sweden

³ Imperial Centre for Inference and Cosmology (ICIC) & Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, United Kingdom

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We perform an analysis of the three-dimensional cosmic matter density field traced by galaxies of the SDSS-III/BOSS galaxy sample. The systematic-free nature of this analysis is confirmed by two elements: the successful cross-correlation with the gravitational lensing observations derived from *Planck 2018* data and the absence of bias at scales $k \simeq 10^{-3} - 10^{-2} h \text{ Mpc}^{-1}$ in the *a posteriori* power spectrum of recovered initial conditions. Our analysis builds upon our algorithm for Bayesian Origin Reconstruction from Galaxies (BORG) and uses a physical model of cosmic structure formation to infer physically meaningful cosmic structures and their corresponding dynamics from deep galaxy observations. Our approach accounts for redshift-space distortions and light-cone effects inherent to deep observations. We also apply detailed corrections to account for known and unknown foreground contaminations, selection effects and galaxy biases. We obtain maps of residual, so far unexplained, systematic effects in the spectroscopic data of SDSS-III/BOSS. Our results show that unbiased and physically plausible models of the cosmic large scale structure can be obtained from present and next-generation galaxy surveys.

Key words: large-scale structure of Universe – methods: statistical – methods: data analysis – gravitational lensing: weak – dark matter

1 INTRODUCTION

The measurement of clustering properties with modern galaxy surveys is achieving unprecedented precision on many scales of interest for cosmology (Ross et al. 2017). However, even for well-controlled galaxy surveys such as SDSS-III/BOSS (Eisenstein et al. 2011), systematic effects affect the largest spatial scales. This is clearly illustrated by the recent work of Kalus et al. (2019): for scales¹ $k \lesssim 10^{-2} h \text{ Mpc}^{-1}$ systematics remain a challenge in clustering analyses. Since sampling noise will be reduced with future surveys such as Euclid (Laureijs et al. 2011), the problem will further increase, hampering our capability to do cosmological inference.

Fortunately, over the last decade, Bayesian forward modelling of large-scale structures has come of age and may provide a way out. This method allows, assuming that the initial conditions are drawn statistically fairly from a Gaus-

sian distribution, to model the detail of the observed galaxy distribution. Notably, the BORG algorithm (Jasche & Wandelt 2013) has been successfully applied to the 2M++ galaxy compilation (Lavaux & Hudson 2011; Lavaux & Jasche 2016; Jasche & Lavaux 2019) and to the SDSS-II main galaxy sample (Abazajian & Survey 2009; Jasche et al. 2010). Other groups (in particular the ELUCID projet, Wang et al. 2014, 2016; Tweed et al. 2017) have developed techniques similar in spirit, meeting some success in applying to the SDSS-II main galaxy sample.

In this work, we apply the newly updated BORG analysis framework jointly to the two galaxy samples of SDSS-III/BOSS, LOWZ and CMASS. Most analyses run separate analysis on each component, which increases sample variance in their measurement. We are not limited by this aspect and we can add as many surveys as needed, provided that no double counting of a single galaxy occurs. We aim at recovering an unbiased ensemble of history of formation of the large-scale structure, and validate the model with Planck lensing maps (Planck Collaboration et al. 2018a). Solving this problem will open up new venues to extract cosmological information, in particular with the likelihood-based

* E-mail: guilhem.lavaux@iap.fr

¹ We use $h = H/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$ with H the Hubble constant at redshift $z = 0$.

ALTAIR extension of BORG (Ramanah et al. 2019) or with the likelihood-free SELFI algorithm (Leclercq et al. 2019).

An important feature of SDSS-III/BOSS is that it provides an excellent test case for the next generation of galaxy surveys, i.e. the Euclid mission (Laureijs et al. 2011) and the Large Synoptic Survey Telescope (LSST LSST Science Collaboration et al. 2009). It will be impossible to control everything in the data acquisition of these surveys, and the huge number of expected observed galaxies with photometric redshifts (~ 20 billions for LSST, ~ 1 billion for Euclid) will make all classical methods signal-dominated and dangerously sensitive to systematic errors (Laureijs et al. 2011; Colavincenzo et al. 2017; Monaco et al. 2018) to reach sub-percent precision the measurement of the density power spectrum. This further highlights the need to control systematic signals. There is also the interesting possibility that unexploited cosmological signal is available in the data of SDSS-III/BOSS.

To achieve a systematic-free inference we make use of a new likelihood (named ‘robust Poisson likelihood’, Porqueres et al. 2019), and a template matching method for systematics (Jasche & Lavaux 2017) to remove the impact of systematic effects on our inference. Usual galaxy survey data analysis rely fully on the existence of maps to correct for large scale, subtle, systematic effects. For example, Leistedt & Peiris (2014) compiled a set of 220 foreground maps of possible contaminants for the inference of the clustering signal of quasars in SDSS-III/BOSS. Later, Elsner et al. (2016, 2017) showed that ‘extended mode’ projection, i.e. the foreground template fitting technique, is almost surely biased, whereas ‘basic mode’ projection is unbiased in most cases. These two approaches resemble what has been done for analysis of Cosmic Microwave Background data obtained from space (Tegmark 1997) and ground observatories (e.g. for ACT and SPT, Fowler et al. 2010; Schaffer et al. 2011). In this work, we use the two kind of procedures at the same time.

Another big issue in analysing galaxy surveys is the derivation of the relation between galaxy population and the large scale dark matter field. This relation is generally called the galaxy bias model (Kaiser 1984; Desjacques et al. 2018). Typical analysis methods are calibrated on mock data from N -body simulations before being applied to galaxy surveys (Chuang et al. 2015; Kitaura et al. 2016; Beutler et al. 2017; Satpathy et al. 2017). A more agnostic procedure would fit this ‘bias’ model jointly with the inference of cosmological parameters, the underlying density field and the eventual residual due to systematic effects (Jasche & Lavaux 2017). However, finding a family of bias models sufficiently generic to capture all the unknown small-scale physics, extensible and fast to evaluate, is non-trivial (Schmidt et al. 2019). In this work, we present a novel bias model that has some of these properties.

This paper is organised as follows. In Section 2, we present the overall organisation of the BORG inference method, its assumptions and the essential new components of the adopted model to represent the galaxy distribution of SDSS-III/BOSS. In particular, we review the properties of our robust likelihood and introduce a new bias model. Next, we present the pre-processed data provided to the BORG inference machine in Section 3. We then describe our results on the systematic-free inference of the large-scale structure

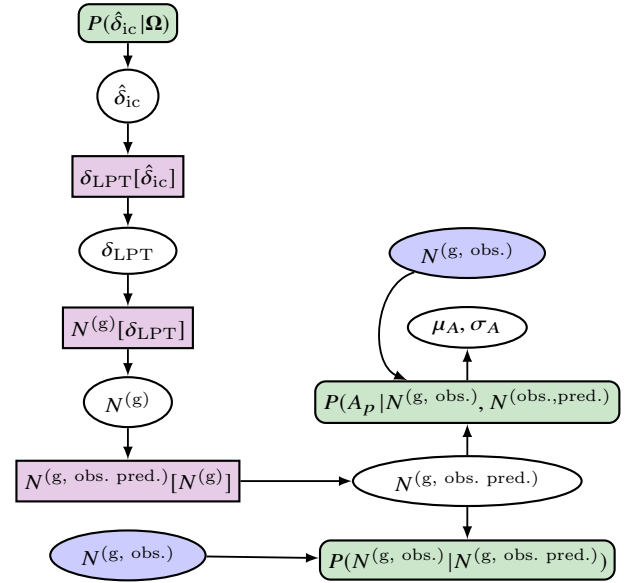


Figure 1. Hierarchical representation of the Bayesian inference framework used for the analysis of SDSS-III/BOSS. Ω represents cosmological parameters, $\hat{\delta}_{ic}$ is the set of Fourier modes encoding the initial conditions at $z \simeq 1000$, δ_{LPT} is the density field obtained from first-order Lagrangian Perturbation Theory, $N^{(g)}$ the galaxy number density field derived from the bias model, $N^{(g, \text{obs. pred.})}$ the galaxy number density field that should be observed after selection and systematic effects, $N^{(g, \text{obs.})}$ the actual SDSS-III/BOSS data, μ_A and σ_A the per-pixel mean and standard deviation of the inferred systematic maps. The details are provided in Section 2. Purple boxes correspond to a deterministic transition from one field to another. Green boxes are probability distributions modelling the field prior, like the top one which has a Gaussian form, or likelihood, like the bottom right ones. White ellipses are statistical variables. The two blue ellipses show the input from the data in the inference.

in Section 4, including the systematic maps that we have derived for the SDSS-III/BOSS sample. We conclude in Section 5.

2 METHOD

This section provides a detailed overview of the method used in this work. We focus on the salient features of the model that we have adopted to analyse the SDSS-III/BOSS data. We note that all the expressions are written as for only one galaxy catalogue. To reduce the number of indices, we have omitted to explicitly mention every-time that the inference is given a set of different independent galaxy catalogue. However the method as implemented does take this into account, and the expressions may be trivially generalised to the multi-catalogue case. We thus omit this in the rest of the section.

In Section 2.1, we give a short presentation of the statistical modelling and sampling algorithms, with references to our previous work where details can be found. Then we present the dynamical model that used in this work in Section 2.2, before describing the galaxy predictive model, i.e. the bias model, in Section 2.3. In Section 2.4, we present the foreground templates used to model known systematic effects. In Section 2.5, we move on our likelihood, designed

to absorb most of the other, unknown, systematic effects. Finally in Section 2.6, we show how this likelihood can be reversed to provide an inference procedure for the systematic maps themselves.

2.1 A probabilistic physical forward model of dark matter dynamics

As mentioned in the introduction, this work describes the extension and application of our previously-developed BORG algorithm. BORG aims at inferring a fully probabilistic and physically plausible model of the three-dimensional matter distribution from observed galaxies in cosmological surveys (see e.g. Jasche & Wandelt 2013; Jasche et al. 2015; Lavaux & Jasche 2016). This framework solves a large-scale Bayesian inverse problem by fitting a dynamical structure formation model to data and inferring the primordial initial conditions from which presently-observed structures formed. The development of this framework was stemmed by convincing evidence from Cosmic Microwave Background observations that the statistics of initial conditions are close to Gaussian (Planck Collaboration et al. 2019). In contrast, the statistics of present large-scale structures are very complex and strongly non-Gaussian. It happens that modelling the change of coordinates relating initial conditions to visible large-scale structures is not so complicated in the cosmological paradigm, and it is even feasible to sample the initial states with a Monte Carlo algorithm, currently the Hamiltonian Markov Chain Monte Carlo algorithm (Jasche & Wandelt 2013). This physical forward modelling approach naturally accounts for the formation of non-linear and non-Gaussian large-scale structures, associated with statistics of the density field beyond 2-point correlations, redshift-space distortions and light-cone effects. As a result, the algorithm provides plausible three-dimensional matter density fields, but also performs a full four-dimensional state inference and recovers the dynamic formation history and velocity fields of the cosmic large-scale structures.

The method also accounts for systematic and stochastic uncertainties, such as survey geometries, selection effects, unknown noise and galaxy biases, as well as foreground contamination (see e.g. Jasche & Wandelt 2013; Jasche et al. 2015; Lavaux & Jasche 2016; Jasche & Lavaux 2017; Porqueres et al. 2019). For further details on the statistical inference machinery and solutions to the described large scale Bayesian inverse problem, the reader is referred to our previous work (Jasche & Wandelt 2013; Jasche et al. 2015; Lavaux & Jasche 2016; Jasche & Lavaux 2019).

We note that our model includes many components (initial Fourier modes amplitudes and phases, galaxy bias and foreground contamination). Parameters related to these components are all injected in a probabilistic framework which is bound together on one side by Bayesian priors, typically Gaussian initial conditions, and by the likelihood of galaxy observations on the other side. A graphical summary of all the steps and connections involved in this Bayesian inference is given in Figure 1. As already detailed many times in our previous work, we use *sampling* algorithms to build a fair ensemble of points in the posterior parameter space, providing globally a numerical approximation to the posterior density distribution. We do *not use an iterative* procedure which would provide a single answer for the reconstruction

problem, and would potentially bias the result as in the case of the Wiener filter (Rybicki & Press 1992). We want to stress this point to reduce misconceptions about our results.

2.2 Dynamics and light cone model

The target resolution of the inferred initial conditions and modelled galaxy distribution, discussed in the Section 3, is about $16h^{-1}$ Mpc. At this resolution, the mass density on the mesh at that scale is only affected by mildly non-linear dynamics at low redshift. We thus limit our model of the dynamics in BORG to first-order Lagrangian Perturbation Theory (LPT), also called the Zel'dovich approximation (Zel'Dovich 1970; Bouchet et al. 1995). This model predicts the matter density field and its dynamics with sufficient accuracy at the scales relevant to this work ($\sim 16h^{-1}$ Mpc, e.g. Bernardeau et al. 2002).

In BORG, we use a particle representation to implement the LPT model. The relation between the final Eulerian position \mathbf{x} and the initial Lagrangian position \mathbf{q} of each particle is given as:

$$\mathbf{x}(\mathbf{q}) = \mathbf{q} + D_+(a_f)\Psi(\mathbf{q}), \quad (1)$$

with $\Psi(\mathbf{q})$ the displacement field derived from the initial conditions proposed by the sampling algorithm, and $D_+(a_f)$ the linear growth factor at the scale factor a_f . We encourage the reader to refer to the previous publications (Jasche & Wandelt 2013; Jasche & Lavaux 2019) for further details on the numerical implementation of the forward and adjoint gradient. For the purpose of this work, we further modify the model by adjusting D_+ depending on the distance to the observer to simulate a light cone. The new evolution equation is thus

$$\mathbf{x}(\mathbf{q}) = \mathbf{q} + D_+(a(|\mathbf{q}|))\Psi(\mathbf{q}), \quad (2)$$

with $a(d)$ the relation between the comoving distance d and the scale factor of the homogeneous Universe at that look-back distance. This model is an approximation to the full problem of light cone building which involves computing intersections of two trajectories, the trajectory of the light emitting object and the geodesic of the photon emitted by that object and detected through observatories. The average error produced by this approximation is given by the typical amplitude of the displacement field $\Psi(\mathbf{q})$. In a Λ CDM universe with Planck cosmology (Planck Collaboration et al. 2018b), that displacement is of order $5h^{-1}$ Mpc, with a maximum of $\sim 20h^{-1}$ Mpc for the fastest moving objects, which is of the same size as a volume element of our mesh (see Appendix A). At the farthest distance, this approximation means that we neglect additional coherent distortion of about $10h^{-1}$ Mpc. At the present resolution, this is still acceptable. However, future improvement on the resolution will require to investigate the detailed impact of the light cone effect on the reconstruction.

2.3 Galaxy bias model

In this work, we introduce a new bias model which is built on a few requirements: i/ galaxy formation is a non-local process in Eulerian coordinates, meaning that the model must be somehow sensitive to the environment, and not necessarily linearly; ii/ it should follow features of the linear bias

model as much as possible on large scales, for better comparison with earlier literature, and (more interestingly) to offer a connection to generic perturbative bias expansion (Desjacques et al. 2018; Schmidt et al. 2019; Elsner et al. 2019); iii/ it must ensure positivity of the final galaxy population field to give physically meaningful predictions. As such, we introduce the following bias model to predict the number of galaxies $N_i^{(g)}$ in a mesh element indexed by i :

$$N_i^{(g)} = \Delta_i^\dagger \mathbf{Q} \Delta_i, \quad (3)$$

with \mathbf{Q} a positive definite matrix and Δ_i a vector formed from local averages of the matter density contrast field δ . In order to guarantee that any sampled matrix is positive definite, we use the Cholesky decomposition of \mathbf{Q} as sampling parameters, i.e. the matrix \mathbf{L} with $\mathbf{Q} = \mathbf{L}\mathbf{L}^\dagger$. The vector Δ_i is defined as follows:

$$\Delta_i^\dagger = \left(1, \delta_i^{(1)}, (\delta_i^{(1)})^2, \dots, \delta_i^{(2)}, (\delta_i^{(2)})^2, \dots, (\delta_i^{(3)}), \dots \right), \quad (4)$$

with $\delta_i^{(\ell)} = A^{(\ell)}(\{\delta_i\})$, $A^{(\ell)}$ being an averaging operation in a neighbourhood of the i -th mesh element for $\ell \geq 2$ or the identity for $\ell = 1$. This can be written more compactly as

$$\Delta_{i,a} = \left(\delta^{(\ell_a)} \right)^{\gamma_a} \quad (5)$$

for $\ell_a \geq 0$ and $\gamma_a \geq 0$. In this work, the averaging is done in practice with an oct-tree structure. The level $\ell = 1$ is directly the density fluctuation at the finest level, i.e. δ . For higher levels, $\ell > 1$, we derive the density fluctuations using the following relation:

$$\delta_{x,y,z}^{(\ell)}(\{\delta_i\}) = \frac{1}{8^\ell} \sum_{a,b,c=0}^{2^\ell-1} \delta_{m^{(\ell)}(x,a), m^{(\ell)}(y,b), m^{(\ell)}(z,c)}, \quad (6)$$

with the coarsening operator

$$m^{(\ell)}(x, a) = 2^\ell \lfloor x/2^\ell \rfloor + a. \quad (7)$$

For the purpose of Hamiltonian Markov Chain exploration used in BORG, we compute analytically and provide the adjoint gradient of the above model in Appendix B.

2.4 Foreground templates

A major point of contention in data analysis is the level of systematic effect contaminating the observational data. The contamination affects the spectroscopic sample of galaxies by hindering a proper uniform target selection from pure photometry. Indeed, to build a galaxy sample, one must generally start with broadband photometry, from which a list of candidates for spectrum measurement is built. Once its spectrum is measured, each candidate object is classified, e.g. as a star or a galaxy. Any bias in target selection can affect the resulting samples of classified objects. For this reason, the final spectroscopic sample of galaxies reflects the biases of the target selection procedure.

In the case of SDSS-III/BOSS, several groups have studied the possible implication of different contaminants (e.g. Ross et al. 2012; Leistedt & Peiris 2014). In this work, we follow the model presented in Jasche & Lavaux (2017) to represent the effect of a small number of systematics maps, which we use to benchmark the effectiveness of the robust

likelihood mechanism presented in the next section. The assumed model is multiplicative, i.e.

$$N_i^{(g, \text{ obs. pred.})} = R_i \prod_a (1 + \alpha_a F_{a,i}) N_i^{(g)}, \quad (8)$$

with $N_i^{(g, \text{ obs. pred.})}$ the predicted mean number of galaxies at mesh element i that is observed given observational constraints (mask and systematics), and $N_i^{(g)}$ the predicted mean number of galaxies from the dynamical model in the same mesh element, as obtained in Equation (3). In the above equation, we have also introduced R_i the linear survey response, which accounts for the mask and the selection effects (radial and angular), $F_{a,i}$ the value of the foreground template a in mesh element i , and α_a the intensity of foreground a . The linear response is generally provided as part of the meta-data of a given survey. It is estimated from the target and spectroscopic sample. In the case of SDSS-III/BOSS, that is just the ratio between those two samples for each angular direction. The parameter α_a is left free sampled directly from its posterior given the data. As mentioned in the introduction of Section 2, there is one parameter for each foreground and for each for *each catalogue* part of the inference problem. This multiplicative foreground model can reasonably model a broad class of systematic effects, such as intergalactic absorption of light by dust, atmospheric effects or fibre collisions. However, it is limited to known effects for which sky models exist.

2.5 Robust likelihood

The SDSS-III/BOSS survey has been designed to optimally study galaxy clustering at the scales of BAOs. While control of systematic effects at these scales has been studied in detail by the SDSS collaboration (e.g. Reid et al. 2016), there exists no equally-good understanding of the impact of systematic effects at the largest scales of the galaxy distribution, typically for modes of wavenumber $k \lesssim 10^{-2} h \text{ Mpc}^{-1}$. So far, state-of-the-art data analysis methods have had limited success in removing some of the large-scale systematic effects inherent to the observations (Kalus et al. 2019). These results indicate that there probably exists a scale in data beyond which galaxy clustering is not understood because it is not modelled sufficiently well using known foreground templates. To address this issue, Porqueres et al. (2019) developed a new likelihood, based on Poisson statistics, which is designed to be robust against unknown foreground contamination at a scale given *a priori*. The underlying idea relies on the assumption that the physically modelled galaxy distribution can be related to the observed one up to some overall scaling over patches on which the unknown foreground modulation is quasi constant. These patches can be chosen in any convenient way for the analysis. In practice, we use a three-dimensional extrusion of pixels of a HEALPIX map, yielding a 3d patch map. This allows us to group pixels with quasi constant foreground amplitudes into sets \mathcal{A}_m , where m runs over indices of an HEALPIX map. The effective predicted Poisson intensity is $A_{p_i} N_i^{(\text{ obs. pred.})}$, where i is a mesh element index of the 3d grid covering the considered volume of Universe, p_i the index of the patches containing i (otherwise said $i \in \mathcal{A}_{p_i}$), and $N_i^{(g, \text{ obs.})}$ the raw galaxy count intensity predicted by the dynamical model and the galaxy

bias model. We build the following probabilistic model:

$$P(\{N_i^{(\text{g, obs.})}\}, \{A_p\} | \{N_i^{(\text{obs. pred.})}\}) = \prod_p \left[\prod_{i \in \mathcal{A}_p} \text{Poisson}(A_p N_i^{(\text{g, obs. pred.})}) \right] \pi(A_p), \quad (9)$$

with $\{N_i^{(\text{g, obs.})}\}$ the number count of galaxies observed in the voxel i . We choose $\pi(A) \propto 1/A$ as the prior probability of the amplitude of the unknown systematic, which thus follow a Jeffreys' prior. The problem of large scale structure inference uses the marginalised version of that probability, and was shown to be resilient to unknown systematics in a test on mock data (Porqueres et al. 2019). After marginalisation over $\{A_p\}$, the new likelihood takes a simple form:

$$P(\{N_i^{(\text{g, obs.})}\} | \{N_i^{(\text{obs. pred.})}\}) \propto \prod_p \prod_{i \in \mathcal{A}_p} \left(\frac{N_i^{(\text{obs. pred.})}}{\sum_{j \in \mathcal{A}_p} N_j^{(\text{obs. pred.})}} \right)^{N_i^{(\text{g, obs.})}}. \quad (10)$$

We immediately notice that this likelihood is insensitive to absolute scales in the predicted galaxy number intensity $\{N_i^{(\text{obs. pred.})}\}$, which is an appealing feature: the ratio cancels any contribution over a scale corresponding to the assumed smoothness of the foreground contamination. Robustness tests are described in more details in Porqueres et al. (2019).

2.6 Systematic map inference

The robust likelihood is designed to ignore information on some spatial scale. In the case of this work, we limit ourselves to ignore information above some angular scale, even if the framework would also work also for complex 3d scales. However, we may still use the inferred density field to solve the reverse problem of inferring the systematic effects that were ignored within the Markov Chain Monte Carlo analysis (MCMC, see Section 4). In doing so, we obtain complete maps of the unknown systematic effects down to some angular scale. For one patch \mathcal{A}_p , we may derive the conditional probability of the value taken by A_p from Equation (9)

$$P(A_p | \{N_i^{(\text{g, obs.})}\}, \{N_i^{(\text{obs. pred.})}\}) \propto \frac{1}{A_p} \prod_{j \in \mathcal{A}_p} \text{Poisson}(A_p N_j^{(\text{obs. pred.})}). \quad (11)$$

For most purposes, we are only interested in the first two moments of the above distribution, the mean and the variance. These may be computed analytically:

$$\mu_{A_p | N^{(\text{obs. pred.})}} = \langle A_p \rangle = \frac{\sum_{i \in \mathcal{A}_p} N_i^{(\text{obs})}}{\sum_{i \in \mathcal{A}_p} N_i^{(\text{obs. pred.})}}, \quad (12)$$

$$\begin{aligned} \sigma_{A_p | N^{(\text{obs. pred.})}}^2 &= \langle (A_p - \mu_{A_p})^2 \rangle \\ &= \frac{\sum_{i \in \mathcal{A}_p} N_i^{(\text{obs})}}{\left(\sum_{i \in \mathcal{A}_p} N_i^{(\text{obs. pred.})} \right)^2}. \end{aligned} \quad (13)$$

In the above, we have used the following identity to compute the integral over the Poisson distribution:

$$\int_0^{+\infty} dx x^\alpha \exp(-\beta x) = \alpha! \beta^{-\alpha-1}. \quad (14)$$

We note that we did not specify the derivation of the set \mathcal{A}_p for each patch of interest. In our case, we are interested in computing sky maps at different redshift of detectable systematic effects, given our model of large scale structures. We use the HEALPIX pixelization to represent these maps. Each patch thus corresponds to the cosmological volume that projects in each pixel of the sought map. We build the set \mathcal{A}_p by throwing 100 uniformly distributed rays at random within each pixel and recording the voxels that are traversed. Because voxel has a finite size, many rays for different HEALPIX pixels will traverse the same voxels. This means that the maps derived from this procedure will have nearby pixels with highly correlated values. That is not a fundamental limitation but a choice of representation of the systematic map that we aim to derive. The value for the patches derived from the posterior analysis would be completely decorrelated if we had decided to choose a non-overlapping set of voxels to compute pixel values.

We note that the average and the variance per pixel given in Equations (12) and (13) are for one particular model of large-scale structure given by the set of values $\{N_i^{(\text{g, obs. pred.})}\}$. BORG provides an ensemble of plausible values for this field. The probability for the set of pixels $\{A_p\}$ is thus:

$$\begin{aligned} P(\{A_p\} | \{N_i^{(\text{g, obs.})}\}) &= \int \left(\prod_{j=1}^{N_g} dN_j^{(\text{obs. pred.})} \right) P(\{N_j^{(\text{obs. pred.})}\} | \{N_i^{(\text{obs})}\}) \times \\ &P(\{A_p\} | \{N_i^{(\text{obs})}\}, \{N_j^{(\text{obs. pred.})}\}) \\ &\simeq \frac{1}{N_{\text{sample}}} \sum_c P(\{A_p\} | \{N_i\}, \{N_{i,c}^{(\text{obs. pred.})}\}), \end{aligned} \quad (15)$$

with $N_{i,c}^{(\text{obs. pred.})}$ the predicted observed galaxy intensity in voxel i for the Markov chain sample c , N_{sample} the number of considered samples in the MCMC, and N_g the number of mesh element to represent the matter density field. Thus, the marginalised mean and variance at each pixel position is computed by taking the average over the Markov chain of the mean and variance given by Equations (12) and (13).

3 THE SDSS-III/BOSS DATA

We apply our Bayesian inference approach to galaxies observed by the Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013), the third generation of the Sloan Digital Sky Survey (SDSS-III, Eisenstein et al. 2011). The BOSS survey is dedicated to observing the three-dimensional clustering of 1.37 million galaxies with spectroscopic redshifts covering about 10 000 deg² of the sky over two contiguous regions in the Northern and Southern Galactic caps. This work uses the final data release DR12²

² <https://data.sdss.org/sas/dr12/booss/lss/>

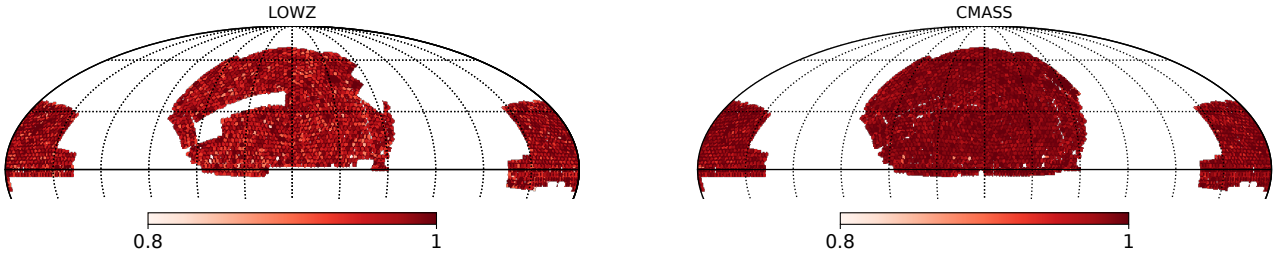


Figure 2. Completeness maps of SDSS-III/BOSS for the LOWZ sample (left panel) and CMASS sample (right panel). These completeness maps are directly derived from the DR12 repository and rendered on an HEALPIX mesh at $N_{\text{side}} = 2048$. We note the usual vetoed regions in LOWZ corresponding to a problem in the target selection that occurred during the first year of data acquisition (Parejko et al. 2013).

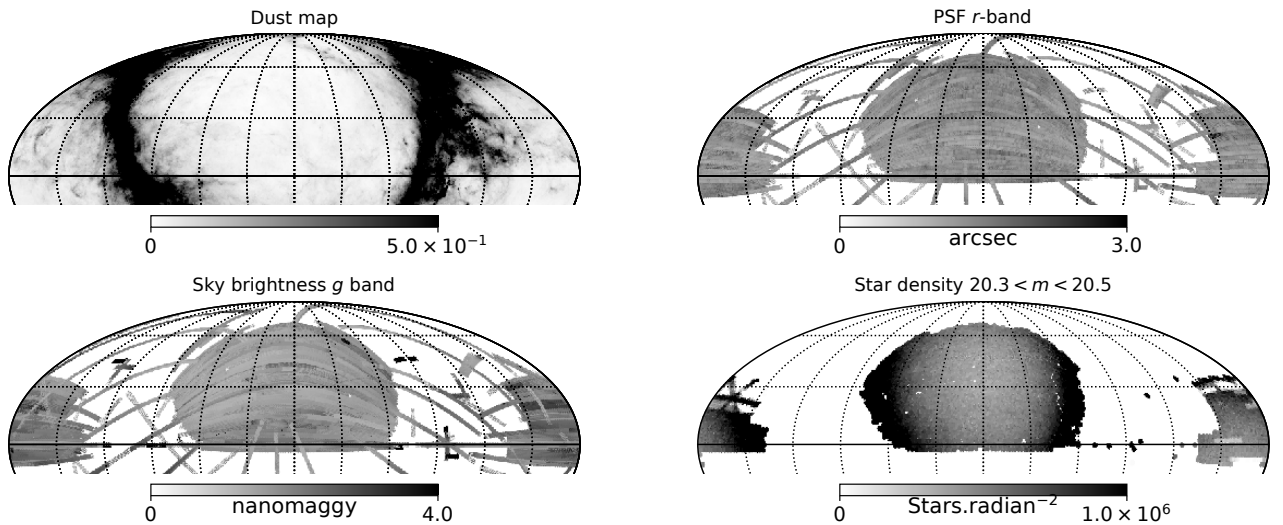


Figure 3. Four of the eleven known systematic maps that we have used in the inference presented in this work. Each of these maps was rendered at $N_{\text{side}} = 2048$ from a MANGLE representation. The above are dust induced reddening (top left), point spread function (PSF) in the r band (top right), sky flux in the g band (bottom left), density of stars with an apparent magnitude $20.3 < i < 20.5$ (bottom right). We note the typical striping induced by the SDSS scanning strategy for the top right and bottom left maps.

Name	Definition
dust	Dust induced reddening (Schlegel et al. 1998)
sky flux	Photometric sky flux in the indicated band (5 maps)
airmass r	Air mass above telescope, r band
psf r	Point spread function, r band
star 0	Density of stars with $20.5 < i < 20.3$
star 1	Density of stars with $20.3 < i < 20.1$
star 2	Density of stars with $20.1 < i < 19.0$

Table 1. Name convention for the 11 foreground templates used in this work. The coefficients α_a attached to each of these templates in Equation (8) are sampled jointly with the bias parameters and the matter density field.

of SDSS-III/BOSS, containing the data of all six years of the survey (Alam et al. 2015). Galaxies were targeted uniformly in a low-redshift sample with $z < 0.45$ (LOWZ). To select additional massive galaxies in the redshift range $0.4 < z < 0.8$, several colour cuts were applied to the SDSS-III imaging data in the (u, g, r, i, z) bands (Fukugita et al. 1996). Accord-

ing to the passively evolving model, these selections result in a sample (CMASS) that has a constant stellar mass limit over the redshift range $0.4 < z < 0.8$ (Maraston et al. 2009). Large stellar masses imply strong galaxy biases with respect to the underlying dark matter density field. This property induces that each galaxy provides strong indication of large scale matter fluctuations, yielding more information on large scale structure analysis than their lower stellar mass counterpart. Detailed descriptions of BOSS targeting criteria, data reduction methods and the construction of the large-scale structure catalogue are described in Eisenstein et al. (2011); Dawson et al. (2013); Alam et al. (2015); Reid et al. (2016). More specifically, we use large-scale structure catalogues provided by the BOSS galaxy clustering working group (Anderson et al. 2014; Reid et al. 2016). Their sample assigns weights to galaxies to correct for non-cosmological fluctuations imprinted on the target catalogue by imperfections in the acquisition of spectroscopic redshifts due to fibre collisions, precluding simultaneous assignments of spectroscopic fibres to targets closer than $62''$ (Ross et al. 2011;

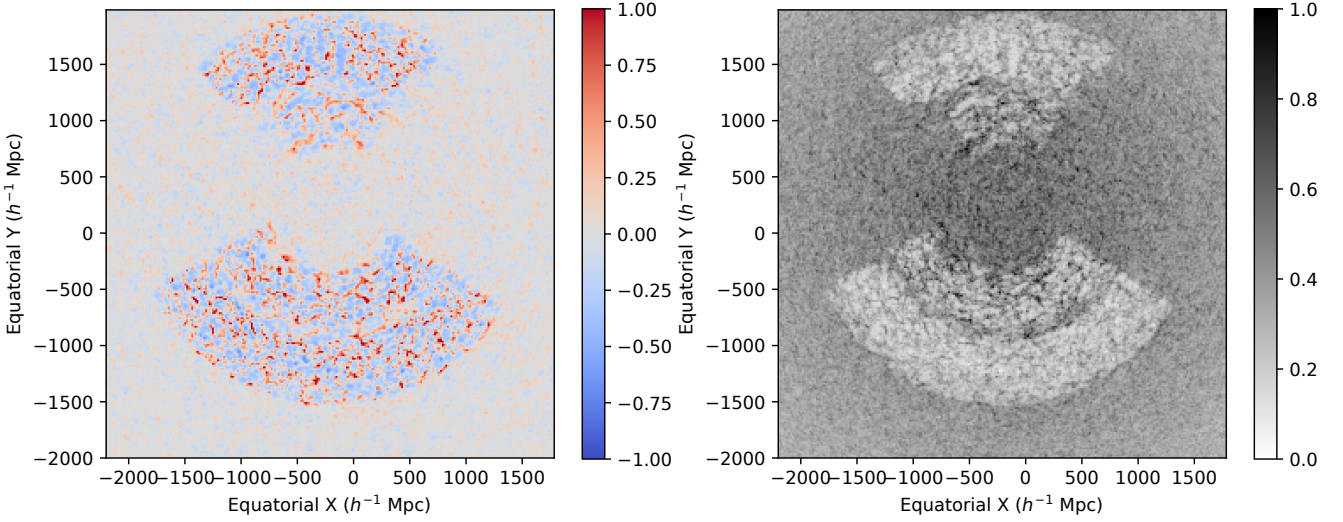


Figure 4. Slices through the inferred three-dimensional ensemble mean density field (*left panel*) and the corresponding standard deviations of density amplitudes (*right panel*) obtained from the Markov Chain. As can be seen, the algorithm recovers the filamentary large-scale structures in regions sampled by galaxies of the SDSS-III/BOSS survey, while matter density approaches cosmic mean in unobserved regions. Correspondingly, the map of standard deviations shows low variance in observed regions and correctly provides higher uncertainty in unobserved regions. The plot illustrates that BORG provides detailed reconstructions of matter density fields and corresponding uncertainty quantification. The coordinates are all comoving assuming the cosmology given in Section 4. The residual small fluctuations outside the observed region in the left panel are due to the finite length of the Markov Chain.

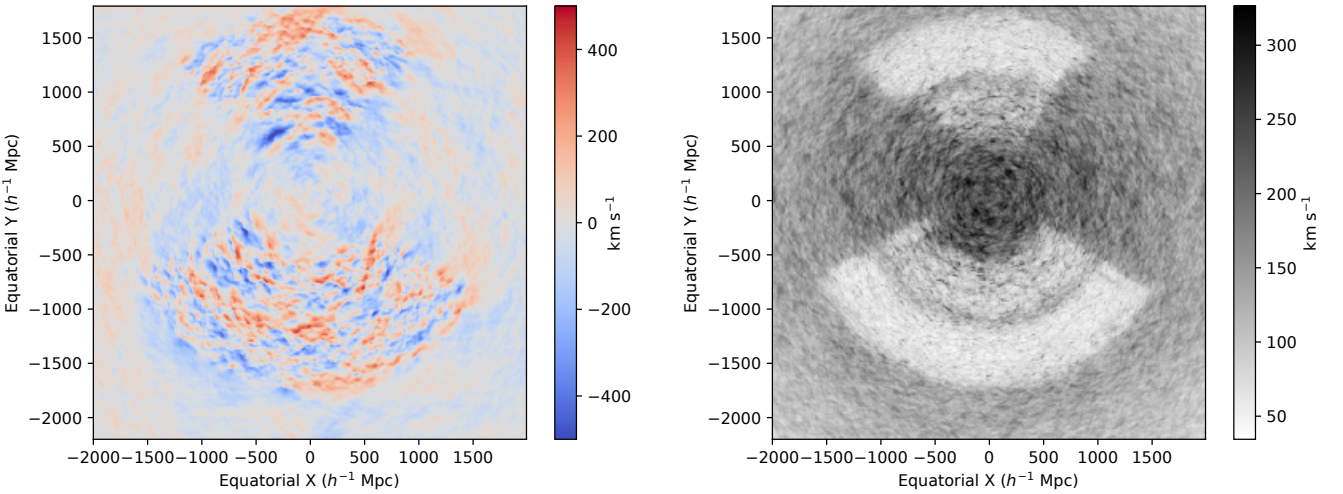


Figure 5. Slices through the cosmic velocity field derived with BORG using data from SDSS-III/BOSS. *Left panel*: ensemble average of the line-of-sight component of the velocity field. *Right panel*: standard error of the mean of the velocity field (standard deviation derived from the posterior). The effect of the light cone is visible (giving a factor of ~ 2 difference in the standard deviation between $z = 0$ and $z = 0.7$).

Ho et al. 2012; Ross et al. 2012). Additional weights are assigned to compensate systematic effects between observed galaxy number densities and the seeing (Reid et al. 2016). To account for survey geometry and spectroscopic completeness, we used the MANGLE software (Swanson et al. 2008) to create corresponding HEALPIX (Gorski et al. 2005) maps at $N_{\text{side}} = 2048$ shown in Fig. 2. To account for redshift dependence in radial selection functions we split the LOWZ and CMASS sample into four redshift bins and two galac-

tic caps each, making a total 16 sub-catalogues. We have thus 4 redshift bins for each of the following catalogue: LOWZ north galactic cap (NGC), LOWZ south galactic cap (SGC), CMASS NGC and CMASS SGC. The four distance bins for LOWZ are chosen as $[600, 750]h^{-1}$ Mpc, $[750, 900]h^{-1}$ Mpc, $[900, 1050]h^{-1}$ Mpc, $[1050, 1200]h^{-1}$ Mpc. Similarly we have four sub-catalogues for CMASS with distance bins chosen as $[1000, 1200]h^{-1}$ Mpc, $[1200, 1400]h^{-1}$ Mpc, $[1400, 1600]h^{-1}$ Mpc, $[1600, 1800]h^{-1}$ Mpc.

The BORG algorithm permits us to treat the respective systematic effects, such as redshift dependent galaxy biases, selection effects and foreground contamination, for each of these 16 galaxy sub-samples. In the following we will provide more detail to our data analysis procedure.

While not strictly required by our robust likelihood framework, we also derive systematic maps from the meta-data of the SDSS-III/BOSS photometric database. As mentioned in Section 2.4, we make them part of the data model to learn about specific features that could still be there even after the cleaning performed by the robust likelihood. We have included 11 foregrounds which have been classically considered by the SDSS-III/BOSS collaboration (e.g. Ross et al. 2012) and which still potentially contaminate the data, despite applying weights to galaxies. The inferred amplitude values will be used to assess the amount of residual corrections that are still necessary at small scales. The known systematic effects that we consider are summarised in the Table 1. We note that we consider the impact of the sky flux independently in each of the SDSS photometric band (u, g, r, i, z), thus the line “sky flux” corresponds actually to five maps. Each of these maps was derived using the MANGLE (Swanson et al. 2008) geometry file describing the SDSS-III/BOSS large structure sample.³ Each MANGLE polygon was assigned a weight depending on the meta data of the corresponding photometric tile. We rendered the maps on an HEALPIX mesh at $N_{\text{side}} = 2048$. This corresponds to a precision of ~ 2 arcminutes. As an illustration, we show a subset of four of the eleven foregrounds maps in Figure 3. These maps show, in Mollweide projection, the vector $F_{a,i}$ of Equation (8).

4 RESULTS

In this section, we detail various aspects of our results. In all the following, we have assumed a cosmology close to what the Planck collaboration has found with CMB data (Planck Collaboration et al. 2018b), namely $\Omega_r = 0$, $\Omega_K = 0$, $\Omega_M = 0.2889$, $\Omega_b = 0.048597$, $\Omega_\Lambda = 0.7111$, $w = -1$, $n_S = 0.9667$, $\sigma_8 = 0.8159$, $H_0 = 67.74 \text{ km s}^{-1} \text{ Mpc}^{-1}$. We note that the absolute value of the Hubble constant enters only through the transfer function in the power spectrum and does not have an impact on the coordinate transform for example. Early tests indicated that a slightly lower Ω_m seems to be preferred. We have thus pushed Ω_m to the lower acceptable limit and proceeded with the run. We have used an inference box of $4000 h^{-1} \text{ Mpc}$ sampled with a 256^3 mesh grid, setting resolution to $\sim 15 h^{-1} \text{ Mpc}$. The radial completeness is estimated in a similar way to Anderson et al. (2012).

The BORG inference machine is left free to choose the value of the 14 parameters of the bias model (model described in Section 2.3) and the amplitude of the 11 foreground templates, for each sub-catalogue (Section 4.4 and 3 for the description of sub-catalogues), and the amplitude of the 256^3 modes in the initial condition at $z = 1000$. We use a particle set of 512^3 to trace the evolution of matter with

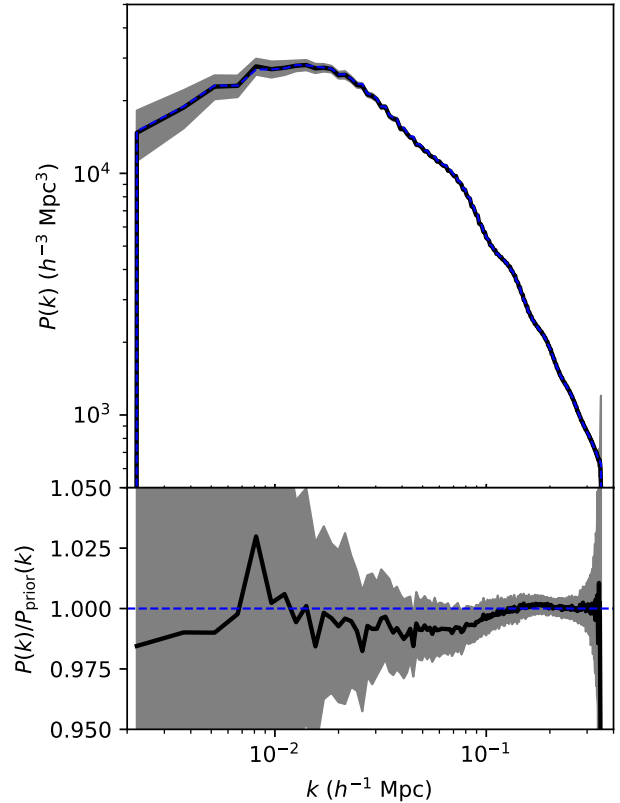


Figure 6. Ensemble average power spectrum of the *a posteriori* samples of initial conditions produced by BORG. *Top panel:* Mean of the power spectra (black line), one standard deviation (shaded grey region), and the prior power spectrum (dashed blue line). *Bottom panel:* same as the top panel but we divide by the prior power spectrum before plotting. The finest mode reachable in the box on the right-hand side is at $k \approx 0.35 h \text{ Mpc}^{-1}$ as can be seen by the strong increase of variance at that scale. It corresponds to a scale that already includes non-linear features. Deviations on large scales, up to the largest mode at $k \approx 10^{-3} h \text{ Mpc}^{-1}$, are below a few percents, confirming the systematic-free nature of the density reconstruction.

Lagrangian perturbation theory. We include both the light cone model of Section 2.2 and the model of redshift-space distortion effects at the level of particles. The bias parameters were given a prior centred on zero with a variance unity on elements of \mathbf{L} , in order to avoid the chain to explore too wide a parameter space initially and cause numerical issues. We have checked that this choice does not impact our result by verifying the *a posteriori* information content: the standard deviation of the sampled parameters are all less 0.7 which is less than 1. We have generated $\sim 10\,400$ MCMC samples from the posterior distribution. By an analysis of the *a posteriori* power spectrum of the initial density (Appendix D), we are confident that the chain has reasonably burned in after $\sim 2\,000$ samples. We use all the samples with an identifier greater than 2 000 for the results shown in this section.

³ https://data.sdss.org/sas/dr12/boos/lss/geometry/boos_geometry_2014_05_28.fits

4.1 Inferred 3D matter density fields

The main purpose of our BORG inference machine is to derive a probable, dynamical, physical, model of the matter distribution of the observed universe. As such, the first data product that we investigate is the inferred matter density that is required to explain the data. This comes in two forms: the initial conditions, post-recombination but still in the linear regime of the dynamics; and the evolved matter density at the moment the photon is detectable in our light cone. BORG samples all possible realisations of the initial conditions that satisfy the observational constraints. We focus our analysis mostly on the first moment of the posterior distribution for each considered parameter. However, more information is available on the posterior. In Figure 4, we show the ensemble average of all realisations of the evolved matter density. We represent a plane parallel to and close to the celestial equatorial plane ($\text{DEC}=0^\circ$), chosen to include the full shape of the light cone probed by the SDSS-III/BOSS. The full density field cube of Figure 4 will be made available on Zenodo at publication time. We recognise the typical structure of SDSS data: the south galactic cap part (SGC) at $Y > 0$, and the north galactic cap (NGC) at $Y < 0$. We see a separation between the LOWZ and CMASS components of SDSS-III/BOSS at a comoving distance $\sim 900h^{-1}$ Mpc. The distinction is more pronounced in the right panel, showing the standard deviation of fluctuations compared to the mean field: LOWZ clearly yields a noisier estimate of the matter density than CMASS. Despite the low resolution of our run ($\sim 15h^{-1}$ Mpc per voxel-side), we see that the density field is non-Gaussian, with some filamentary structure.

In Figure 5, we show the line-of-sight component of the velocity field, which we infer from the data and the dynamical model. The velocity field is derived using the simplex-in-cell estimator presented in Abel et al. (2012); Hahn et al. (2015); Leclercq et al. (2017) (based on the Delaunay tessellation of elementary Lagrangian cubes of particles into six tetrahedra), applied to the particle tracers generated by BORG for each element of the MCMC. We show the ensemble mean average (left panel) and the standard deviation with respect to the mean (right panel). We note that the light cone model also induces a modulation depending from the distance to the observer, located at the centre of the figure. In the right panel for example, the standard deviation of unobserved region is significantly higher close to the observer than in outer regions, as expected in perturbation theory, which predicts a factor ~ 2 difference between $z = 0.2$ and $z = 0.7$. Visual inspection of the maps does not show any particular anomaly. We note a slight excess of infall towards the observer in the lower-left region of the left panel, which could be due to an insufficient number of samples in the ensemble averaging of the Markov Chain.

We show in the top panel of Figure 6 the power spectrum of a *posteriori* initial conditions, both the ensemble average and the standard deviation. We also plot our prior on the cosmological power spectrum, obtained from the Eisenstein & Hu (1999) fitting function for our choice of cosmological parameters. In the bottom panel, we show the deviations of the *a posteriori* power spectrum from our prior. The prior is not strictly enforced in our inference framework, but only used as a guideline in the absence of informative data. We note that, contrary to previous attempts (e.g. Ross et al.

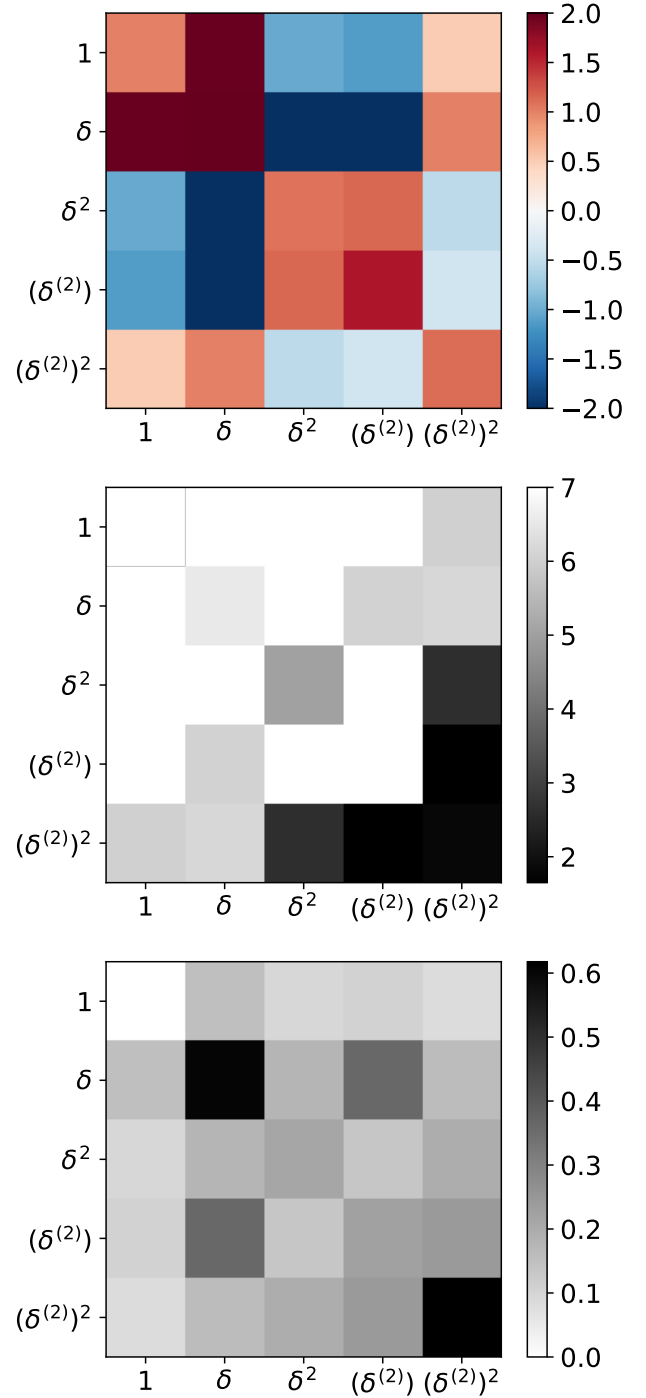


Figure 7. Bias parameters inferred for the CMASS NGC lowest redshift bin. The parameters are presented as the matrix \mathbf{Q} appearing in Equation (3). *Top panel:* matrix of mean parameters. *Middle panel:* ratio of the top matrix to the bottom matrix, showing the signal-to-noise matrix for each coefficient. *Bottom panel:* matrix of standard deviations for each coefficient. The inferred matrix represented here corresponds to a quadratic form: each coloured coefficient is associated with the product of the symbols indicated in the corresponding row and column in the bias model (see Equation (3)). The off-diagonal terms contribute twice while the on-diagonal terms contribute only once. The coefficient corresponding to 1×1 (top-left corner) is special as it is fixed to a value of unity and not sampled.

2012; Kalus et al. 2019), we do not observe strong deviations of the power spectrum at $k \lesssim 10^{-2} h \text{ Mpc}^{-1}$. Such deviations are typical of cases where systematic effects are improperly accounted for (Porqueres et al. 2019). If our model did not include systematic effects, the *a posteriori* power spectrum would have received contributions from them at the largest scales, as was noted by all previous attempts as well as our own investigations (e.g. Jasche & Lavaux 2017; Kalus et al. 2019; Porqueres et al. 2019). Given that past Cosmic Microwave Background missions (Bennett et al. 2013; Planck Collaboration et al. 2018b) have not observed any anomalous power at large angular scales deviating from the predictions of the Λ CDM model, we conclude the power spectrum deviations originally observed in SDSS-III/BOSS were due to systematic effects. The fact that these deviations vanish when using our framework for known and unknown systematic effects indicates that the presence of excess large-scale power is unnecessary to explain data. Furthermore, the posterior captures more information than what is in the prior, which is indicated by the shifted mean in the bottom panel of Figure 6.

4.2 The galaxy bias model

The BORG forward model includes a new quadratic-form bias, described in Section 2.3. As it is multi-scale, a direct comparison to previous analyses is not straightforward. In Figure 7, we give an example of the bias parameters that we have inferred, with corresponding uncertainties. These parameters correspond to the coefficients forming the \mathbf{Q} matrix for the sub-catalogue holding the CMASS NGC in a distance bin of $[1000, 1200] h^{-1} \text{ Mpc}$. This is thus the closest to a central slice in SDSS-III/BOSS. The matrix shown in Figure 7 gives the coefficient of the field produced by the product of its row and column labels in the bias model. For example, coefficients in the top row correspond to coupling the constant (“1”) with something else. If we choose the second column (labelled “ δ ”), we obtain the coefficient in the quadratic form that corresponds to the term $1 \times \delta$. In the second row and the last column, we find the coefficient of the term $\delta \times (\delta^{(2)})^2$. It is a quadratic form, thus off-diagonal terms shall be counted twice owed to the symmetry. In Figure 7, the top panel is the ensemble average for each coefficient, the bottom panel is the standard deviation with respect to that mean. The middle panel is the ratio of the top to the bottom panel, corresponding to the signal-to-noise ratio.

These bias parameters indicate that there is evidence for scale dependence, as the data require the model to have a non-vanishing second level (terms in $\delta^{(2)}$) to be represented fairly. We also note that it requires some compensation between scales as, for example, in the top row of the top panel of Figure 7. There, we have a positive coefficient in the second column (“ δ ”) followed by a negative coefficient in the fourth column (“ $\delta^{(2)}$ ”). There is thus some evidence of scale-dependent biasing. We leave further interpretation to future work.

As our model is non-local and non-linear (Equation 3), it is not straightforwardly related to classical linear biasing. However, assuming that the dark matter density is mostly the same at the two levels (1) and (2), the model has a linear term linking the matter density to galaxy number count

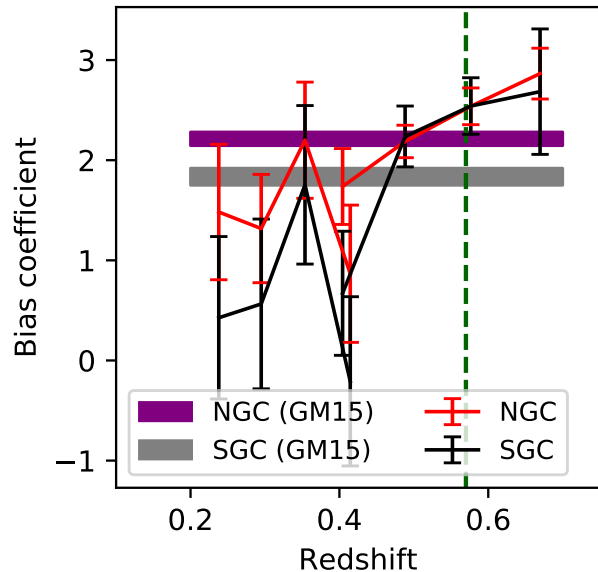


Figure 8. Evolution of the linear component of the bias model with redshift from the two hemispheres. We show here the quantity closest to linear biasing in our model, the sum $2(Q_{1,\delta} + Q_{1,\delta^{(2)}})$, as discussed in the text. In CMASS (at redshifts $0.4 \leq z \leq 0.7$), the evolution shows evolution with redshift, with a clear increase in the CMASS redshift range. There is little residual difference between data from the northern and southern hemisphere. The measurements of the linear bias coefficient b_1 from Gil-Marín et al. (2015, GM15) are shown by purple and grey bands, with the reference redshift indicated by a vertical dark green dashed line.

which behaves like $2(Q_{1,\delta} + Q_{1,\delta^{(2)}})$. In Figure 8, we show the evolution of this combination of the bias parameters of our model for the different sub-catalogues, organised in redshift bins. In LOWZ, there is no clear trend for this combination of bias parameters with redshift, but in CMASS, we observe that the equivalent of the linear bias evolves by nearly a factor of two between redshift 0.4 and 0.7. Additionally, there is no big discrepancy between the NGC and SGC part of CMASS, which indicates that the systematic effects between these two regions of the sky have likely been taken care of.

We can only partially compare our results to Gil-Marín et al. (2015) who provide the bias in very large bins, and without light cone correction. As we work with a fixed cosmological prior, we set their σ_8 to their reference maximum likelihood measurement and focus on the parameter b_1 . They report their measurement at an effective redshift $z_{\text{eff}} = 0.57$ (indicated by a vertical dashed line), corresponding to our second before last bin in Figure 8. Their reported b_1 value are indicated by two horizontal bands (purple for NGC, grey for SGC), which corresponds to the measurements reported in table 1 and 2 of Gil-Marín et al. (2015). Given the respective approximations involved, both measurements agree very well.

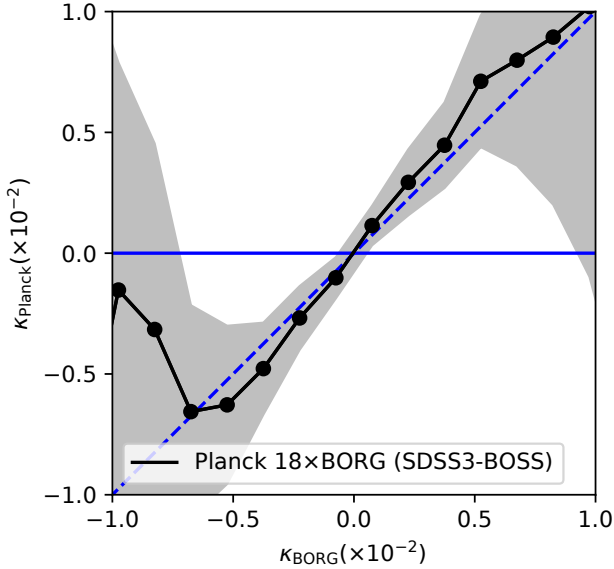


Figure 9. Correlation of the inferred lensing potential from SDSS-III/BOSS and the reconstructed lensing convergence from Planck temperature and polarisation maps. We show here the correlation, for each line-of-sight, between the value of the convergence that is obtained from the Planck lensing map derived in Planck Collaboration et al. (2018a) and the value that we compute from the gravitational potential derived from BORG. Following the procedure used in the Planck analysis, we restricted the modes to $\ell = 8 - 2048$ to derive the pixel based map. The blue dashed line indicates a perfect correlation. Further details are given in Section 4.3.

4.3 Cross analysis with CMB lensing

As presented above, the BORG algorithm provides very detailed reconstructions of the matter distribution over a cosmological volume. As demonstrated by the analysis of *a posteriori* power spectra in Section 4.1, inferred initial density fields follow the expected statistical properties at all Fourier modes considered in this work. This is owed to the robust treatment of unknown foreground effects which could otherwise introduce erroneous features at large scales in the matter distribution. In our previous work (Jasche & Lavaux 2019), we have demonstrated that BORG uses the physical forward modelling approach to perform dynamical mass estimates of galaxy clusters. We have found compatible mass profiles with the one derived from gold standard weak lensing measurements, X-ray observations or classical application of the virial theorem to galaxy velocity dispersion. To provide an independent test on whether BORG recovered the underlying large-scale dark matter field from SDSS-III/BOSS galaxy clustering data, we here perform a simple cross-correlation analysis with CMB weak lensing data provided by the Planck satellite mission. Achieving this goal requires to first generate posterior templates of weak lensing convergence maps from our inferred mass distributions. We use the classical expression to derive lensing convergence from the matter density fluctuation, which we reproduce

here:

$$\kappa(\hat{n}) = \frac{3}{2} \Omega_m \left(\frac{H_0}{c} \right)^2 \times \int_0^{\chi_{\text{CMB}}} \frac{d\chi}{a(\chi)} \frac{f_K(\chi) f_K(\chi_{\text{CMB}} - \chi)}{f_K(\chi_{\text{CMB}})} \delta_m(\chi, \hat{n}), \quad (16)$$

with Ω_m the matter density at redshift $z = 0$, $H_0 = 100 \text{ km s}^{-1} \text{ Mpc}^{-1}$, c the speed of light, χ the comoving distance, $f_K(\chi)$ the angular diameter distance, χ_{CMB} the comoving distance to the last scattering surface, \hat{n} the direction in which we observe the convergence. A schematic derivation is provided in Appendix C. Having only to rely on the local density fluctuation greatly simplifies the derivation of the convergence map by only taking integrals on lines of sight of the density contrast.

For our cross analysis we use publicly available CMB lensing convergence map⁴ obtained by the Planck satellite mission (Planck Collaboration et al. 2014, 2018a). We have checked that the results are consistent between the 2015 and 2018 maps. The result of the cross-analysis is shown in Figure 9. We show there a direct comparison, for each line of sight, of the convergence computed from the temperature and polarisation maps of the CMB sky observed by the Planck satellite and the one derived from the BORG analysis using Equation (16). Random samples of the observational noise for the Planck convergence has been taken into account, as well as the fluctuations allowed by the BORG posterior constrained by SDSS-III/BOSS data. The correlation procedure automatically cancels out the noise in the lensing map reconstructed from Planck mission data. The grey band is generated by the BORG posterior.

Other groups have already reported some correlation between the CMB lensing map obtained by the Planck collaboration and tracers of large-scale structures with Sloan Digital Sky Survey data. One of these test is provided in Singh et al. (2017). However, their comparison between large-scale structures and Planck CMB lensing is done at a much smaller scale than here by focusing on galaxy clusters. Their signal is typically vanishing starting from $\sim 10h^{-1} \text{ Mpc}$ from a galaxy, whereas in our case our voxels have a size of $\sim 16h^{-1} \text{ Mpc}$. This shows the future potential of an inferred density map such as the one we are providing, and that we have barely scratched the surface of the amount of available information. He et al. (2018) attempted a first detection of matter filaments using galaxies of SDSS-III/BOSS through the use of cross-correlation of the angular power spectrum. This detection is however done only at the level of correlation between cross-angular spectra. At larger distances, Han et al. (2019) found some evidence of correlation between the Quasar catalog from SDSS-IV and the same lensing map that we use. Thus we expect that a further extension of the present inference in the SDSS-IV regime would yield even better comparison.

A few other notable examples are the correlation with the CIB-WISE data (Yu et al. 2017), and similarly the correlation with the 2MASS-PhotoZ sample (Bianchini & Reichardt 2018). In these two cases, the sample either covers a larger fraction of the sky or has more galaxies and span

⁴ <http://pla.esac.esa.int/pla/>

a redshift range that is comparable to SDSS-III/BOSS. Additionally Yu et al. (2017) use the CIB contribution which peaks at much farther distances and provide a good template for the lensing convergence, which explains their very high correlation to the Planck lensing map. The WISE component that is used in that work is providing only $\sim 10\%$ of the correlation. Bianchini & Reichardt (2018) finds also some correlation although it is much weaker owed to the redshift distribution of the galaxies of the 2MPZ which is limited to $z \lesssim 0.2$.

The above agreement is showing that the mass distribution that BORG derives from SDSS-III/BOSS is supported by both an independent data-set and an independent physical effect that measure the same quantity. A detailed analysis of CMB-Large scale structure cross analysis will be presented in a forthcoming publication.

4.4 Mean inferred systematic properties

In this section, we discuss our results concerning the contamination of the SDSS-III/BOSS sample with known and unknown systematic effects. We remind the reader that we use two techniques at the same time for taking into account these effects, leading to a clean reconstruction of the matter density field: the template based approach (also known as ‘extended mode’ projection, see Leistedt & Peiris 2014) and the robust likelihood (closer to ‘basic mode’ projection, see Porqueres et al. 2019).

In Figure 10, we show the mean and standard deviation for each individual foreground coefficients multiplying the indicated templates independently at different redshift and for the NGC and SGC side. We note that for a large fraction of these coefficients, no signal is really detectable with our robust likelihood. For example, the dust contamination is completely flat and compatible with zero. However some of these coefficients exhibit positive, redshift dependent, signal. That is the case for the point spread function (PSF) in the r band (top row, middle panel). In this case there is a clear difference between NGC and SGC as well. Another template that has clear correlation with data is the skyflux in the u band (third row, left panel). There is a monotonic increase in the contamination level of the SDSS-III/BOSS data in the NGC, while SGC seems to be more immune. We have chosen to use the same choice of foreground templates as the one studied by Ross et al. (2012), notably the slices of star density. Though our results are not directly comparable owed to the different procedure to analyse the data, Figure 11 of Ross et al. (2012) is the most evocative. Generally speaking, this other study showed that the CMASS sample is more contaminated than the LOWZ sample. The star density and the seeing/PSF were among the top contaminant. Here we clearly have evidence of this in the subplot labelled “psf_r” and “star_0”. The effect of sky brightness seems larger in our analysis than in the original SDSS analysis. We note that we used the weights provided by the SDSS-III/BOSS collaboration to correct our sample of galaxies before doing the inference, thus some of these contaminations have already been compensated. Our plots may be understood as additional residual contamination that were not accounted for in the galaxy weighting of SDSS-III/BOSS. To conclude this discussion, these results highlight the power of our in-

ference method to detect and correct defects in the data acquisition.

In Figures 11 and 12, we show the correlation coefficients between these same foreground coefficients and the amplitude of modes at different scales. In both figures, we have ordered in increasing redshift from left-to-right, LOWZ being in the top rows and CMASS in the bottom rows. The foregrounds templates are indicated on the y -axis and the scale on the x -axis. We see that despite seeing in most cases a null detection in Figure 10, the amplitudes of coefficients tend to correlate heavily with a lot of modes in the reconstructed initial conditions. This correlation is not stable with redshift nor with scales. The most notable example is the skyflux in the u band: it is mostly positively correlated with density Fourier modes up to $k \sim 0.15h \text{ Mpc}^{-1}$ and then becomes negatively correlated at higher k , for a lot of sub-catalogues. The influence of the airmass in the r band is also an example of contaminant that changes significantly with redshift in both LOWZ and CMASS. Other foregrounds give an impact that is more focused either spatially or in redshift.

Our Bayesian inference approach has another seducing aspect: unknown foreground contamination may also be reconstructed from posterior samples of the Markov chain, as discussed in Section 2.6. This possibility was already mentioned by Monaco et al. (2018) for Euclid-like surveys. But it is already possible for the SDSS-III/BOSS sample of galaxies. Here we have used equations (12) and (13) to estimate the ensemble mean and signal-to-noise ratio maps for the four redshift bins of the LOWZ and CMASS samples of SDSS-III/BOSS. The corresponding inference results for unknown foreground contamination are presented in Figures 13 and 14. As can be seen, these maps clearly contain spatial structure despite having marginalised over already 11 foregrounds per sub-catalogue. The systematic maps show clear iso-declination striping, which does not seem to follow the drift-scan strategy of the SDSS photometry. The drift-scan strategy is visible for example in Figure 3. These stripe modulate the signal at the level of 30% of multiplicative correction on the sky. It is most prominent for the highest signal-to-noise redshift bin at $z = 0.29$ and $z = 0.35$ for LOWZ and $z = 0.49$ and $z = 0.58$ for CMASS.

While the striping structure at each redshift bin are fairly represented by some pattern of iso-declination modulation for both NGC and SGC, the pattern itself between the two north and south caps look different. For example in Figure 13 (LOWZ), at $z = 0.35$, there is a clear wide blue-stripe ($\sim -30\%$ correction) at $\text{DEC} = +30^\circ$ in the SGC, while it is reddish ($\sim +30\%$) for the NGC. It is not strictly iso-declination all the time either. For example in LOWZ, at $z = 0.29$ the red stripe at $\text{DEC} = +30^\circ$ is widening towards lower DEC while going from left to right of the NGC. Finally, the stripes are not constant with redshift, sometimes inverting completely. That is the case between the two redshift bins of CMASS at $z = 0.49$ and $z = 0.58$ for which the large stripe just above $\text{DEC} = +30^\circ$ is blue in the first case, and red in the second case.

The plausible origin of these systematic effects is likely to be on the ground given the distribution of the stripes. One of such problems are the “contrails” (Finkbeiner et al. 2016). Understanding the detail of the origin of these systematic effects is however beyond the scope of this work.

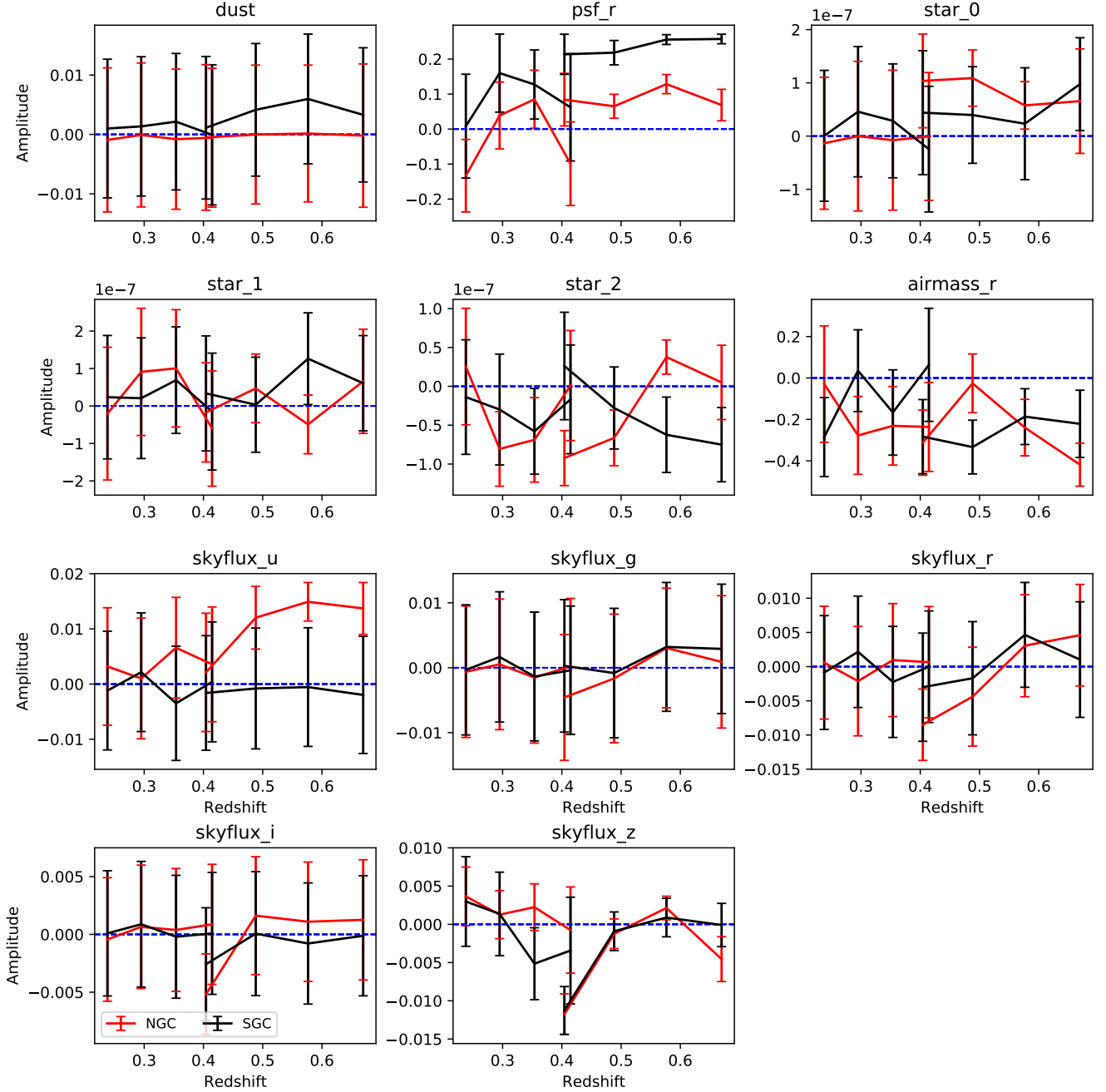


Figure 10. Foreground template coefficients for NGC and SGC as a function of redshift. We show the inferred mean values alongside their standard deviation for each of the 11 foreground templates assigned to each sub-catalogue in our BORG run on SDSS-III/BOSS. As for other figures, the LOWZ component is at redshift $z \leq 0.4$ and the CMASS component is at $z > 0.4$.

The systematic maps will be made available for download on Zenodo after publication.

5 CONCLUSION

With the advent of next-generation galaxy surveys, cosmic large-scale structures will become one of the most important cosmological probes to test the fundamental physics governing the dynamics of our Universe. To ensure contin-

ued scientific progress in cosmology, the acquisition of novel quality data needs to be accompanied by the development of novel methods capable of handling unknown systematic effects and to link complex non-linear structure growth physics with observations. Such model of large scale structures as the one we have derived have many applications for the study and observation of the Universe through different instruments. Some of the applications that were considered in the past are cosmic-web identifications and characterization (Leclercq et al. 2015b, 2017), cosmic voids properties

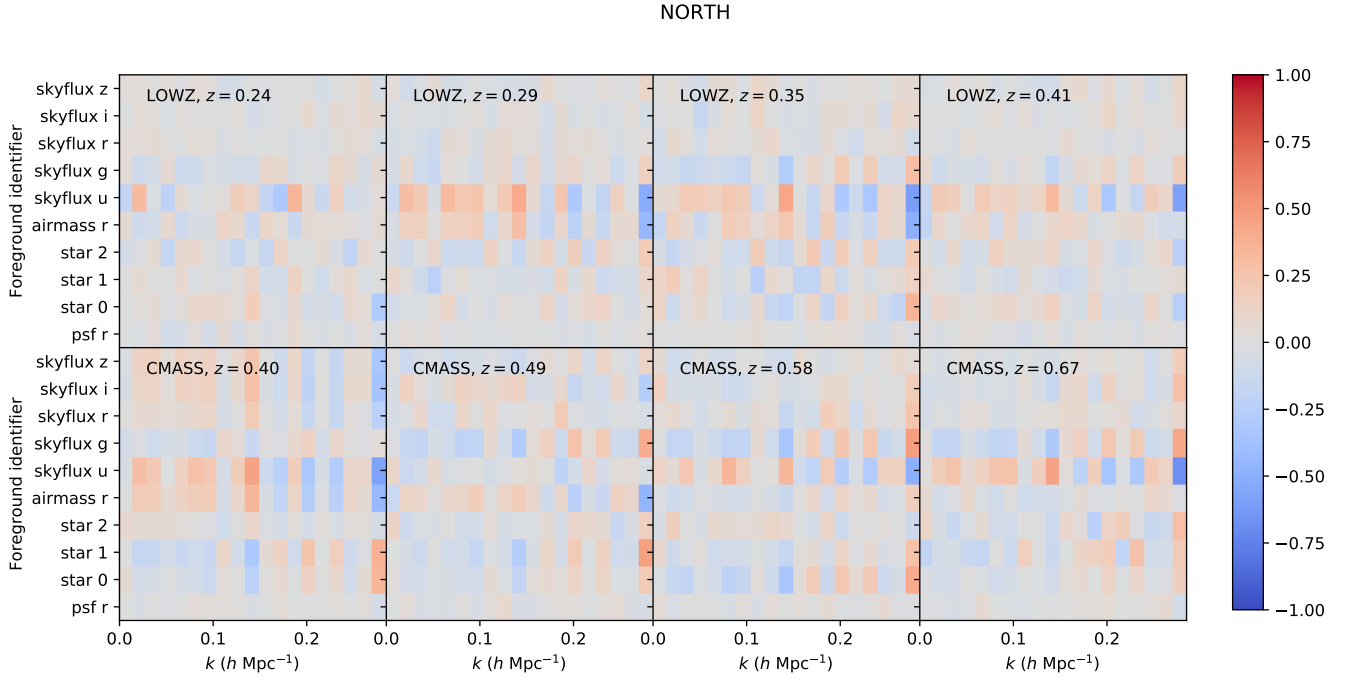


Figure 11. Correlation between foreground template coefficients and the Fourier modes sampled by BORG for the NGC sub-samples. The top row contains the correlation matrix for the LOWZ NGC sub-samples ordered from low (*left panel*) to high (*right panel*) redshift. The second row shows the same quantity but for the CMASS NGC sub-samples.

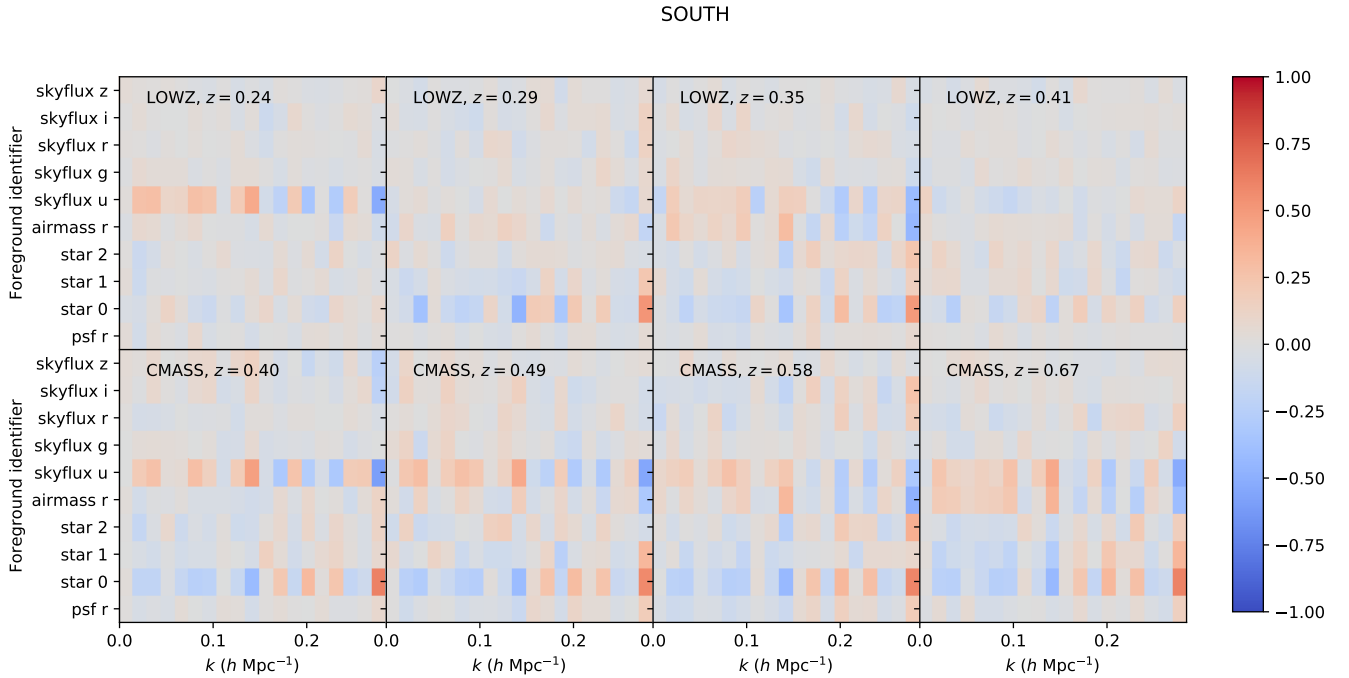


Figure 12. Same as Figure 11 but for the SGC sub-samples.

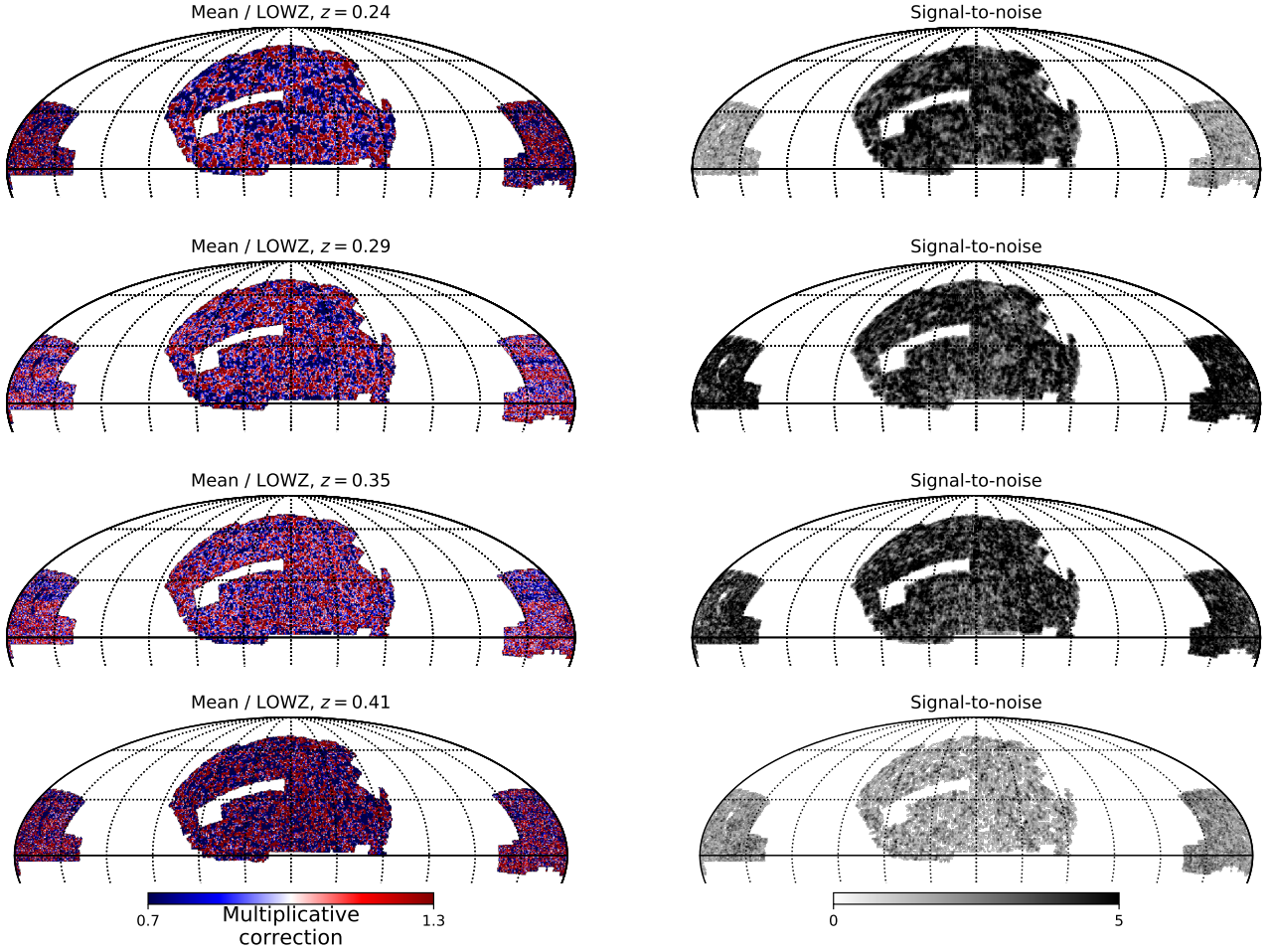


Figure 13. Inferred mean systematic maps (*left column*) as well as the estimated corresponding signal-to-noise for each pixel in these maps (*right panel*), as derived from SDSS-III/BOSS data, in equatorial coordinates. Each of these maps is estimated using Equations (12) and (13), independently for each of the redshift bins. The maps are multiplicative, we clearly note a correlated modulation of order 30% on the sky, indicative of unknown residual systematic effects in the data.

(Leclercq et al. 2015a), cosmic magnetic fields (Hutschenreuter et al. 2018), constraints on fifth-force gravity models (Desmond et al. 2018a,b, 2019), peculiar velocity corrections to Hubble-Lemaître constant deduced from standard sirens (Mukherjee et al. 2019).

While traditional methods focus only on analysing a limited number of low-order statistics of the matter distribution, here we apply a fully Bayesian physical forward modelling approach to extract the significant information entailed in the high-order statistics associated to the filamentary matter distribution underlying the galaxies in surveys.

Specifically, we presented a fully Bayesian analysis of the spatial matter distribution probed by SDSS-III/BOSS data. As described in this work, our method infers physically plausible reconstructions from the data while accounting for systematic effects, such as galaxy biases, light-cone effects, survey geometries and other selection effects. Most notably, we demonstrate the application of a novel robust likelihood approach to data, required to deal with unknown systematic effects in the data, which otherwise would result in the

erroneous reconstruction of the large-scale matter distribution and corresponding velocity fields, posing significant nuisances for cosmological interpretation of observations.

We conducted an analysis of SDSS-III/BOSS data to recover the cosmic large-scale structure within a Cartesian co-moving volume of $4\,000h^{-1}$ Mpc at a resolution of $\sim 15.6h^{-1}$ Mpc. Our analysis simultaneously accounts for data in the southern and northern galactic cap of SDSS-III/BOSS. We carefully accounted for non-linear scale dependencies in galaxy biases and data selection effects by splitting the data into galaxy sub-samples of eight redshift bins, nearly equidistant. For each of these galaxy samples, we treated respective systematic effects separately. To model possible non-linear and non-local effects of the galaxy bias, we proposed a novel multi-power galaxy biasing model, which uses the information of the density field at two different levels of resolution, resulting in a fourteen parameter model per galaxy sub-sample. We determined corresponding bias parameters for each of the galaxy sub-samples, to account for possible redshift evolution. In addition, for each of these galaxy sub-samples, we accounted for survey ge-

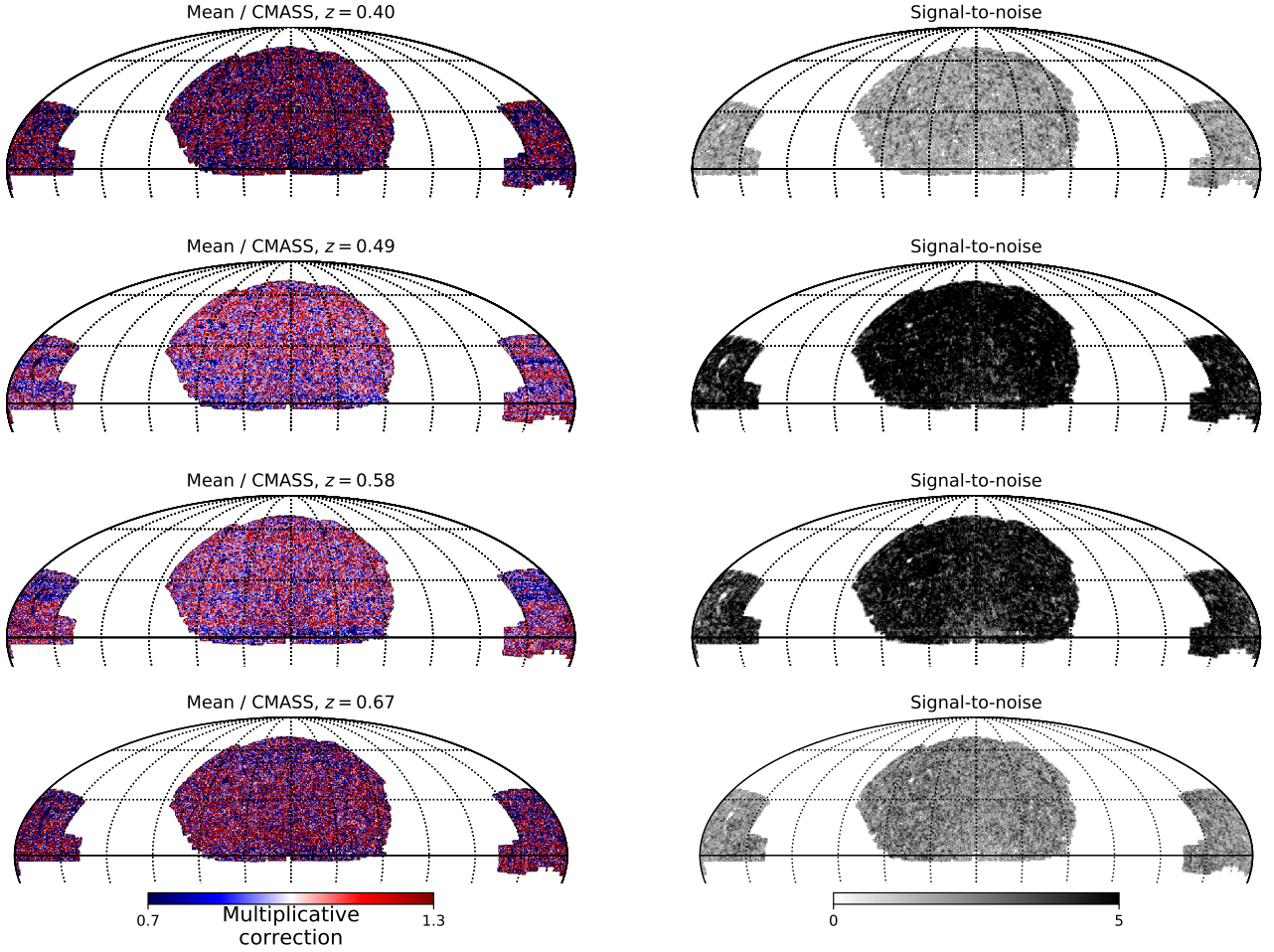


Figure 14. Same as Fig 13 but for the CMASS sample.

ometry, and we self-consistently inferred the amplitudes of eleven known foreground templates as well as the unknown noise levels of the galaxy samples. Besides fitting known foreground contributions, a significant improvement over previous work is that our approach uses a robust likelihood approach to also account for unknown systematic effects affecting the survey. As demonstrated in this work, the detailed handling of unknown systematics in galaxy surveys is crucial to infer cosmologically significant and unbiased information from the largest scales in present and coming galaxy surveys. To confirm the reality of the large-scale dynamics that we recovered, we checked the correlation with lensing measurements obtained from the data of the Planck mission. The near-perfect alignment between the prediction that we derived from SDSS-III/BOSS and Planck lensing provides solid evidence that the inferred dark matter density field is correct in the volume spanned by SDSS-III/BOSS.

In summary, the combination of a Bayesian physical forward modelling approach with a robust likelihood approach to account for unknown systematic effects in data is a successful approach to characterise the cosmic large-scale structure and its dynamic formation. The presented work, therefore, defines a promising path towards a fully physically meaningful analysis of next-generation galaxy surveys.

ACKNOWLEDGEMENTS

We thank Fabian Schmidt, Benjamin Wandelt, David Weinberg, François Bouchet, Stéphane Colombi, Valérie de Laparent, Matthew Lehnert, Suvodip Mukherjee, Peter H. Johansson for useful discussions. This work has been done within the activities of the Domaine d’Intérêt Majeur (DIM) “Astrophysique et Conditions d’Apparition de la Vie” (ACAV), and received financial support from Région Ile-de-France. GL acknowledges financial support from the ILP LABEX, under reference ANR-10-LABX-63, which is financed by French state funds managed by the ANR within the programme “Investissements d’Avenir” under reference ANR-11-IDEX-0004-02. GL also acknowledges financial support from the ANR BIG4, under reference ANR-16-CE23-0002. FL acknowledges funding from the Imperial College London Research Fellowship Scheme. This work was granted access to the HPC resources of CINES (Centre Informatique National de l’Enseignement Supérieur) under the allocation A0020410153 and A0040410153 made by GENCI and has made use of the Horizon cluster hosted by the Institut d’Astrophysique de Paris on which the cosmological simulations were post-processed. GL thanks the hospitality of the

University of Helsinki where part of this work took place. This work is done within the Aquila Consortium⁵.

REFERENCES

- Abazajian K., Survey f. t. S. D. S., 2009, *The Astrophysical Journal Supplement Series*, 182, 543
- Abel T., Hahn O., Kaehler R., 2012, *Monthly Notices of the Royal Astronomical Society*, 427, 61
- Alam S., et al., 2015, *The Astrophysical Journal Supplement Series*, 219, 12
- Anderson L., et al., 2012, *Monthly Notices of the Royal Astronomical Society*, 427, 3435
- Anderson L., et al., 2014, *Monthly Notices of the Royal Astronomical Society*, 441, 24
- Bennett C. L., et al., 2013, *The Astrophysical Journal Supplement Series*, 208, 20
- Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, *Physics Reports*, 367, 1
- Beutler F., et al., 2017, *Monthly Notices of the Royal Astronomical Society*, 464, 3409
- Bianchini F., Reichardt C. L., 2018, *The Astrophysical Journal*, 862, 81
- Bouchet F. R., Colombi S., Hivon E., Juszkiewicz R., 1995, *Astronomy & Astrophysics*, 296, 575
- Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, *Monthly Notices of the Royal Astronomical Society*, 446, 2621
- Colavincenzo M., Monaco P., Sefusatti E., Borgani S., 2017, *Journal of Cosmology and Astroparticle Physics*, 2017, 052
- Dawson K. S., et al., 2013, *The Astronomical Journal*, 145, 10
- Desjacques V., Jeong D., Schmidt F., 2018, *Physics Reports*, 733, 1
- Desmond H., Ferreira P. G., Lavaux G., Jasche J., 2018a, *Physical Review D*, 98
- Desmond H., Ferreira P. G., Lavaux G., Jasche J., 2018b, *Physical Review D*, 98
- Desmond H., Ferreira P. G., Lavaux G., Jasche J., 2019, *Monthly Notices of the Royal Astronomical Society: Letters*, 483, L64
- Eisenstein D. J., Hu W., 1999, *The Astrophysical Journal*, 511, 5
- Eisenstein D. J., et al., 2011, *The Astronomical Journal*, 142, 72
- Elsner F., Leistedt B., Peiris H. V., 2016, *Monthly Notices of the Royal Astronomical Society*, 456, 2095
- Elsner F., Leistedt B., Peiris H. V., 2017, *Monthly Notices of the Royal Astronomical Society*, 465, 1847
- Elsner F., Schmidt F., Jasche J., Lavaux G., Nguyen N.-M., 2019, arXiv:1906.07143 [astro-ph]
- Finkbeiner D. P., et al., 2016, *The Astrophysical Journal*, 822, 66
- Fowler J. W., et al., 2010, *The Astrophysical Journal*, 722, 1148
- Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, *The Astronomical Journal*, 111, 1748
- Gil-Marín H., Noreña J., Verde L., Percival W. J., Wagner C., Manera M., Schneider D. P., 2015, *Monthly Notices of the Royal Astronomical Society*, 451, 539
- Gorski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelman M., 2005, *The Astrophysical Journal*, 622, 759
- Hahn O., Angulo R. E., Abel T., 2015, *Monthly Notices of the Royal Astronomical Society*, 454, 3920
- Han J., Ferraro S., Giusarma E., Ho S., 2019, *Monthly Notices of the Royal Astronomical Society*, 485, 1720
- He S., Alam S., Ferraro S., Chen Y.-C., Ho S., 2018, *Nature Astronomy*, 2, 401
- Ho S., et al., 2012, *The Astrophysical Journal*, 761, 14
- Hutschenreuter S., Dorn S., Jasche J., Vazza F., Paoletti D., Lavaux G., Enßlin T. A., 2018, *Classical and Quantum Gravity*, 35, 154001
- Jasche J., Lavaux G., 2017, *Astronomy & Astrophysics*, 606, A37
- Jasche J., Lavaux G., 2019, *Astronomy & Astrophysics*, 625, A64
- Jasche J., Wandelt B. D., 2013, *Monthly Notices of the Royal Astronomical Society*, 432, 894
- Jasche J., Kitaura F. S., Li C., Enßlin T. A., 2010, *Monthly Notices of the Royal Astronomical Society*, 409, 355
- Jasche J., Leclercq F., Wandelt B. D., 2015, *Journal of Cosmology and Astroparticle Physics*, 2015, 036
- Kaiser N., 1984, *The Astrophysical Journal*, 284, L9
- Kaiser N., 1998, *The Astrophysical Journal*, 498, 26
- Kalus B., Percival W. J., Bacon D. J., Mueller E.-M., Samushia L., Verde L., Ross A. J., Bernal J. L., 2019, *Monthly Notices of the Royal Astronomical Society*, 482, 453
- Kilbinger M., 2015, *Reports on Progress in Physics*, 78, 086901
- Kitaura F.-S., et al., 2016, *Monthly Notices of the Royal Astronomical Society*, 456, 4156
- LSST Science Collaboration et al., 2009, arXiv:0912.0201 [astro-ph]
- Laureijs R., et al., 2011, arXiv:1110.3193 [astro-ph]
- Lavaux G., Hudson M. J., 2011, *Monthly Notices of the Royal Astronomical Society*, 416, 2840
- Lavaux G., Jasche J., 2016, *Monthly Notices of the Royal Astronomical Society*, 455, 3169
- Leclercq F., Jasche J., Sutter P. M., Hamaus N., Wandelt B., 2015a, *Journal of Cosmology and Astroparticle Physics*, 2015, 47
- Leclercq F., Jasche J., Wandelt B., 2015b, *Journal of Cosmology and Astro-Particle Physics*, 2015, 15
- Leclercq F., Jasche J., Lavaux G., Wandelt B., Percival W., 2017, *Journal of Cosmology and Astroparticle Physics*, 2017, 049
- Leclercq F., Enzi W., Jasche J., Heavens A., 2019, arXiv:1902.10149 [astro-ph]
- Leistedt B., Peiris H. V., 2014, *Monthly Notices of the Royal Astronomical Society*, 444, 2
- Lewis A., Challinor A., 2006, *Physics Reports*, 429, 1
- Maraston C., Stromback G., Thomas D., Wake D. A., Nichol R. C., 2009, *Monthly Notices of the Royal Astronomical Society: Letters*, 394, L107
- Monaco P., Di Dio E., Sefusatti E., 2018, arXiv:1812.02104 [astro-ph]
- Mukherjee S., Lavaux G., Bouchet F. R., Wandelt B. D., Nissanke S., Leclercq F., Jasche J., Hotokozaka K., 2019, submitted
- Parejko J. K., et al., 2013, *Monthly Notices of the Royal Astronomical Society*, 429, 98
- Planck Collaboration et al., 2014, *Astronomy & Astrophysics*, 571, A1
- Planck Collaboration et al., 2018b, arXiv:1807.06205 [astro-ph]
- Planck Collaboration et al., 2018a, arXiv:1807.06210 [astro-ph]
- Planck Collaboration et al., 2019, arXiv:1905.05697 [astro-ph, physics:gr-qc, physics:hep-ph, physics:hep-th]
- Porqueres N., Ramanah D. K., Jasche J., Lavaux G., 2019, *Astronomy & Astrophysics*, 624, A115
- Ramanah D. K., Lavaux G., Jasche J., Wandelt B. D., 2019, *Astronomy & Astrophysics*, 621, A69
- Reid B., et al., 2016, *Monthly Notices of the Royal Astronomical Society*, 455, 1553
- Ross A. J., et al., 2011, *Monthly Notices of the Royal Astronomical Society*, 417, 1350
- Ross A. J., et al., 2012, *Monthly Notices of the Royal Astronomical Society*, 424, 564
- Ross A. J., et al., 2017, *Monthly Notices of the Royal Astronomical Society*, 464, 1168
- Rybicki G. B., Press W. H., 1992, *The Astrophysical Journal*, 398, 169
- Satpathy S., et al., 2017, *Monthly Notices of the Royal Astro-*

⁵ <https://www.aquila-consortium.org/>

- nomical Society, 469, 1369
- Schaffer K. K., et al., 2011, *The Astrophysical Journal*, 743, 90
- Schlegel D. J., Finkbeiner D. P., Davis M., 1998, *The Astrophysical Journal*, 500, 525
- Schmidt F., Elsner F., Jasche J., Nguyen N. M., Lavaux G., 2019, *Journal of Cosmology and Astroparticle Physics*, 2019, 042
- Singh S., Mandelbaum R., Brownstein J. R., 2017, *Monthly Notices of the Royal Astronomical Society*, 464, 2120
- Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, *Monthly Notices of the Royal Astronomical Society*, 387, 1391
- Tegmark M., 1997, *Physical Review D*, 55, 5895
- Tweed D., Yang X., Wang H., Cui W., Zhang Y., Li S., Jing Y. P., Mo H. J., 2017, *The Astrophysical Journal*, 841, 55
- Wang H., Mo H. J., Yang X., Jing Y. P., Lin W. P., 2014, *The Astrophysical Journal*, 794, 94
- Wang H., et al., 2016, *The Astrophysical Journal*, 831, 164
- Yu B., Hill J. C., Sherwin B. D., 2017, *Physical Review D*, 96
- Zel'Dovich Y., 1970, *Astronomy & Astrophysics*, 5, 84

APPENDIX A: VARIANCE OF THE DISPLACEMENT FIELD

In a Λ CDM universe, assuming that evolution of large-scale structures is well described by the Zel'Dovich approximation (Zel'Dovich 1970), the statistics of the displacement is simple. Using the continuity equation, we can write

$$\nabla_{\mathbf{q}} \cdot \Psi = -D(t)\delta(\mathbf{q}), \quad (\text{A1})$$

with \mathbf{q} the Lagrangian coordinates, which at high redshift are close to the Eulerian coordinates, $D(t)$ the growth function, Ψ the displacement field. The one-point variance of the displacement field becomes thus

$$\begin{aligned} \langle \Psi_a^2(t, \mathbf{q}) \rangle &= \int \frac{d^3 \mathbf{k} d^3 \mathbf{k}'}{(2\pi)^6} e^{i(\mathbf{k}+\mathbf{k}') \cdot \mathbf{q}} \langle \hat{\Psi}_a(\mathbf{k}) \hat{\Psi}_a(\mathbf{k}') \rangle \\ &= D^2(t) \int \frac{d^3 \mathbf{k} d^3 \mathbf{k}'}{(2\pi)^6} \frac{-k_a k'_a}{|\mathbf{k}|^2 |\mathbf{k}'|^2} e^{i(\mathbf{k}+\mathbf{k}') \cdot \mathbf{q}} \langle \hat{\delta}(\mathbf{k}) \hat{\delta}(\mathbf{k}') \rangle \\ &= D^2(t) \int \frac{d^3 \mathbf{k}}{(2\pi)^3} P(k) \frac{k_a^2}{k^4} \\ &= \frac{D^2(t)}{3} \int \frac{d^3 \mathbf{k}}{(2\pi)^3} \frac{1}{k^2} P(k) \end{aligned} \quad (\text{A2})$$

$$= \frac{D^2(t)}{12\pi^2} \int_{k=0}^{+\infty} dk P(k), \quad (\text{A3})$$

with $P(k)$ the power spectrum of matter density fluctuations at high redshift. For a Λ CDM universe, with Planck 2018 cosmology, the square root of that variance is $5.96 h^{-1}$ Mpc. An acceptable typical upper bound to the displacement field may be at ~ 3 times that value, which leads to $17.9 h^{-1}$ Mpc.

APPENDIX B: ADJOINT GRADIENT OF THE BIAS MODEL

Computing of the adjoint gradient, or back-propagation in machine learning terminology, consists in linearly transforming an error vector back to the adequate parameter space of interest. In BORG, that consists in transporting the error vector from the likelihood space, which touches galaxy distribution, to the initial condition. The bias model step relate

the matter density to the expected galaxy distribution, before the effect of the pipeline of detection by the instrument. We assume that we are provided an error vector v_i , per mesh element. The new error vector \tilde{v}_q will be derived as follow:

$$\tilde{v}_q = \sum_i v_i \frac{\partial N_i^{(g)}}{\partial \delta_q} = 2 \sum_i v_i \frac{\partial \Delta_i}{\partial \delta_q} \mathbf{Q} \Delta_i. \quad (\text{B1})$$

In general the element of the vector Δ_i take the following form

$$(\Delta_i)_a = (\delta_i^{(\ell_a)})^{\gamma_a}, \quad (\text{B2})$$

with j_a the density averaging level at the component a and γ_a the power rising of the component a . The detail of that ordering is given in Section 2.3. The special case $j_a = 0$ corresponding to $\delta^{(0)} = 1$. Thus we may derive the derivative of the vector Δ_i by looking at each component:

$$\frac{\partial \Delta_{i,a}}{\partial \delta_q} = \frac{\partial \delta_i^{(j_a)}}{\partial \delta_q} \times \begin{cases} \gamma_a \left(\delta_i^{(j_a)} \right)^{\gamma_a - 1}, & \text{if } \gamma_a \geq 1, j_a \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{B3})$$

$$= \frac{\partial \delta_i^{(j_a)}}{\partial \delta_q} g_a \left(\delta_i^{(j_a)} \right) \quad (\text{B4})$$

Finally the derivative of the averaging operator is

$$\frac{\partial \delta_i^{(\ell_a)}}{\partial \delta_q} = \frac{1}{8^\ell} \times \begin{cases} 1 & \text{if } q \in \mathcal{V}_i^{(j_a)}, \\ 0 & \text{otherwise,} \end{cases} \quad (\text{B5})$$

with $\mathcal{V}_i^{(j_a)}$ the vicinity set of i at the level j_a of the oct-tree. This vicinity set is defined implicitly from Equation (7). We may compute it explicitly by doing the matrix-vector multiplication with the vector \mathbf{v} :

$$\tilde{v}_q = \frac{1}{8^\ell} \sum_{\alpha, \beta} \sum_{a, b, c=0}^{2^{\ell-1}} v_{f_i(q, a, b, c)} \times g_\alpha \left(\delta_{f_i(q, a, b, c)}^{(\ell_a)} \right) Q_{\alpha, \beta} \left(\delta_{f_i(q, a, b, c)}^{(\ell_a)} \right)^{\gamma_\alpha}. \quad (\text{B6})$$

This gives an explicit algorithm to compute the adjoint-gradient with this new bias model.

APPENDIX C: LENSING EQUATION

In this appendix we give a brief reminder of the derivation of Equation (16). If we consider the Newtonian potential Ψ defined at comoving distances χ and angular direction $\hat{\mathbf{n}}$ on the sky (Kaiser 1998; Lewis & Challinor 2006; Kilbinger 2015), then the sky displacement of one photon, at first order of perturbation in Ψ and on the geodesic trajectory followed by that photon is:

$$\alpha(\hat{\mathbf{n}}) = \frac{1}{c^2} \times \int_0^{\chi_{\text{CMB}}} d\chi \frac{f_{\mathbf{K}}(\chi_{\text{CMB}} - \chi)}{f_{\mathbf{K}}(\chi_{\text{CMB}}) f_{\mathbf{K}}(\chi)} (\nabla_{\hat{\mathbf{n}}} \Psi)(\chi \hat{\mathbf{n}}; \chi_{\text{CMB}} - \chi) \quad (\text{C1})$$

where χ is the comoving radial distance and

$$f_{\mathbf{K}}(\chi) = \begin{cases} \sin(\chi) & \text{for } K = +1, \text{ closed universe;} \\ \chi & \text{for } K = 0, \text{ flat universe;} \\ \sinh(\chi) & \text{for } K = -1, \text{ open universe.} \end{cases} \quad (\text{C2})$$

The convergence is defined as the sky divergence of the sky displacement:

$$\kappa(\hat{\mathbf{n}}) = \nabla_{\hat{\mathbf{n}}} \alpha(\hat{\mathbf{n}}), \quad (\text{C3})$$

Furthermore, the three-dimensional potential Ψ is related to the matter density contrast $\delta_m(\mathbf{x})$ via the Poisson equation in comoving coordinates,

$$\nabla_{\mathbf{x}}^2 \Psi = \frac{3}{2a(\chi)} \Omega_m(\chi) H_0^2 \delta_m(\mathbf{x}; \chi). \quad (\text{C4})$$

The above equation is valid in the usual perturbative regime of the metric, which is the case for the entirety of this work. By moving the divergence inside the integral, we obtain

$$\kappa(\hat{\mathbf{n}}) = -\frac{1}{c^2} \times \int_0^{\chi_{\text{CMB}}} d\chi \frac{f_K(\chi_{\text{CMB}} - \chi) f_K(\chi)}{f_K(\chi_{\text{CMB}})} \times \left(\nabla_{\hat{\mathbf{n}}}^2 \Psi(f_K(\chi) \hat{\mathbf{n}}; \chi_{\text{CMB}} - \chi) \right), \quad (\text{C5})$$

As generally done in the scientific literature and explicitly justified in Kilbinger (2015), we replace the 2D Laplacian by the 3D Laplacian because we expect the second-order radial derivatives to average to zero at the scale that we consider. Thus we have a simplified expression for the convergence

$$\kappa(\hat{\mathbf{n}}) = \frac{3}{2} \Omega_m \left(\frac{H_0}{c} \right)^2 \times \int_0^{\chi_{\text{CMB}}} d\chi \frac{f_K(\chi) f_K(\chi_{\text{CMB}} - \chi)}{f_K(\chi_{\text{CMB}})} \frac{\delta_m(\chi, \hat{\mathbf{n}})}{a(\chi)} \quad (\text{C6})$$

This greatly simplifies the derivation of the convergence map by only taking integrals on line of sights of the density contrast.

APPENDIX D: TESTING THE WARM-UP PHASE OF THE SAMPLER

As described in our previous works (Jasche & Wandelt 2013; Lavaux & Jasche 2016; Jasche & Lavaux 2019), we initialize the Markov chain with an over-dispersed state, that is far remote from the target regions in the parameter space. This permits us to test the sampler behaviour during the initial warm-up phase and confirm it has approached the stationary regime before starting to record Markov samples for the analysis. Over-dispersed initial states are prepared by initialising the Markov chain with a random Gaussian initial density field scaled by a factor 1/10, which translates to 1/100 in Figure D1. To follow the sampler behaviour during its warm-up phase, we follow the traces of posterior power spectrum amplitudes throughout the initial sampler steps. As demonstrated by Figure D1, initially power spectrum amplitudes at the different modes of Fourier-space perform a coherent drift towards preferred regions in parameter space. After about 1,000 Markov transition steps the chain has reached a stationary distribution and power spectrum amplitudes oscillate around their expected fiducial values. From that moment, we start recording samples from the stationary distribution to perform the analysis presented in this work.

This paper has been typeset from a $\text{T}_{\text{E}}\text{X}/\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ file prepared by the author.

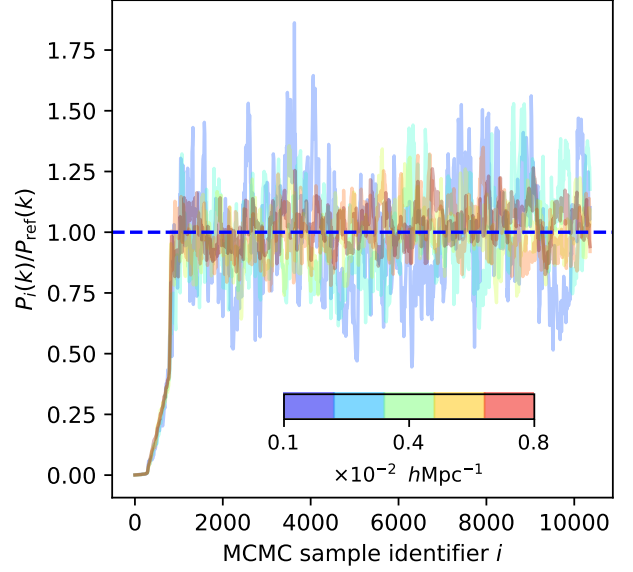


Figure D1. Amplitudes of the *a posteriori* primordial matter power spectrum at different Fourier modes traced during the warm-up phase of the MCMC sampler. As can be seen, initially, modes perform a coherent drift towards the high probability region in posterior distribution and start oscillating around their fiducial values once the Markov chain has reached a stationary state.