Bayesian physical reconstruction of initial conditions from large-scale structure surveys

Jens Jasche, Benjamin D. Wandelt

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **432,** 894–913 (2013)
Advance Access publication 2013 April 24

doi:10.1093/mnras/stt449

Bayesian physical reconstruction of initial conditions from large-scale structure surveys

Jens Jasche[1]⋆ and Benjamin D. Wandelt[1,2,3,4]

[1]*CNRS, UMR7095, Institut d'Astrophysique de Paris, F-75014 Paris, France*
[2]*UPMC Univ Paris 06, UMR7095, Institut d'Astrophysique de Paris, F-75014 Paris, France*
[3]*Department of Physics, 1110 W Green Street, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[4]*Department of Astronomy, 1002 N Gregory Street, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

Accepted 2013 March 11. Received 2012 October 29; in original form 2012 March 19

ABSTRACT

We present a fully probabilistic, physical model of the non-linearly evolved density field, as probed by realistic galaxy surveys. Our model is valid in the linear and mildly non-linear regimes and uses second-order Lagrangian perturbation theory to connect the initial conditions with the final density field. Our parameter space consists of the 3D initial density field and our method allows a fully Bayesian exploration of the sets of initial conditions that are consistent with the galaxy distribution sampling the final density field. A natural by-product of this technique is an optimal non-linear reconstruction of the present density and velocity fields, including a full propagation of the observational uncertainties. A test of these methods on simulated data mimicking the survey mask, selection function and galaxy number of the Sloan Digital Sky Survey Data Release 7 main sample shows that this physical model gives accurate reconstructions of the underlying present-day density and velocity fields on scales larger than $\sim$6 Mpc h^{-1}. Our method naturally and accurately reconstructs non-linear features corresponding to three-point and higher order correlation functions such as walls and filaments. Simple tests of the reconstructed initial conditions show statistical consistency with the Gaussian simulation inputs. Our test demonstrates that statistical approaches based on physical models of the large-scale structure distribution are now becoming feasible for realistic current and future surveys.

Key words: methods: numerical – methods: statistical – large-scale structure of Universe.

1 INTRODUCTION AND MOTIVATION

Ongoing and planned large-scale structure (LSS) surveys will measure the distribution of galaxies at an unprecendented level of accuracy in the coming decade. These surveys are expected to vastly enhance our constraints on the physics of cosmogenesis, neutrino physics and dark energy phenomenology.

How do we compare cosmological models to these surveys? We have an observationally well-supported physical model of the initial conditions. According to this model, a homogeneous and isotropic density field with small, very nearly Gaussian and nearly scale-invariant correlated density perturbations arose from quantum perturbations in the very early Universe. Gravitational evolution in an expanding background processed these initial conditions into an evolved density field, at first through linear transfer and then through non-linear structure formation. LSS surveys catalogue the positions of observed tracers of this evolved density field in redshift space.

⋆E-mail: jasche@iap.fr

It is now standard to model the initial Gaussian density perturbations statistically in terms of the early universe processes that created them, such as the physics of inflation, the change from matter to radiation-dominated universe, neutrino free-streaming and the acoustic oscillations of photon–baryon plasma. Within the standard cosmology, the evolution and growth of the initial perturbations in an expanding Universe is well understood in principle, and directly linked to its dominant constituents such as dark matter and dark energy. It therefore seems natural to analyse LSS surveys directly in terms of the simultaneous constraints they place on the initial density field and the physical evolution that links the initial density field to the observed tracers of the evolved density field.

For a variety of good reasons the current state of the art of statistical analyses of LSS surveys is far removed from this ideal. There are some areas where significant progress seems very difficult. In particular, a detailed physical model of the way galaxies arise in response to the spatial fluctuations in the dark matter distribution is not computationally tractable (the 'bias' problem). Even for the dark matter alone, reversing the non-linear evolution that link the initial and evolved density field is a fundamentally ill-posed problem (see e.g. Nusser & Dekel 1992; Crocce & Scoccimarro 2006).

As a consequence, the state of the art in the analysis of galaxy surveys addresses these problems in isolation. In the standard approach, the link between theory and observation is made through the power spectrum. This requires solving two separate problems: the data analysis problem of inferring the power spectrum from an observed sample of tracers given a survey mask and selection function (see e.g. Feldman, Kaiser & Peacock 1994; Eriksen et al. 2004; Tegmark et al. 2004; Wandelt, Larson & Lakshminarayanan 2004; Percival 2005; Jasche et al. 2010a; Elsner & Wandelt 2012); and the much more difficult theoretical problem of modelling the power spectrum *and the form of its likelihood* for the non-linearly evolved and biased galaxy density field (see e.g. Baugh, Gaztanaga & Efstathiou 1995; Peacock & Dodds 1996; Smith et al. 2003; Jeong & Komatsu 2006; Heitmann et al. 2010, and references therein).

Three-dimensional inference of the matter distribution from observations requires modelling the statistical behaviour of the mildly non-linear and non-linear regime of the matter distribution. The exact statistical behaviour of the matter distribution in terms of a probability distribution for the fully evolved density field is not known. Previous approaches therefore relied on phenomenological approximations such as multivariate Gaussian or log-normal distributions incorporating a cosmological power spectrum to accurately account for the correct two-point statistics of the density fields. Both of these distributions can be considered as maximum entropy prior on a linear and logarithmic scale, respectively, and are therefore well justified for Bayesian analysis. However, these priors only parametrize the two-point statistics of the matter distribution. Since LSS formation through gravitational clustering is essentially a deterministic process described by Einstein's equations and since the only stochasticity in the problem enters in the generation of initial conditions, it seems reasonable to account for the increasing statistical complexity of the evolving matter distribution by a dynamical model.

In this paper we describe progress towards such an approach that uses data to constrain a set of a priori possible dynamical, three-dimensional histories. We use second-order Lagrangian perturbation theory (2LPT) as a physical model of the gravitational dynamics that link the initial three-dimensional Gaussian density field to the observed, non-Gaussian density field. In Bayesian parlance our prior for the evolved density is the initial Gaussian density field evolved by a 2LPT model. Using the powerful sampling techniques recently developed by Jasche & Kitaura (2010) we can use this model as prior information and explore the range of initial Gaussian density fields that are statistically consistent with the data, modelled as a Poisson sample from evolved density fields.

Our method will also automatically generate reconstructions of the large-scale velocity field since our model incorporates dynamics. Since the approach is implemented in a fully Bayesian framework we do not produce unique reconstructions, but a set of *samples* which can be interpreted as a probabilistic representation of the information the observations contain about the underlying density (initial and evolved) and the velocity field. In particular, the variations between samples represent the uncertainties that remain in the reconstruction owing to the modelled statistical and systematic errors in the data.

1.1 Comparison to prior work

In the recent past several papers have pointed out the promise of the log-normal model in fitting to observations of the non-linear density field (see e.g. Jasche & Kitaura 2010; Jasche et al. 2010b; Kitaura, Jasche & Metcalf 2010). While the log-normal approach provides a good model of the one-point and two-point functions of the field we will show that Gaussian statistics evolved by 2LPT reproduces those successes but, in addition, reproduces features naturally that are associated with the higher order n-point functions such as filaments and walls. This is not surprising since it is well known that 2LPT reproduces the exact one- and three-point statistics of fully non-linear density fields at large scales, and also approximates higher order statistics very well (see e.g. Moutarde et al. 1991; Buchert, Melott & Weiss 1994; Bouchet et al. 1995; Scoccimarro 2000; Bernardeau et al. 2002; Scoccimarro & Sheth 2002).

The field of velocity field reconstructions has a long history (see e.g. Bertschinger et al. 1990; Nusser & Dekel 1992; Dekel et al. 1999; Frisch et al. 2002; Brenier et al. 2003; Lavaux 2008; Mohayaee & Sobolevskiĭ 2008; Kitaura et al. 2012). The contribution of our approach is the imbedding of a non-Gaussian model in a probabilistic framework. Zel'dovich and Monge–Ampère–Kantorovitch (MAK) are, respectively, perturbative and non-perturbative attempts to reconstruct the displacement field linking the initial conditions from tracers of LSS and as such also generate estimates of the velocity field. Our approach goes beyond these works in several ways: we combine the inference with a detailed non-Gaussian model of realistic survey features (mask, selection function and shot noise); we implement explicitly a Gaussian prior for the initial density field and the Bayesian exploration gives a quantitative characterization of the uncertainties in our inferences.

Significant effort has also been invested in establishing accurate representations of the observed Universe in numerical simulations, by constraining simulations by observations (see e.g. Kravtsov, Klypin & Hoffman 2002; Klypin et al. 2003; Dolag et al. 2005; Martinez-Vaquero et al. 2009; Gottloeber, Hoffman & Yepes 2010; Lavaux 2010; Libeskind et al. 2010). Many of these approaches rely on integrating the observed present-day density field backwards in time to the initial state. Such an approach is generally hindered due to incomplete observations of the final state and by spurious erroneous enhancement of decaying mode power in the initial conditions during backward integration (Nusser & Dekel 1992). The fully probabilistic approach, proposed in this work, naturally accounts for uncertainties of only partially observed final states, by exploring physical reasonable solutions, filtered by the 2LPT model, for the initial conditions which can all lead to the same or similar final observations. Furthermore, our method solely depends on forward evaluations of the model, which therefore accurately handles the issue of decaying mode power. Also note that unique recovery of initial conditions is generally not possible on all scales due to the chaotic nature of the dynamical LSS formation process on small scales (see e.g. Nusser & Dekel 1992; Crocce & Scoccimarro 2006). These uncertainties will also be accurately accounted for by our method while exploiting information on the initial conditions on all scales accessible to the 2LPT model.

The paper is structured as follows. In Section 2 we discuss the design of posterior distributions for LSS inference and show that the complex problem of modelling accurate prior distributions for the evolved non-Gaussian matter distribution can be recast as an initial conditions problem once a physical model for LSS formation is specified. Furthermore, we will present the resultant 2LPT–Poissonian posterior distribution for the inference of the three-dimensional matter distribution from galaxy surveys. Section 3 provides a brief overview over the Hamiltonian sampling approach employed in the inference framework described in this work, and in Section 4 we present the relevant derivations of the Hamiltonian forces required for an efficient numerical implementation of the hybrid Monte Carlo (HMC) sampler. In the following Section 6, we

describe the generation of an artificial galaxy survey, inspired by the Sloan Digital Sky Survey Data Release 7 main sample (Abazajian et al. 2009). In Section 7, we describe the application of our method to this simulated data in order to provide a proof of concept and to estimate the behaviour of the algorithm in a realistic setting. We will conclude the paper with a summary and a discussion of the results in Section 8.

2 THE 2LPT–POISSONIAN POSTERIOR

2.1 The non-Gaussian density prior

As already pointed out in the Introduction, inferring the three-dimensional LSS from observations requires the design of suitable prior distributions for the fully gravitationally evolved density field. Standard approaches such as Wiener filtering employ Gaussian priors, which are assumed to be suitable for the inference of the largest scales (see e.g. Lahav 1994; Zaroubi 2002; Erdoğdu et al. 2004; Kitaura & Enßlin 2008; Kitaura et al. 2009; Jasche et al. 2010a). For the inference of the density field in the non-linear regime log-normal priors have been proposed and successfully applied to LSS inference problems (Jasche & Kitaura 2010; Jasche et al. 2010b; Kitaura et al. 2010). More recently, Kitaura (2012) proposed to use Edgeworth expansions to construct prior distributions incorporating also third-order moments of the distribution. All of these approaches are based on heuristic approximations to the full problem. Currently, a closed form description of the present day density field in terms of a multivariate probability distribution does not exist.

While there exist considerable difficulties in constructing a suitable probability distribution for the present day density field, the initial seed fluctuations at redshifts $z \sim 1000$ obey Gaussian statistics to great accuracy, in agreement with theories of inflation and observations (see e.g. Linde 2008; Komatsu et al. 2011). Therefore, the complicated nature of the present matter distribution solely originates from deterministic physical processes during structure formation. Generally, gravitational interactions introduce mode coupling and phase correlations, such that the statistical behaviour of the present day density strongly deviate from a Gaussian distribution (see e.g. Peacock 1999).

Since initial and final conditions are linked via deterministic structure formation processes, it seems reasonable to formulate the inference problem in terms of simpler statistics at the initial conditions, rather than approximating the complex statistical behaviour of the non-linear matter distribution. More specifically, given a suitable model of LSS formation $G(a, \delta^{\rm i})$ we can obtain a prior distribution for the final density contrast $\delta^{\rm f}$ for a given cosmic scale factor a by marginalizing over the initial conditions:

$$\begin{aligned}\mathcal{P}(\{\delta_l^{\rm f}\}) &= \int {\rm d}\{\delta_l^{\rm i}\}\ \mathcal{P}(\{\delta_l^{\rm f}\}, \{\delta_l^{\rm i}\}) \\ &= \int {\rm d}\{\delta_l^{\rm i}\}\ \mathcal{P}(\{\delta_l^{\rm i}\})\ \mathcal{P}(\{\delta_l^{\rm f}\} \mid \{\delta_l^{\rm i}\}),\end{aligned} \tag{1}$$

where, for a deterministic structure formation model, the conditional probability is given by Dirac delta distributions:

$$\mathcal{P}\left(\{\delta_l^{\rm f}\} \,\middle|\, \{\delta_l^{\rm i}\}\right) = \prod_l \delta^{\rm D}\left(\delta_l^{\rm f} - G(a, \delta^{\rm i})_l\right). \tag{2}$$

Given a model $G(a, \delta^{\rm i})$ for structure formation, a prior distribution for the present-day density field can be obtained by a two-step sampling process, by first generating an initial conditions realization from the prior distribution $\mathcal{P}(\{\delta_l^{\rm i}\})$ and then propagating the initial state forward in time with a suitable model of LSS formation. This process amounts to generating samples from the joint prior distribution of the final and initial conditions:

$$\mathcal{P}(\{\delta_l^{\rm f}\}, \{\delta_l^{\rm i}\}) = \mathcal{P}\left(\{\delta_l^{\rm i}\}\right) \prod_l \delta^{\rm D}\left(\delta_l^{\rm f} - G\left(a, \delta^{\rm i}\right)_l\right). \tag{3}$$

By discarding the initial density realization, one obtains a realization from the prior distribution for the non-linear final state. Assuming, a multivariate Gaussian process with zero mean and covariance matrix $\mathbf{S}$ to generate the initial density field $\delta^{\rm i}$ the joint prior distribution is given as

$$\begin{aligned}\mathcal{P}\left(\{\delta_l^{\rm i}\}, \{\delta_l^{\rm f}\} \,\middle|\, \mathbf{S}\right) &= \mathcal{P}\left(\{\delta_l^{\rm i}\} \,\middle|\, \mathbf{S}\right) \prod_l \delta^{\rm D}\left(\delta_l^{\rm f} - G\left(a, \delta^{\rm i}\right)_l\right) \\ &= \frac{{\rm e}^{-\frac{1}{2}\sum_{lm}\delta_l^{\rm i}\mathbf{S}^{-1}_{lm}\delta_m^{\rm i}}}{\det(2\pi\mathbf{S})} \prod_l \delta^{\rm D}\left(\delta_l^{\rm f} - G\left(a, \delta^{\rm i}\right)_l\right).\end{aligned} \tag{4}$$

In this work, we will employ a 2LPT model to approximately describe gravitational LSS formation (also see Appendix B for an overview over the 2LPT model). By employing a Lagrangian model of structure formation, we particularly account for non-local effects of gravitational mass transport from initial to final positions. 2LPT is able to recover the exact one-, two- and three-point statistics at large scales, and also approximates higher order statistics very well (see e.g. Moutarde et al. 1991; Buchert et al. 1994; Bouchet et al. 1995; Scoccimarro 2000; Scoccimarro & Sheth 2002). The 2LPT model therefore naturally reproduces physically reasonable higher order statistics in the matter inference problem without requiring the introduction of additional parameters for the description of higher order statistics. Our approach therefore provides a physically meaningful way of matching the higher order statistics of the evolved density field while requiring no further knowledge other than the initial two-point statistics.

2.2 The large-scale structure likelihood

Above we demonstrated that the task of modelling accurate prior distributions for the statistical behaviour of the present-day matter distribution can be recast into an initial conditions inference problem once a model for LSS formation is specified.

The methods described in this work are general and can be adapted for the inference from any particular probe of the three-dimensional LSS. We will illustrate our method for the case of optical galaxy redshift surveys.

Galaxies tend to follow the gravitational potential of the cosmic matter distribution and thus can be considered as matter tracers. The statistical uncertainty due to the discrete nature of the galaxy distribution can be modelled as an inhomogeneous Poisson process (see e.g. Layzer 1956; Peebles 1980; Martínez & Saar 2002). Also note that Poissonian likelihoods have already been successfully employed for non-linear density field inference (see e.g. Jasche & Kitaura 2010; Jasche et al. 2010b; Kitaura et al. 2010, for details). Following this approach, the corresponding Poissonian likelihood distribution can be expressed as

$$\mathcal{P}\left(\{N_k^{\rm g}\} \,\middle|\, \{\lambda_k\}\right) = \prod_k \frac{(\lambda_k)^{N_k^{\rm g}}\, {\rm e}^{-\lambda_k}}{N_k^{\rm g}!}, \tag{5}$$

where $N_k^{\rm g}$ is the observed galaxy number at position $\boldsymbol{x}_k$ in the sky and λ_k is the expected number of galaxies at this position. The mean galaxy number is related to the final density field $\delta_k^{\rm f}$ via

$$\lambda_k = \lambda_k(\delta) = R_k \bar{N}\left(1 + B(\delta^{\rm f})_k\right), \tag{6}$$

where R_k is a linear response operator, incorporating survey geometries and selection effects, $\bar{N}$ is the mean number of galaxies in the volume and $B(x)_k$ is a non-linear, non-local, bias operator at position $\boldsymbol{x}_k$ (also see Jasche & Kitaura 2010; Jasche et al. 2010b, for further discussions).

The joint posterior distribution for the initial conditions $\delta_l^{\rm i}$ and the final density field $\delta_l^{\rm f}$ given the galaxy observations is then obtained by the multiplying equation (4) and (5):

$$\mathcal{P}\left(\left\{\delta_l^{\rm i}\right\},\left\{\delta_l^{\rm f}\right\}\middle|\left\{N_i\right\},\mathbf{S}\right) = \frac{\mathrm{e}^{-\frac{1}{2}\sum_{lm}\delta_l^{\rm i}\mathbf{S}_{lm}^{-1}\delta_m^{\rm i}}}{\det\left(2\pi\mathbf{S}\right)}\prod_l\delta^{\rm D} \times\left(\delta_l^{\rm f}-G(a,\delta^{\rm i})_l\right)\prod_k \times\frac{\left(\lambda_k\left(\delta^{\rm f}\right)\right)^{N_k^{\rm g}}\mathrm{e}^{-\lambda_k\left(\delta^{\rm f}\right)}}{N_k^{\rm g}!}. \quad (7)$$

We see that given a model of structure formation $G(a,\ \delta^{\rm i})$, the final density field $\delta_l^{\rm f}$ is a free by-product of the inference process. Consequently, marginalizing equation (7) over $\delta_l^{\rm f}$ yields the desired target posterior distribution for LSS inference:

$$\mathcal{P}\left(\left\{\delta_l^{\rm i}\right\}\middle|\left\{N_i\right\},\mathbf{S}\right) = \frac{\mathrm{e}^{-\frac{1}{2}\sum_{lm}\delta_l^{\rm i}\mathbf{S}_{lm}^{-1}\delta_m^{\rm i}}}{\det\left(2\pi\mathbf{S}\right)} \times\prod_k\frac{\left(\lambda_k\left(G(a,\delta^{\rm i})\right)\right)^{N_k^{\rm g}}\mathrm{e}^{-\lambda_k\left(G(a,\delta^{\rm i})\right)}}{N_k^{\rm g}!}. \quad (8)$$

This result requires several remarks. First, A nearly trivial, but nevertheless important, conclusion to draw from equation (8) is that LSS inference depends solely on the initial conditions $\delta_l^{\rm i}$. Consequently, the complex task of designing suitable prior distributions for the inference of the evolved density field can be recast into an initial value problem by modelling a suitable physical model to account for the complexity of the final state.

Secondly, note that inferring the initial density field does not involve backward in time integration of the physical model. The inference process exclusively requires model evaluations in the forward time direction, counter to the widely held notion that inference of initial conditions requires backward integration of the equations of motion. Nevertheless, traditional approaches of initial conditions inference, also known as 'time machines', rely on the inversion of the flow of time in the model equations (see e.g. Nusser & Dekel 1992). As pointed out by Nusser & Dekel (1992), the disadvantage of backward integration is that it may lead to artificial increase of decaying modes amplitudes introducing erroneous artificial density and velocity fluctuations in the initial conditions. Also note that LSS surveys only provide limited information on the full final state due to survey geometries and statistical uncertainties. These problems generally hinder a unique backward integration of the partially observed final state to the initial conditions.

To alleviate this problem, and to ensure physical meaningful backward integration of non-linear models, one has to augment unobserved regions in the data with statistically meaningful information mimicking the unobserved part of the evolved density field. This in turn requires accurate knowledge on the multivariate probability distribution for the evolved final state $\delta_l^{\rm f}$, which is not known at present.

Such problems are naturally prevented by casting the reconstruction of initial conditions as the statistical inference problem expressed in equation (8). Since the proposed method solely depends on forward model evaluations, unobserved regions and statistical uncertainties can be easily dealt with in the initial conditions, which amounts to modelling simple, uncorrelated Gaussian processes. Further, the posterior distribution proposed in equation (8) accounts for systematics, such as survey geometry, selection effects and biases but also for statistical uncertainties such as the noise of the galaxy distribution and cosmic variance.

We therefore see that statistical uncertainties do not allow a unique inference of the initial conditions from partially observed final states. Consequently, our approach aims at exploring the highly non-Gaussian and non-linear posterior distribution $\mathcal{P}\left(\{\delta_l^{\rm i}\}|\{N_i\},\mathbf{S}\right)$ of the initial density field $\delta_l^{\rm i}$ conditional on galaxy observations N_l in order to quantify the uncertainty and significance of the inferred initial conditions. The overall inference process is numerically non-trivial. It is made possible by sophisticated non-linear analysis methods, which will be described in the following.

3 HAMILTONIAN SAMPLING

As described in the previous section, exploration of the initial conditions posterior distribution requires numerically efficient Markov chain Monte Carlo (MCMC) methods to accurately account for all non-linearities and non-Gaussianities involved in the inference process. Unfortunately, unlike as in the Gibbs sampling approach for LSS proposed in Jasche et al. (2010a), direct sampling from this posterior is not possible. One therefore has to rely on a sampling procedures with an accept–reject step for the exploration of the high dimensional parameter space encountered in this problem. A major drawback of traditional algorithms of this type, e.g. Metropolis–Hastings, is their dominant random walk behaviour and a possible high rejection rate if no suitable proposal distribution can be designed. In this work, we usually deal with about 10^6, or more, free parameters $\delta_l^{\rm i}$ which correspond to the initial density contrast amplitudes at the volume elements of the analysed volume. Because of this high dimensionality of the problem, a low acceptance rate of the Metropolis–Hastings algorithm would result in a prohibitive computational cost for the method. Given this situation, we propose to use a HMC method, which in the absence of numerical errors, would yield an acceptance rate of unity. The HMC method exploits techniques developed to follow classical dynamical particle motion in potentials (Duane et al. 1987; Neal 1993, 1996). In this fashion the Markov sampler follows a persistent motion through the parameter space, suppressing the random walk behaviour. This enables us to sample with reasonable efficiency in high dimensional spaces (Hanson 2001). Furthermore, the HMC has already been successfully applied to non-linear LSS inference problems (see e.g. Jasche & Kitaura 2010; Jasche et al. 2010b).

In the following, we will just briefly outline the basic idea of the hybrid Hamiltonian sampling algorithm. More detailed description of the algorithm and its application in LSS inference and in general can be found in Duane et al. (1987), Neal (1993), Hanson (2001), Jasche & Kitaura (2010) and Jasche et al. (2010b).

3.1 The HMC

Suppose, we wish to generate samples from a probability distribution $\mathcal{P}(\{x_i\})$, where $\{x_i\}$ is a set consisting of the N elements x_i. If we interpret the negative logarithm of this posterior distribution as a potential

$$\psi(x) = -\ln(\mathcal{P}(x)), \quad (9)$$

and by introducing a 'momentum' variable p_i and a 'mass matrix' $\mathbf{M}$, as nuisance parameters, we can formulate a Hamiltonian

describing the dynamics in the multidimensional phase space. Such a Hamiltonian is then given as

$$H = \sum_i \sum_j \frac{1}{2}\, p_i\, \mathbf{M}^{-1}_{ij}\, p_j + \psi(x). \tag{10}$$

As can be seen in equation (10), the form of the Hamiltonian is such that the joint distribution is separable into a Gaussian distribution in the momenta $\{p_i\}$ and the target distribution $\mathcal{P}(\{x_i\})$ as

$$\mathrm{e}^{-H} = \mathcal{P}(\{x_i\})\, \mathrm{e}^{-\frac{1}{2}\sum_i \sum_j\, p_i\, \mathbf{M}^{-1}_{ij}\, p_j}. \tag{11}$$

For this reason, marginalization over all momenta will again yield the original target distribution $\mathcal{P}(\{x_i\})$.

In order to generate a random variate from this joint distribution, being proportional to $\exp(-H)$, one first draws a set of momenta from the distribution defined by the kinetic energy term that is an N dimensional Gaussian with a covariance matrix $\mathbf{M}$. Then one follows the deterministic dynamical evolution of the system, given a starting point $(\{x_i\},\ \{p_i\})$ in phase space for some fixed pseudo-time τ according to Hamilton's equations:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \frac{\partial H}{\partial p_i}, \tag{12}$$

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \frac{\partial H}{\partial x_i} = -\frac{\partial \psi(x)}{\partial x_i}. \tag{13}$$

The integration of this equations of motion yields the new position $(\{x_i'\},\ \{p_i'\})$ in phase space. This new point is accepted according to the usual acceptance rule:

$$\mathcal{P}_A = \min\left[1, \exp(-\left(H(\{x_i'\},\ \{p_i'\}) - H(\{x_i\},\ \{p_i\})\right))\right]. \tag{14}$$

Since the equations of motion provide a solution to a Hamiltonian system, energy or the Hamiltonian given in equation (10) is conserved, and therefore the solution to this system provides an acceptance rate of unity. In practice, numerical errors can lead to a somewhat lower acceptance rate. Samples of the desired target distribution are then obtained by simply discarding the auxiliary momenta $\{p_i\}$, which amounts to marginalization over these nuisance parameters. The particular implementation of the hybrid Hamiltonian Monte Carlo sampler for the problem described in this work will be discussed below.

4 EQUATIONS OF MOTION FOR LSS INFERENCE

As described above, the HMC approach permits to explore the non-linear LSS posterior by following Hamiltonian dynamics in the high dimensional parameter space. The corresponding forces required to evaluate these Hamiltonian trajectories can be derived from the LSS posterior given in equation (8). Consequently, the Hamiltonian potential $\Psi(\{\delta^{\mathrm{i}}_l\})$ can be written as

$$\begin{aligned}\Psi\left(\left\{\delta^{\mathrm{i}}_l\right\}\right) &= -\ln\left(\mathcal{P}\left(\left\{\delta^{\mathrm{i}}_l\right\}\,\middle|\,\{N_i\}, \mathbf{S}\right)\right)\\ &= \Psi_{\mathrm{prior}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right) + \Psi_{\mathrm{likelihood}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right),\end{aligned} \tag{15}$$

with the potential $\Psi_{\mathrm{prior}}(\{\delta^{\mathrm{i}}_l\})$ is given as

$$\Psi_{\mathrm{prior}}(\{\delta^{\mathrm{i}}_l\}) = \frac{1}{2}\sum_{lm} \delta^{\mathrm{i}}_l \mathbf{S}^{-1}_{lm}\, \delta^{\mathrm{i}}_m, \tag{16}$$

and $\Psi_{\mathrm{likelihood}}(\{\delta^{\mathrm{i}}_l\})$ is given as

$$\begin{aligned}\Psi_{\mathrm{likelihood}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right) = \sum_k R_k \bar{N}_{\mathrm{gal}}\, &\left(1 + G\left(a, \delta^{\mathrm{i}}\right)_k\right)\\ &- N_k \ln\left(R_k \bar{N}_{\mathrm{gal}}\, \left(1 + G(a, \delta^{\mathrm{i}})_k\right)\right).\end{aligned} \tag{17}$$

Given the above definition of the Hamiltonian potential $\Psi(\{\delta^{\mathrm{i}}_l\})$ one can obtain the required Hamiltonian forces by differentiating with respect to δ^{i}_p:

$$\frac{\partial \Psi\left(\left\{\delta^{\mathrm{i}}_l\right\}\right)}{\partial \delta^{\mathrm{i}}_p} = \frac{\partial \Psi_{\mathrm{prior}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right)}{\partial \delta^{\mathrm{i}}_p} + \frac{\partial \Psi_{\mathrm{likelihood}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right)}{\partial \delta^{\mathrm{i}}_p}. \tag{18}$$

Here the prior term is given by

$$\frac{\partial \Psi_{\mathrm{prior}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right)}{\partial \delta^{\mathrm{i}}_p} = \sum_j \mathbf{S}^{-1}_{pj}\, \delta^{\mathrm{i}}_j. \tag{19}$$

In contrast the likelihood term cannot be derived as trivially. A detailed derivation for the likelihood term can be found in Appendix D. The likelihood term $\Psi_{\mathrm{likelihood}}(\{\delta^{\mathrm{i}}_l\}))$ can be expressed as

$$\frac{\partial \Psi_{\mathrm{likelihood}}\left(\left\{\delta^{\mathrm{i}}_l\right\}\right)}{\partial \delta^{\mathrm{i}}_p} = -D^1\, J_p + D^2 \sum_{a>b} \left(\boldsymbol{\tau}_p^{aabb} + \boldsymbol{\tau}_p^{bbaa} - 2\boldsymbol{\tau}_p^{abab}\right), \tag{20}$$

where the vector J_p and the tensor $\boldsymbol{\tau}_p^{abcd}$ are defined in Appendix D.

Finally, the equations of motion for the Hamiltonian system given in equations (12) and (13) can be written as

$$\frac{\mathrm{d}\delta^{\mathrm{i}}_n}{\mathrm{d}t} = \sum_j \mathbf{M}^{-1}_{nj}\, p_j, \tag{21}$$

and

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = -\sum_j \mathbf{S}^{-1}_{nj}\, \delta^{\mathrm{i}}_j + D^1\, J_n + D^2 \sum_{a>b} \left(\boldsymbol{\tau}_n^{aabb} - \boldsymbol{\tau}_n^{bbaa} - 2\boldsymbol{\tau}_n^{abab}\right). \tag{22}$$

New samples from the LSS posterior can then be obtained by following the dynamical evolution of the Hamiltonian system in phase space.

5 NUMERICAL IMPLEMENTATION

We named our numerical implementation of the initial conditions sampler Bayesian Origin Reconstruction from Galaxies (BORG). It utilizes the FFTW3 library for fast Fourier transforms (FFTs) and the GNU Scientific Library (GSL) for random number generation (Frigo & Johnson 2005; Galassi et al. 2009). In particular, we use the Mersenne Twister MT19937, with 32-bit word length, as provided by the GSL_RNG_MT19937 routine, which was particularly designed for MCMC simulations (Matsumoto & Nishimura 1998).

5.1 The leapfrog scheme

The numerical implementation of the HMC sampler employed in this work largely follows the implementation described in Jasche & Kitaura (2010). Generally the numerical integration scheme is required to meet some conditions in order to achieve optimal efficiency of the sampler. High acceptance rates require the numerical integration scheme to be highly accurate in order to conserve the total Hamiltonian. Low accuracy in the integration scheme will generally lower the acceptance rate. Additionally, the integrator must be symplectic, meaning exactly reversible, in order to ensure

the Markov chain satisfies detailed balance (Duane et al. 1987). For this reason, we implemented the leapfrog scheme to integrate the Hamiltonian system. It is also numerically robust, and allows for simple propagation of errors. In particular, we implement the kick–drift–kick scheme. The equations of motions are integrated by making n steps with a finite step size ϵ, such that $\tau = n\epsilon$:

$$p_m\left(t+\frac{\epsilon}{2}\right) = p_m(t) - \frac{\epsilon}{2}\left.\frac{\partial\psi(\{\delta_k^{\rm i}\})}{\partial\delta_l^{\rm i}}\right|_{\delta_m^{\rm i}(t)}, \tag{23}$$

$$\delta_m^{\rm i}(t+\epsilon) = \delta_m^{\rm i}(t) - \frac{\epsilon}{m_i}\, p_m\left(t+\frac{\epsilon}{2}\right), \tag{24}$$

$$p_m(t+\epsilon) = p_m\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\left.\frac{\partial\psi(\{\delta_k^{\rm i}\})}{\partial\delta_l^{\rm i}}\right|_{\delta_m^{\rm i}(t+\epsilon)}. \tag{25}$$

We iterate these equations until $t = \tau$. Also note that it is important to vary the pseudo-time interval τ to avoid resonant trajectories. We do so by drawing n and ϵ randomly from a uniform distribution.

5.2 Hamiltonian mass

The HMC possesses a large number of tunable parameters contained in the 'mass' matrix $\mathbf{M}$ which have an important effect on the performance of the sampler. The Hamiltonian mass defines the inertia of individual parameters when moving through the parameter space. Consequently, too large masses will result in slow exploration efficiency, while too light masses will result in large numerical errors of the integration scheme leading to high rejection rates.

Generally, if the individual $\delta_l^{\rm i}$ would be Gaussian distributed, a good choice for HMC masses is to set them inversely proportional to the variance of that specific $\delta_l^{\rm i}$ (Taylor, Ashdown & Hobson 2008). For non-Gaussian distributions it is reasonable to use some measure of the width of the distribution (Taylor et al. 2008). For example, Neal (1996) proposes to use the curvature at the peak. A suitable approach to define Hamiltonian masses is to perform an approximate stability analysis of the numerical leapfrog scheme (see e.g. Taylor et al. 2008; Jasche & Kitaura 2010). Following this approach, we will expand the Hamiltonian forces, given in equation (18), around a mean signal ξ_l^i, which is assumed to be the mean initial density contrast once the sampler moved beyond the burn-in phase. As described in Appendix F approximating the Hamiltonian forces to linear order amounts to approximating the target distribution by a Gaussian distribution. According to the discussion in Appendix F, the Hamiltonian masses should be set as

$$\mathbf{M}_{mj} = \mathbf{S}^{-1}_{mj} - \delta^K_{mj}\, D^1\left.\frac{\partial J_j(s)}{\partial s_j}\right|_{s_j=\xi_j}, \tag{26}$$

where J_j is defined in Appendix D. Calculation of the leapfrog scheme requires inversions of $\mathbf{M}$. Considering the high dimensionality of the problem, inverting and storing $\mathbf{M}^{-1}$ is computationally impractical. For this reason, we construct a diagonal 'mass matrix' from equation (26). We found that choosing the diagonal of $\mathbf{M}$, as given in equation (26), in its Fourier basis yields faster convergence for the sampler than a real space representation since it accounts for the correlation structure of the underlying density field.

6 GENERATING MOCK OBSERVATIONS

In the previous sections we presented the derivation and the implementation of our method. Here we will describe the generation of mock data sets that will be used to test our method. Following closely the description in Jasche & Kitaura (2010), we will first generate a realization for the density contrast $\delta_l^{\rm i}$ from a normal distribution with zero mean and a covariance matrix corresponding to a cosmological power spectrum in a three-dimensional Cartesian box with $N_{\rm side} = 128$, corresponding to $N_{\rm vox} = 2097\,152$ volume elements, and a comoving box length of $L = 750\,{\rm Mpc}\,h^{-1}$. The density field will then be scaled back to an initial time corresponding to a cosmological scale factor $a_{\rm init} = 0.001$ by multiplication with a cosmological growth factor $D^+(a_{\rm init})$, described in Appendix A. In particular, we use a cosmological power spectrum for the underlying matter distribution, with baryonic wiggles, calculated according to the prescription described in Eisenstein & Hu (1998, 1999). We assume a standard Λ cold dark matter (ΛCDM) cosmology with a set of cosmological parameters ($\Omega_{\rm m} = 0.22$, $\Omega_\Lambda = 0.78$, $\Omega_{\rm b} = 0.04$, $h = 0.702$, $\sigma_8 = 0.807$ and $n_{\rm s} = 0.961$). The Gaussian initial conditions are then propagated forward in time using 2LPT as described in Appendix B. From the resultant particle distribution the final density contrast field $\delta_l^{\rm f}$ is constructed via the cloud in cell (CIC) method (see e.g. Hockney & Eastwood 1988).

An artificial galaxy catalogue is then generated by simulating the inhomogeneous Poisson process given by equation (5) on top of the final density field $\delta_l^{\rm f}$. In order to set up a realistic testing environment, we seek to emulate the main features of the Sloan Digital Sky survey as closely as possible. We employ the survey geometry of the Sloan Digital Sky Survey Data Release 7 depicted in the right-hand panel of Fig. 1. The mask has been calculated using the MANGLE code provided by Swanson et al. (2008) and has been stored on a HEALPix map with $n_{\rm side} = 4096$ (Górski et al. 2005). Further, we assume that the radial selection function follows from a standard Schechter luminosity function with standard r-band parameters ($\alpha = -1.05$, $M_* - 5\log_{10}(h) = -20.44$), and we only include galaxies within an apparent Petrosian r-band magnitude range $14.5 < r < 17.77$ and within the absolute magnitude ranges $M_{\rm min} = -17$ to $M_{\rm max} = -23$. The corresponding radial selection function $f(z)$ is then obtained by integrating the Schechter luminosity function over the range in absolute magnitude:

$$f(z) = \frac{\int_{M_{\rm min}(z)}^{M_{\rm max}(z)} \Phi(M)\,{\rm d}M}{\int_{M_{\rm min}}^{M_{\rm max}} \Phi(M)\,{\rm d}M}, \tag{27}$$

where $\Phi(M)$ is given in Appendix C. The resulting selection function for the simulated galaxy sample is depicted in the left-hand panel of Fig. 1. The survey response operator R_i, required to simulate the Poisson process, can be calculated as the product of the survey geometry and the selection function at each point in the three-dimensional volume:

$$R_i = M_i\, F_i = M(\alpha_i, \delta_i) f^{\prime}(z_i). \tag{28}$$

Finally, we choose $\bar{N} = 9.93$, and perform the Poisson sampling on the grid resulting in a total number of galaxies $N_{\rm tot} = 484\,227$.

7 TESTING

In this section, we describe the application of our method to the artificial data set described in Section 6. The primary intention of the following tests is to evaluate the performance of our method in realistic situations, taking into account observational systematics and uncertainties.

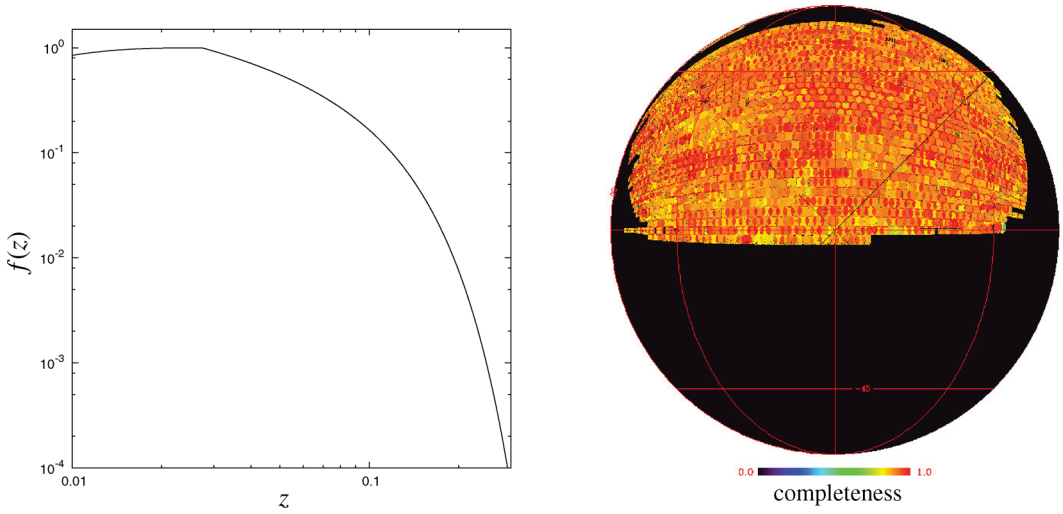

Figure 1. Selection function $f(z)$ as a function of redshift z (left-hand panels) and the two-dimensional completeness map for the SDSS DR7 (right-hand panel).

7.1 Testing convergence and correlations

The Metropolis–Hastings sampler in general and the HMC in particular are designed to have the target distribution, in our case the LSS posterior distribution, as its stationary distribution (see e.g. Metropolis et al. 1953; Hastings 1970; Neal 1993). For this reason, the sampling process will provide us with samples from the specified LSS posterior distribution after an initial burn-in phase. The length of this initial burn-in phase has to be assessed using numerical experiments.

Generally, burn-in manifests itself as a systematic drift of the sampled parameters towards the true parameters from which the artificial data set was generated. This behaviour can be monitored by following the evolution of parameters in subsequent samples (see e.g. Eriksen et al. 2004; Jasche et al. 2010a). To test this initial burn-in behaviour we will arbitrarily reduce the power of the random initial density field $\delta_i^{\rm i}$ by a factor of 0.1, and observe the drift towards the true underlying values by following successive power spectra measured from the samples. In Fig. 2 successive power spectra of the first 800 samples are presented. The plot nicely demonstrates the systematic drift towards the true underlying solution by successive restoration of the true power in the initial density field.

More specifically, we can quantify the initial burn-in behaviour by comparing the ith power spectrum measurement $P_i(k)$ in the chain to its true underlying value $P^0(k)$ via

$$\xi\left(P_i(k)\right) = 1. - \frac{P_i(k)}{P^0(k)}. \tag{29}$$

The results for this exercise are presented in the lower panel of Fig. 2. It can be nicely seen that the algorithm restores the correct power an all scales and drifts towards preferred regions in parameter space to commence exploration of the LSS posterior. It is also important to remark that in this test, we do not observe any particular hysteresis for the poorly constrained large-scale modes, meaning they do not remain at their initially set values but efficiently explore the parameter space. This already indicates the ability of our algorithm to account for artificial mode coupling as introduced by the survey geometry.

Burn-in also manifests itself in the initial acceptance rate as shown in the left-hand panel of Fig. 3. The dip in the initial acceptance rate function corresponds to the point when the sampling algorithm restored the full power of the initial density field. This has a simple explanation. Initially, since the power was reduced by a factor of 10, the system behaved more or less linear since the displacement of 2LPT particles is small. Once the correct power is restored the displacement of particles increases, leading to a higher non-locality of the system. When the dip in the acceptance rate occurs, the sampler starts exploring physically more reasonable

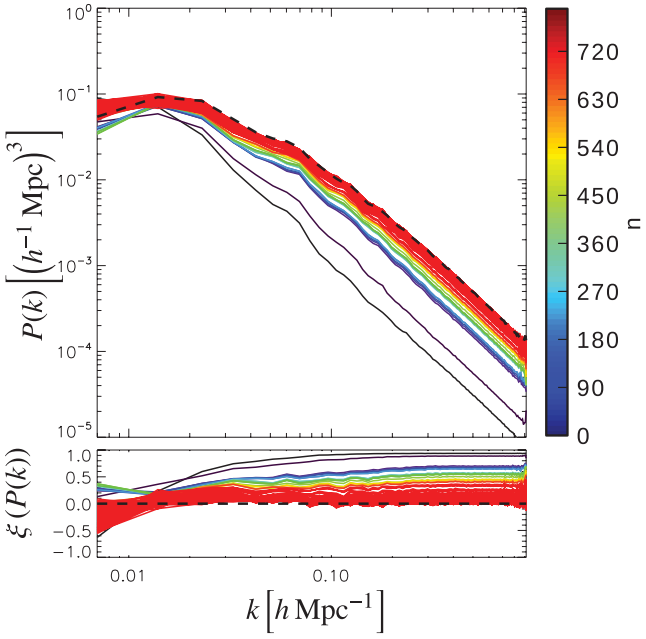

Figure 2. The plot demonstrates the initial burn-in drift of successive power spectra, measured from the initial density fields, towards the true underlying solution. Successive samples are colour coded corresponding to their sample number as indicated by the colour bar on the right. Black dashed lines correspond to the true underlying values. Lower panels depict the successive deviation ξ from the true values, as described in the text, for the measured power spectra. The sequence of 800 successive samples, visualizes how the sampler approaches the true underlying values and starts exploring the parameter space around them.

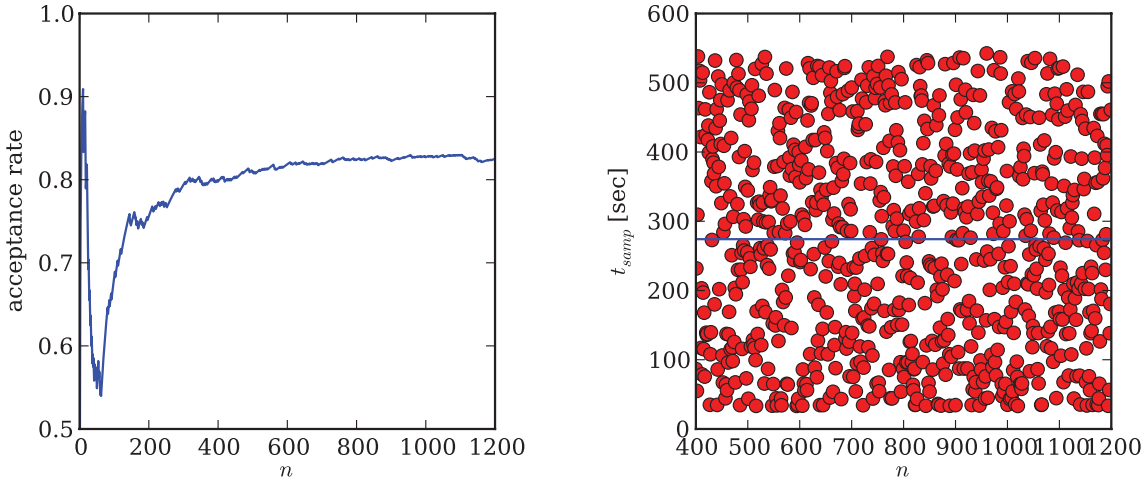

Figure 3. Acceptance rates for successive samples (left-hand panel) and the execution time per sample (right-hand panel). It can be seen that the acceptance rates drops during the initial burn in phase and finally stabilizes at about 84 per cent. The left-hand panel demonstrates the scatter in the execution times of individual samples. The average execution time is about 300 s as indicated by the solid blue line.

states in the initial conditions which can explain the observations. This process is accompanied by an initially lower acceptance rate. Once the sampler moves to a reasonable region in parameter space the acceptance rate approaches asymptotically a rate of about 84 per cent. This high acceptance rate, combined with the fast de-correlation properties, we will demonstrated next, shows that our sampler makes sampling from this multimillion dimensional, non-linear posterior numerically feasible.

In particular, these tests show that the algorithm requires an initial burn-in phase of on the order of 600 samples before providing samples from the target distribution. Also note that the initial burn-in period can be shortened by initializing the algorithm with an initial density field which is closer to the true solution.

Another important issue to study when dealing with MCMC methods is the mixing efficiency of the algorithm. Generally, successive samples in the chain will not be independent but correlated with previous samples. Consequently, the correlation between successive samples determines the amount of independent samples which can be drawn from the chain. We study this effect by following a similar approach as described in Eriksen et al. (2004) or Jasche et al. (2010a).

Assuming all parameters in the Markov chain to be independent of one another one can estimate the correlation between subsequent density samples by calculating the autocorrelation function:

$$C(\delta)_n = \left\langle \frac{\delta^{\mathrm{i}} - \langle\delta\rangle}{\sqrt{\mathrm{Var}\,(\delta)}} \frac{\delta^{\mathrm{i}+n} - \langle\delta\rangle}{\sqrt{\mathrm{Var}\,(\delta)}} \right\rangle, \tag{30}$$

where n is the distance in the chain measured in iterations (also see e.g. Eriksen et al. 2004; Jasche et al. 2010a, for a similar discussion). The results for this analysis are presented in Fig. 4, where we plot the correlation coefficients for individual density amplitudes selected by their signal-to-noise ratio. It can be generally seen that the mixing efficiency is a little lower in regions with low signal-to-noise ratio but the mixing efficiency of the algorithm is very good overall.

We can further define a correlation length of the Markov sampler as the distance in the chain n_{c} beyond which the correlation coefficient $C(\delta)_n$ has dropped below a threshold of $C^{\mathrm{th}}(\delta)_n = 0.1$. As can be seen in Fig. 4 the correlation length is about 200 samples, demonstrating the high mixing efficiency of the algorithm. Despite the high dimensionality of the problem considered here, these tests demonstrate that exploring LSS posterior for the initial

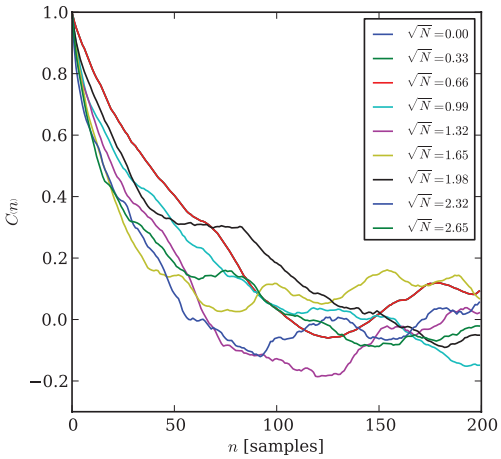

Figure 4. Correlation length for different signal-to-noise ratio values $\sqrt{N}$, as indicated in the legend.

conditions from observations exhibiting systematic uncertainties and uncertainties are numerically feasible with our method.

7.2 Large-scale structure inference

In this section we will discuss the results obtained from the application of our MCMC algorithm to the artificial data set, as described in Section 6.

7.2.1 Inferred three-dimensional density fields

We performed our analysis in 128^3 cubic Cartesian box with side length of 750 Mpc h^{-1}, yielding $\sim 2 \times 10^6$ parameters to infer. In the course of the sampling process, our algorithm, therefore, provides matter field realizations for the initial and final 2LPT density fields, with a grid resolution of about $\sim$6 Mpc h^{-1}, constrained by observed data. Also note that by employing a 2LPT model, initial density fields are naturally inferred at their initial Lagrangian coordinates, while final density fields are recovered at their corresponding final Eulerian coordinates. Our analysis resulted in $\sim 2 \times 10^4$ samples for initial and final density fields, which

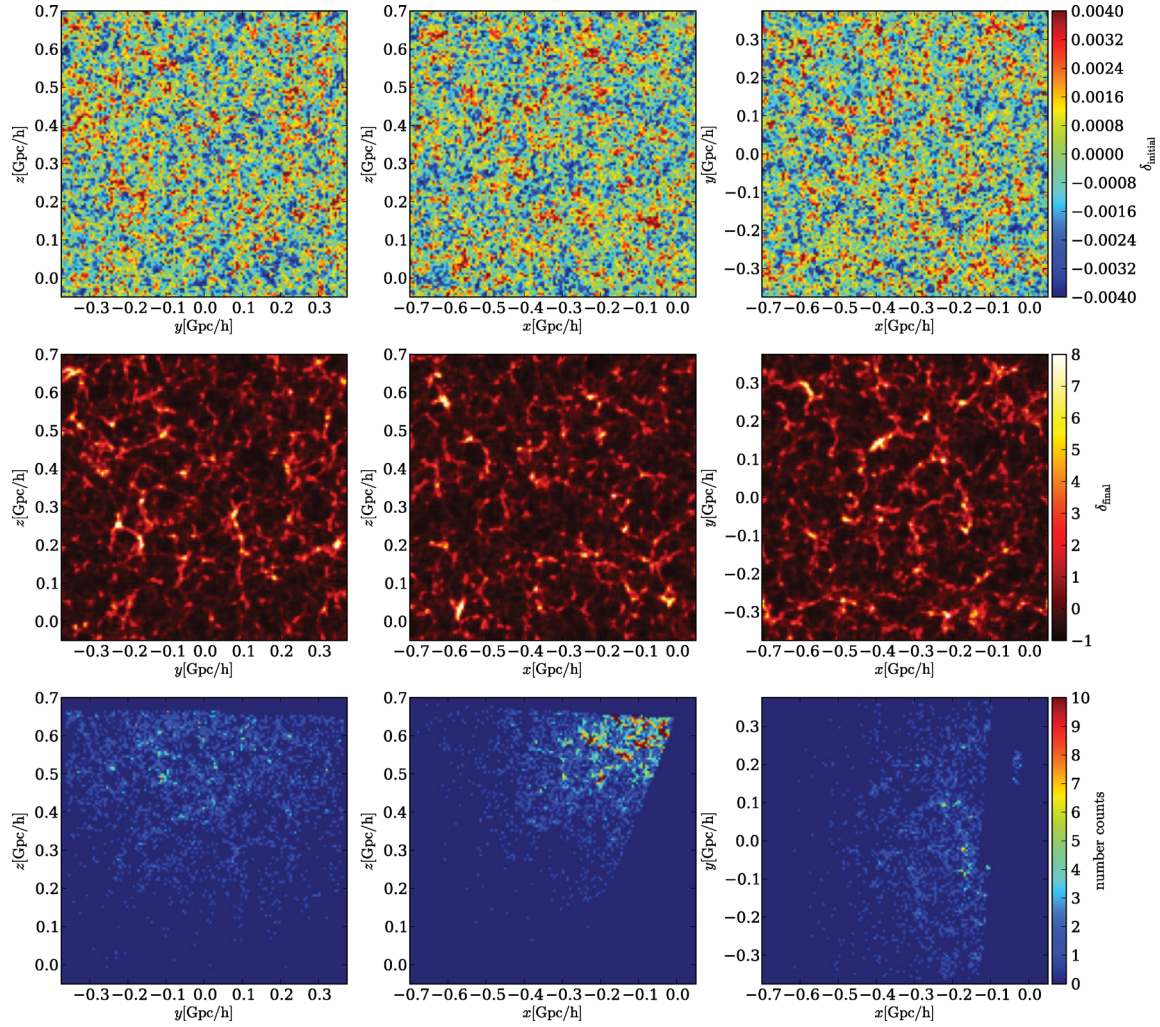

Figure 5. Three slices through a sample of the initial density field (top panels), the final density field (middle panels) and through the corresponding data cube represented by the galaxy number counts (lower panels). The plots nicely demonstrate the correlation between the final density field and the data. To some extent, one can observe the connection between large structures in the initial conditions and the final density field.

can be considered as full physical density fields, at least to the degree as permitted by the validity of the 2LPT model. In particular, as indicated by the inferred power spectra in Fig. 2, the individual samples possess correct power, and do not show any sign of attenuation due to survey characteristics such as survey geometry and selection effects. In Fig. 5 we show slices from three different sides through samples of the initial and corresponding final density fields as well as through the data. It is immediately visible that the statistics of the initial and final matter fields differ greatly. While the initial density field appears to be very Gaussian, the final density field is clearly non-Gaussian. This demonstrates how the physical 2LPT model for structure formation is able to account for the growing statistical complexity of the density distribution during the evolution from the initial to the final state. Furthermore, comparison of the final density field (middle panels of Fig. 5) to the data (lower panels of Fig. 5) demonstrates the recovered structures from the data. One can nicely see that the algorithm tries to extrapolate unobserved filaments between clusters based on the physically reasonable assumptions provided by the 2LPT model. In general, the algorithm nicely recovers the filamentary structure of the matter distribution.

More importantly, Fig. 5 also illustrates that the algorithm accurately accounts for survey geometry and selection effects by augmenting unobserved or poorly observed regions with statistically correct information. Consequently, inferred initial and final density fields possess equal power throughout their entire domains and are not affected by selection or mask artefacts, thus representing physical matter field realizations. It is particularly interesting that unobserved and observed regions in the inferred final density fields do not appear visually distinct, a consequence of the excellent approximation of the 2LPT not just to the first but also higher order moments (Moutarde et al. 1991; Buchert et al. 1994; Bouchet et al. 1995; Scoccimarro 2000; Scoccimarro & Sheth 2002). This is a great advantage over previous methods based on Gaussian or log-normal models where the assumption of a cosmological power spectrum only specifies the two-point statistics correctly. In particular, the reader may want to compare with fig. 2 in Jasche et al. (2010b), where unobserved regions are augmented with a log-normal model unable to represent filamentary structures.

To further illustrate the degree of information which can be extracted from observations, subject to selection effects and survey geometry, in Fig. 6 we show ensemble mean density fields estimated from 1.5×10^4 samples. The plot reveals that on average highly detailed structures can be recovered. In comparison with Fig. 5, one can see that in poorly or not observed regions, the ensemble mean

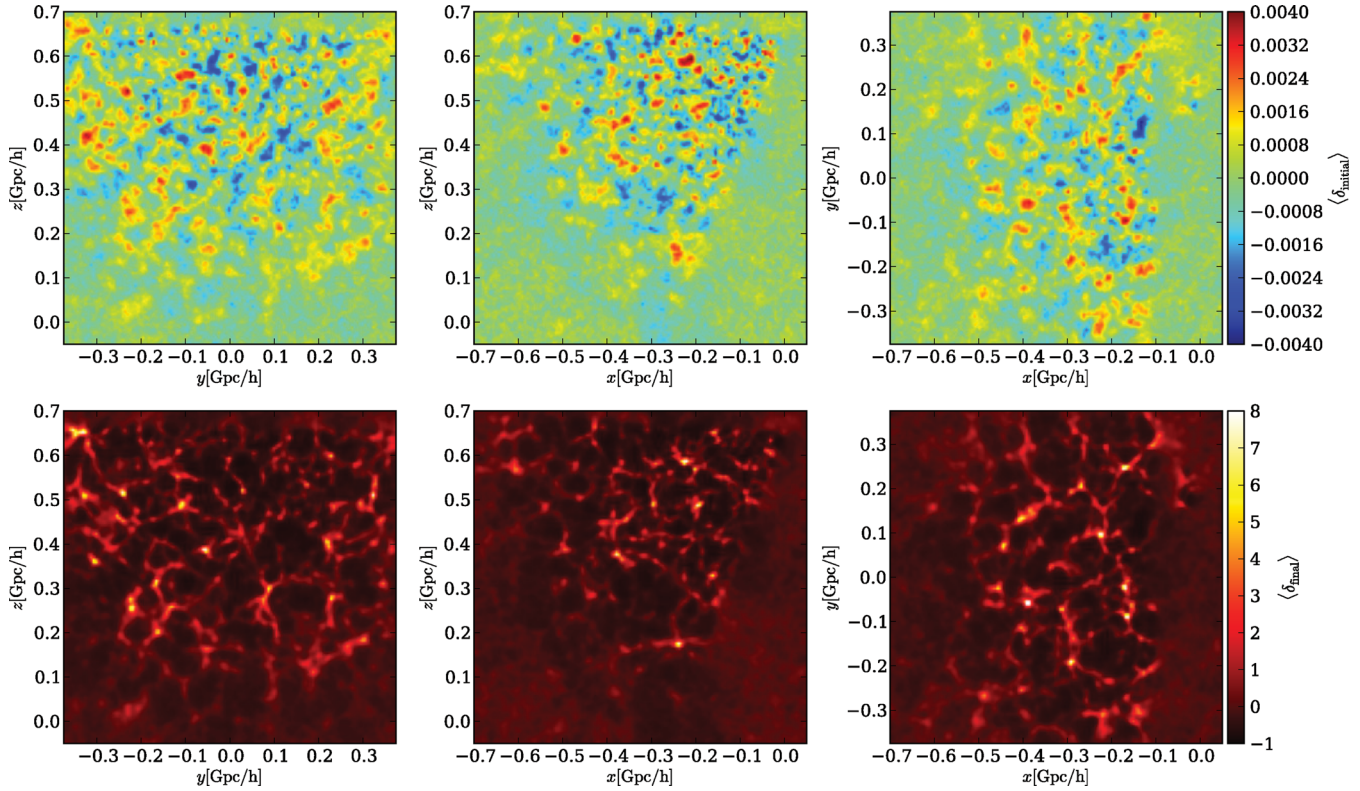

Figure 6. Three slices from different perspectives through three-dimensional ensemble mean fields estimated from 1.5×10^4 samples. Upper and lower panels show slices through the initial and final mean density fields, respectively.

density field approaches the cosmic mean density. This is expected, since regions which do not provide any observational information should on average reflect the cosmic mean. Note, however, that the uncertainty in these regions is accurately accounted for in the inference process, by augmenting these regions with statistically correct information as demonstrated by Fig. 5.

Next, we will discuss the statistical behaviour of inferred initial and final density fields. In Fig. 7 we compare the one-point distribution of the inferred initial and final density field measured from the corresponding samples. It can be seen that while the initial density contrast follows Gaussian statistics, the final distribution is highly skewed and represents the expected log-normal like behaviour. These results therefore supports our initial claim that the complex problem of modelling a prior distribution for the present fully non-linear density field can be exchanged for an initial conditions inference problem when assuming a physical model which accounts for the increasing statistical complexity of the matter distribution during structure formation.

7.2.2 *Testing accuracy of inferred initial conditions*

An important task of the present work is to test the accuracy of inferred initial conditions from galaxy observations. In particular, not only survey characteristics such as survey geometry and selection effects but also noise have to be accurately translated to the initial conditions. This is a non-trivial task, since gravitational structure formation is a non-local and non-linear process. As a consequence, also the information content of the observed data will be distributed differently in initial and final fields. Although, the total amount of information is conserved, by following for example tracer particles from high density, and thus high signal-to-noise ratio, regions backward in time, one sees that the same amount of information is distributed over a larger region in the initial conditions. The analogue result applies to underdense regions. This means in particular that the signal-to-noise ratio for a given comoving Eulerian volume is a function of time. In our approach, we account for these non-local processes by incorporating a 2LPT model into the inference process.

In general it is important to estimate the accuracy of inferred initial conditions, in particular to test for possible biases. However, note that the agreement between the recovered and true initial density fields will crucially depend on the signal-to-noise ratio at the various regions of the initial conditions. To estimate a proxy for the signal-to-noise ratio in the initial conditions we calculate the ratio Σ between the absolute value of the ensemble mean and the ensemble variance given as

$$\Sigma = \frac{|\langle \delta_{\rm initial} \rangle|}{\sqrt{\langle (\delta_{\rm initial} - \langle \delta_{\rm initial} \rangle)^2 \rangle}}. \tag{31}$$

To test whether our method is biased with respect to the true underlying initial density field $\delta^{\rm true}_{\rm initial}$ we analyse the scatter $\Delta\delta_{\rm initial} = \delta^{\rm true}_{\rm initial} - \delta_{\rm initial}$ between the true underlying and inferred density fields. In particular, we estimated the posterior distribution $\mathcal{P}(\Delta\delta_{\rm initial}|\Sigma)$ of deviations from the true values, conditional on the signal-to-noise ratio parameter Σ. The result of this exercise is demonstrated in Fig. 8. As can be clearly seen, the distribution $\mathcal{P}(\Delta\delta_{\rm initial}|\Sigma)$ is centred on the vertical zero axis and thus demonstrates our method to be unbiased with respect to the true underlying initial density field. Moreover, the plot also quantifies the accuracy by which the initial density field can be inferred depending on the signal-to-noise ratio parameter Σ. In particular, as expected regions

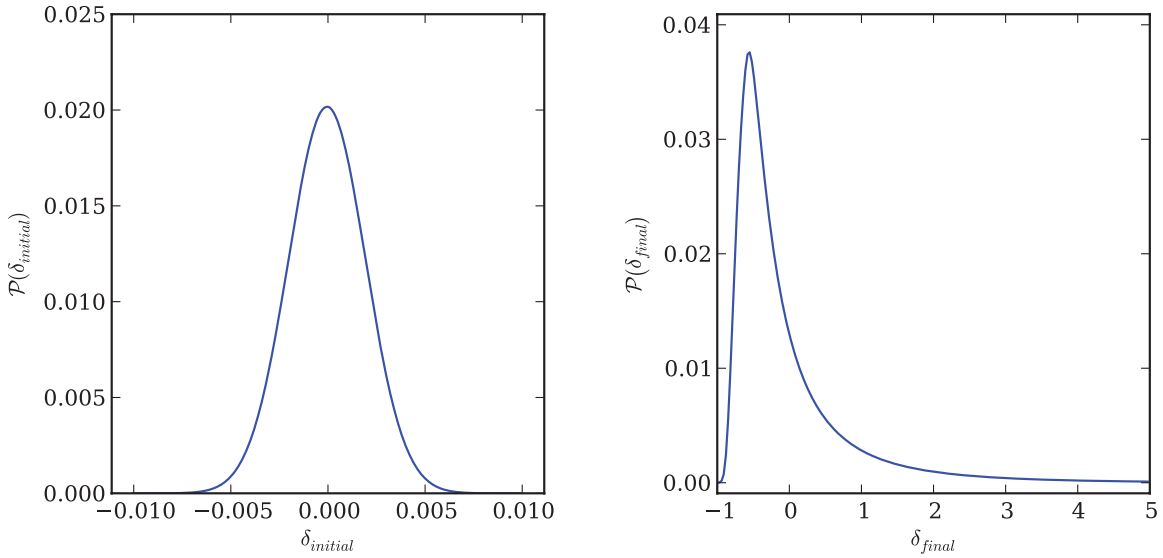

Figure 7. One-point distributions for the density contrast in the initial field (left-hand panel) and for the final field (right-panel) measured from the samples. It can be seen that, while the inferred initial density field follows a Gaussian distribution, the final field exhibits the highly skewed log-normal like behaviour.

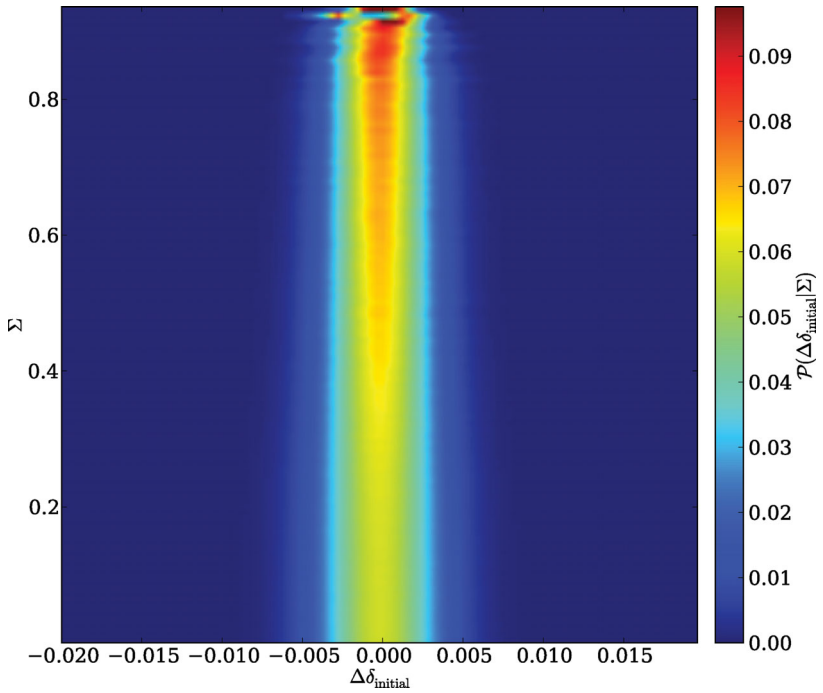

Figure 8. Posterior distribution for the deviations $\Delta\delta_{\text{initial}} = \delta^{\text{true}}_{\text{initial}} - \delta_{\text{initial}}$ of the inferred initial density amplitudes from their true underlying values conditional on the signal-to-noise ratio parameter Σ. As expected the accuracy of inferred δ_{initial} values depends on Σ. In particular, regions with higher Σ also yield higher accuracy for inferred density amplitudes δ_{initial}. Also note that the plot shows no bias, demonstrating that posterior results are unbiased with respect to the true underlying initial density field.

of higher signal-to-noise ratio can be recovered with greater accuracy compared to regions with low Σ.

To further quantify the accuracy of the recovered density field, we estimate the correlation coefficient $r(x)$ between density samples and the true underlying solution as a function of some parameter x. The correlation coefficient is given as

$$r(k_x) = \frac{\left\langle \delta_0^x \, \langle\delta\rangle^x \right\rangle}{\sqrt{\left\langle \left(\delta_0^x\right)^2 \right\rangle} \sqrt{\left\langle (\langle\delta\rangle^x)^2 \right\rangle}}, \tag{32}$$

where we will choose x to be the signal-to-noise ratio $\sqrt{N}$ for the final density field and a specific smoothing scale k_{th} for the initial density field. We also present the correlation coefficient $r(x)$ of inferred and true initial density fields for various values of the signal-to-noise ratio parameter Σ. The results for these tests are demonstrated in Fig. 9. The left-hand panel of Fig. 9 depicts the correlation between the true underlying final density field and the final density samples as a function of the signal-to-noise ratio $\sqrt{N}$. It can be seen that the correlation with the truth is generally higher for higher signal-to-noise ratios. Even in zones that contain

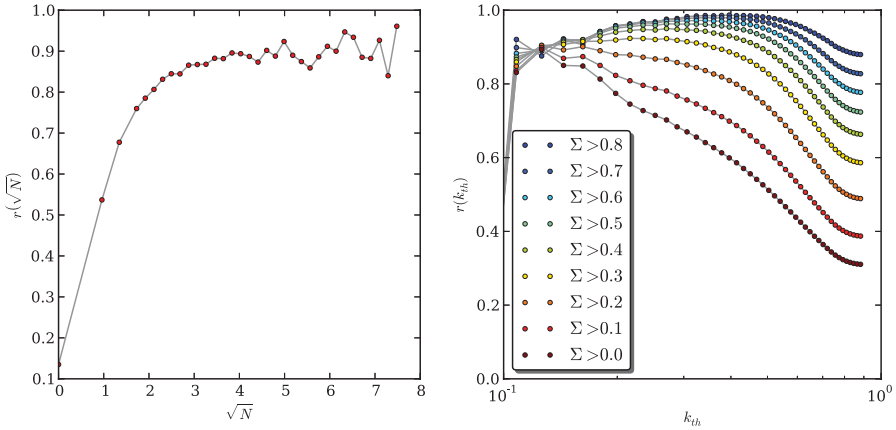

Figure 9. Cross-correlation coefficient between the true final density field and a sample as a function of signal-to-noise ratio $\sqrt{N}$ (left-hand panel) and the cross-correlation between the true underlying initial density field and the inferred ensemble mean initial density field as a function of smoothing scale $k_{\rm th}$ for different values of the signal-to-noise ratio parameter Σ (right-hand panel). It is interesting to remark that the correlation between true underlying and samples of the density field still amounts to about 55 per cent in regions where only a single galaxy has been observed. Also note that the accuracy of inferred initial density fields depends on survey characteristics through the signal-to-noise ratio parameter Σ. Particularly regions with high Σ can be inferred with more than 90 per cent accuracy on all scales considered in our analysis.

just a single galaxy we still get a correlation of about 55 per cent. It is also remarkable that the algorithm still provides a 10 per cent correlation with the true underlying density field in regions which have not been sampled by galaxies such as centres of voids or masked regions. The right-hand panel of Fig. 9 demonstrates the cross-correlation between the true underlying initial conditions and the inferred ensemble mean initial density field as a function of filter scale $k_{\rm th}$, for various values of Σ, when smoothed with a spherical top hat filter in Fourier space. These results clearly demonstrate that the accuracy of inferred structures in the initial conditions strongly depends on survey characteristics through the signal-to-noise ratio value Σ. In particular, Fig. 9 reveals that for the highest signal-to-noise ratio regions, the method recovers structures in the initial conditions with an accuracy of about 90 per cent throughout all scales considered in this analysis. Note that even for the lowest signal-to-noise ratio regime we still yield a 30 per cent correlation between the true and inferred initial density fields on a scale of about $\sim$6 Mpc h^{-1}. In general, even though only roughly half of the analysed volume has been observed, the large scales of the initial conditions can be much easier recovered than small-scale features. This is in agreement with expectations, since the largest scales behave more linearly than the smaller scales and hence are easier to recover and also because the Lagrangian dynamics involve the gravitational potential which carries information on the largest scales. Particular the shot noise contribution at the smallest scales in the final galaxy observation will smear out features in the initial conditions, since the 2LPT displacement vector for the particles will fluctuate on these scales.

7.2.3 Inferred dynamics

Importantly, the algorithm provides dynamical information on the LSS given the 2LPT model. In Fig. 10, we show the comparison between the true underlying velocity field and the velocity field inferred by a randomly selected sample. It can be seen that the algorithm is able to recover the true underlying velocity field in detail. This is expected, since as demonstrated by Fig. 9, our method yields on average an accuracy of about 90 per cent at recovering the largest scales in the initial conditions, and since bulk velocities are mostly sensitive to the largest scales by being related to the gravitational potential. As a consequence our inference is able to accurately infer the velocity field in noisy or even completely masked regions. This clearly demonstrates the strength of this approach in extrapolating physically reasonable states of the matter distribution even into poorly observed regions.

8 DISCUSSION AND CONCLUSION

We describe a new method to perform dynamical LSS inference from galaxy redshift surveys employing a 2LPT model. In Section 2 we demonstrated that the problem of constructing suitable prior distributions for the non-linear density field is directly linked to the problem of inferring initial conditions, once a dynamical model for LSS formation is given. In this approach the evolved non-linear density field acts as a mere nuisance parameter in the inference process, which shifts the problem of designing prior distributions to physical modelling of the matter evolution dynamics.

Since the method we propose provides an approximation to the non-linear dynamics the algorithm automatically provides information on the dynamical evolution of the large-scale matter distribution. By exploring the space of dynamical *histories* compatible with both data and model our approach therefore marks the passage from Bayesian three-dimensional density inference to full four-dimensional state inference.

Particularly, in this work we have employed a 2LPT model as an approximate dynamical description of the LSS evolution on the large scales. As described in the literature, the 2LPT model describes the one-, two- and three-point statistics correctly and represents higher order statistics very well (see e.g. Moutarde et al. 1991; Buchert et al. 1994; Bouchet et al. 1995; Scoccimarro 2000; Scoccimarro & Sheth 2002). Hence, the algorithm proposed in this work can exploit higher order statistics, modelled through the 2LPT model, to provide physically reasonable matter field realizations conditional on the observed galaxy distribution.

It is also important to remark that the inference process described in Section 2 requires at no point the inversion of the flow of time

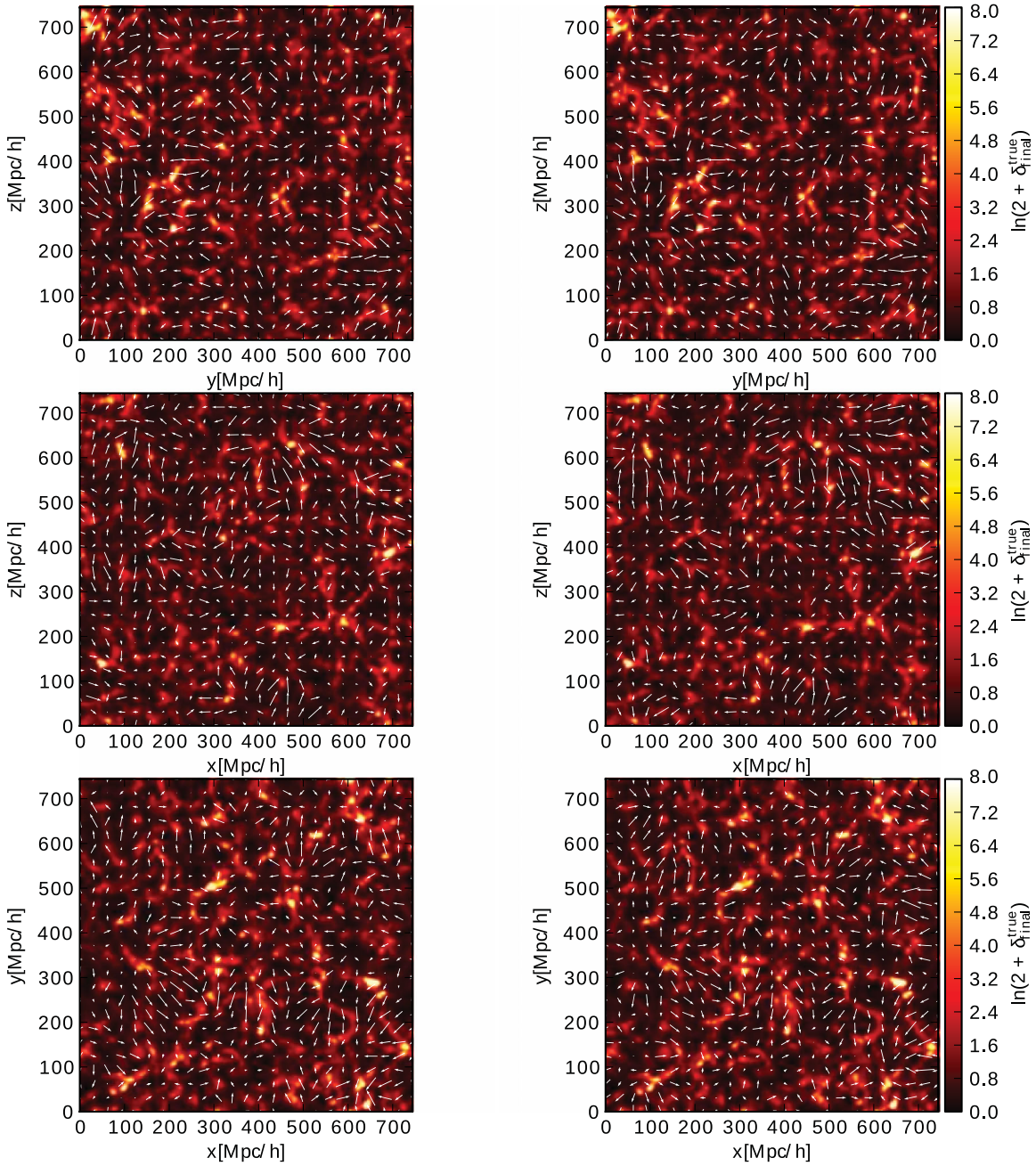

Figure 10. Three slices through the true underlying density field from three different sides overplotted with the two-dimensional projection of the true velocity (left-hand panels) and a sample velocity field (right-hand panels). It can be seen that the algorithm is able to infer the true underlying dynamics of the system to great detail in noisy and even unobserved regions, when compared to the corresponding data panels in Fig. 5.

in the dynamical model. The inference process therefore solely depends on forward propagation of the model, which consequently alleviates many of the problems encountered in previous approaches to the reconstruction of initial conditions, such as spurious decaying mode amplification. Rather than inferring the initial conditions by backward integration in time our approach builds a non-linear filter using the dynamical forward model as a prior. This prior singles out physically reasonable LSS states from the space of all possible solutions.

The resultant inference procedure is numerically highly non-trivial, since the LSS posterior distribution has to be evaluated in very high dimensional space. Typically we are dealing with 10^6–10^7 parameters, corresponding to the voxels used to discretize the domain. In Section 3, we described an efficient HMC implementation for the LSS inference problem when employing a dynamical model for LSS formation. Further, we discussed some details of the numerical implementation in Section 5.

To provide a proof of concept we test the algorithm in an artificial scenario, based on the characteristics of the Sloan Digital Sky Survey Data Release 7. In particular, as described in Section 6, we use the SDSS DR7 completeness map and realistic selection functions based on the Schechter luminosity function to generate a realistic testing environment essentially emulating the SDSS DR7 main sample.

The major aim of testing the algorithm, described in Section 7, was to estimate the method's performance in a realistic scenario. An important issue to test when dealing with MCMC methods is the question of how many independent samples can be drawn from the chain. The high efficiency of our HMC scheme permits to explore the posterior distribution with a typical acceptance rate of about 84 per cent while maintaining the correlation length of the chain at or below 300 steps. We estimate the length of the burn-in phase to be about 600 steps. In summary, our tests reveal that the proposed analysis approach is not only within numerical reach but is efficient enough to work well with present-day computational resources.

The properties of the inferred LSS fields were studied in Section 7.2. It is clear upon visual inspection that our approach returns far more physical reconstructions than previous methods based solely on two-point information (see e.g. Lahav 1994; Zaroubi 2002; Erdoğdu et al. 2004; Kitaura & Enßlin 2008; Kitaura et al. 2009, 2010; Jasche & Kitaura 2010; Jasche et al. 2010a,b). This is particularly obvious for unobserved regions which are augmented with statistically correct information, in order to account for survey geometry and cosmic variance. In the present approach augmented regions are visually indistinguishable from regions containing data. Also note, that measurements of the posterior power spectra do not show any sign of attenuation due to survey geometry or selection effects, indicating that the posterior density fields possess equal power throughout their entire domains. Therefore, the individual density field samples can be regarded as full physical matter field realizations conditional on observations, at least to the degree represented by the 2LPT model.

By studying the one-point distributions of the inferred initial and final density fields we demonstrated that the algorithm correctly recovers the Gaussian initial conditions from a galaxy observation which does not exhibit Gaussian but highly skewed log-normal-like statistics. This demonstrates that the algorithm correctly accounts for the mode coupling and phase correlations originally introduced to the matter distribution by gravitational structure formation. In addition, it supports our initial claim that the approach of searching for phenomenological approximations to the full probability distribution for the non-linear matter field can be efficiently reformulated as an initial condition problem once a physical model for LSS is employed.

In general, the accuracy of inferred initial density fields crucially depends on survey characteristics, such as survey geometry, selection effects and noise. All these quantities are only given at the epoch of observations but not at the initial conditions. Our method therefore has to account for the non-linear and non-local transfer of information from the observations to the initial conditions. Since this is a complex approach, it is important to test the accuracy of inferred initial density fields. To demonstrate that our method is unbiased with respect to the true underlying initial density field we estimated the posterior distribution $\mathcal{P}\left(\Delta\delta_{\rm initial}|\Sigma\right)$ of the deviations of inferred initial conditions from their true values conditional on a signal-to-noise ratio parameter Σ. This test clearly demonstrated that the inferred results are unbiased with respect to the underlying density field, since the distribution is centred on the vertical zero axis. Also, as expected, the accuracy of inferred initial conditions depends on the signal-to-noise ratio value Σ. In particular, inferred density fields in high signal-to-noise ratio regions have higher accuracy compared to regions of lower signal-to-noise ratio values.

To further evaluate the accuracy of recovered density fields, we studied the correlation between the true underlying and samples of the final density field as a function of the signal-to-noise ratios. As expected, the correlation can reach more than 90 per cent in the high signal-to-noise ratio regime, where signal-to-noise ratio $\sim$7. In addition, the algorithm still provides a correlation of about 55 per cent between the true underlying final density field and the samples in regions where only a single galaxy has been observed. Also note that in regions where the signal-to-noise ratio is zero, which are either centres of voids or unobserved regions, the algorithm still provides a 10 per cent correlation. This is a clear manifestation of improved inference due to the incorporation of a physical model of LSS formation, which exploits additionally three-point and higher moment statistics of the density distribution. These tests further demonstrate that the algorithm correctly accounts for systematics such as the survey geometry and selection effects.

We also tested the average accuracy of recovered structures in the initial conditions by estimating the cross-correlation between the inferred ensemble mean initial density field and the true underlying initial density field as a function of filter scale and for various signal-to-noise ratio values. This test revealed that structures in the highest signal-to-noise ratio regimes can be recovered with about 90 per cent accuracy throughout the entire range of scales considered in our analysis. Even for the lowest signal-to-noise ratio regime we still observe about 30 per cent correlation between the true underlying and the inferred ensemble mean initial density field. Also note that on average the largest scales can be recovered with an accuracy of about 90 per cent, even though only roughly half of the analysed volume has been observed.

Along with the inferred density fields the algorithm also provides dynamical information on the large-scale flows. By comparing the true underlying velocity field to the inferred velocity field of an arbitrary sample we demonstrated that the algorithm accurately recovers large-scale flows, even in noisy or even unobserved regions. This feature can be easily explained by the notion that our method accurately infers the large-scale initial density field and the fact that bulk velocities depend predominantly on the largest scales through the gravitational potential. This clearly demonstrates the strength of the method in extrapolating physically reasonable states into poorly observed regions. Nevertheless, it should be remarked that the 2LPT model, as an approximation, does not capture the exact physical behaviour of the actual structure formation processes and thus introduces model errors. Note that since 2LPT and true structure formation both describe deterministic processes, the model error is not stochastic but deterministic. Describing the model error as a stochastic process will decouple the inference process from observations by wiping out information in the data which cannot be accurately represented by the model. In turn, this only means that information on the initial density field should only be extracted in regions where the 2LPT model is applicable, which amounts to large scales or low-density regions. This, however, can also be done in post-processing. Note that our method is generally able to deal with deterministic and stochastic model errors, by exchanging the Dirac delta distribution in equation (2) with the corresponding statistics. We are currently exploring various approaches to deal with deterministic model errors and to improve the accuracy of the 2LPT model in the inference framework. The results of this work will be subject to a future publication.

The method we describe forms the basis for a sophisticated and extensible dynamical LSS inference framework. In future work we will demonstrate the application of the algorithm to a real galaxy survey accounting for additional systematics such as luminosity or colour-dependent bias. Note that the algorithm as described in this work can be easily extended to account for any kind of non-linear and non-local bias. In particular, the 2LPT model, as employed in this work, can already be interpreted as a non-local, non-linear

bias model between the initial conditions and the galaxy observations. It would also be possible to incorporate a halo-model-based galaxy bias model in the fashion as described by Scoccimarro & Sheth (2002). The combination of the algorithm described in this work and the photometric redshift sampling method proposed in Jasche & Wandelt (2012) will lead to immediate improvements for the inferred photometric redshifts, since the combination of both algorithms will exploit higher order statistics, whereas the algorithm described in Jasche & Wandelt (2012) is solely based on two-point statistics. In a similar fashion, dynamical velocity information provided by the 2LPT model can be used to correct for redshift uncertainties in spectroscopic surveys.

Since the algorithm is fully Bayesian, it provides inferred initial and final density fields and also the means of estimating their significance and uncertainties by a sampled representation of the initial conditions posterior distribution. The algorithm will therefore provide accurate information on the initial conditions from which the observed LSS originates. These initial density fields may be useful for a variety of scientific projects such as constrained simulations (see e.g. Kravtsov et al. 2002; Klypin et al. 2003; Dolag et al. 2005, 2012; Martinez-Vaquero et al. 2009; Gottloeber et al. 2010; Lavaux 2010; Libeskind et al. 2010). Since the 2LPT model reconstructs the initial tidal field it may also open up a new way to study the angular momentum build-up of galaxies through tidal torque theory (see e.g. the review by Schäfer 2009, and references therein). Also note that the validity of the 2LPT model improves with increasing redshift. Therefore, the proposed method may also be of interest for density field inference from 21-cm surveys at redshifts of about $z \sim 6$ (see e.g. Lidz et al. 2007).

In conclusion, we presented a new Bayesian dynamical LSS inference algorithm which will provide the community with accurate measurements of the three-dimensional initial density field as well as estimates of the dynamical behaviour of the LSS.

ACKNOWLEDGEMENTS

JJ is partially supported by a Feodor Lynen Fellowship by the Alexander von Humboldt foundation. BDW acknowledges support from NSF grants AST 07-08849 and AST 09-08693 ARRA, and a Chaire d'Excellence from the Agence Nationale de Recherche and computational resources provided through XSEDE grant AST100029.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Baugh C. M., Gaztanaga E., Efstathiou G., 1995, MNRAS, 274, 1049
Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Bertschinger E., Dekel A., Faber S. M., Dressler A., Burstein D., 1990, ApJ, 364, 370
Bouchet F. R., Colombi S., Hivon E., Juszkiewicz R., 1995, A&A, 296, 575
Brenier Y., Frisch U., Hénon M., Loeper G., Matarrese S., Mohayaee R., Sobolevskiĭ A., 2003, MNRAS, 346, 501
Buchert T., 1989, A&A, 223, 9
Buchert T., Melott A. L., Weiss A. G., 1994, A&A, 288, 349
Crocce M., Scoccimarro R., 2006, Phys. Rev. D, 73, 063520
Dekel A., Eldar A., Kolatt T., Yahil A., Willick J. A., Faber S. M., Courteau S., Burstein D., 1999, ApJ, 522, 1
Dolag K., Grasso D., Springel V., Tkachev I., 2005, J. Cosmol. Astropart. Phys., 1, 9
Dolag K., Erdmann M., Müller G., Walz D., Winchen T., 2012, arXiv: e-prints
Doroshkevich A. G., 1970, Afz, 6, 581
Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Phys. Lett. B, 195, 216
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Eisenstein D. J., Hu W., 1999, ApJ, 511, 5
Elsner F., Wandelt B. D., 2012, A&A, 540, L6
Erdoğdu P. et al., 2004, MNRAS, 352, 939
Eriksen H. K. et al., 2004, ApJS, 155, 227
Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23
Frigo M., Johnson S. G., 2005, Proc. IEEE, 93, 216
Frisch U., Matarrese S., Mohayaee R., Sobolevski A., 2002, Nat, 417, 260
Galassi M., Davies J., Theiler J., Gough B., Jungman G., 2009, GNU Scientific Library: Reference Manual, 3 edn., for GSL Version 1.12. Network Theory Ltd, Bristol
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Gottloeber S., Hoffman Y., Yepes G., 2010, arXiv:e-prints
Hanson K. M., 2001, in Sonka M., Hanson K. M., eds, Proc. SPIE Conf. Ser. Vol. 4322, Medical Imaging 2001: Image Processing. SPIE, Bellingham, p. 456
Hastings W. K., 1970, Biometrika, 57, 97
Heitmann K., White M., Wagner C., Habib S., Higdon D., 2010, ApJ, 715, 104
Hockney R. W., Eastwood J. W., 1988, Computer Simulation Using Particles. Taylor & Francis, Bristol, PA
Jasche J., Kitaura F. S., 2010, MNRAS, 407, 29
Jasche J., Wandelt B. D., 2012, MNRAS, 425, 1042
Jasche J., Kitaura F. S., Ensslin T. A., 2009, arXiv:e-prints
Jasche J., Kitaura F. S., Wandelt B. D., Enßlin T. A., 2010a, MNRAS, 406, 60
Jasche J., Kitaura F. S., Li C., Enßlin T. A., 2010b, MNRAS, 409, 355
Jenkins A., 2010, MNRAS, 403, 1859
Jeong D., Komatsu E., 2006, ApJ, 651, 619
Kitaura F.-S., 2012, MNRAS, 420, 2737
Kitaura F. S., Enßlin T. A., 2008, MNRAS, 389, 497
Kitaura F. S., Jasche J., Li C., Enßlin T. A., Metcalf R. B., Wandelt B. D., Lemson G., White S. D. M., 2009, MNRAS, 400, 183
Kitaura F.-S., Jasche J., Metcalf R. B., 2010, MNRAS, 403, 589
Kitaura F.-S., Angulo R. E., Hoffman Y., Gottlöber S., 2012, MNRAS, 425, 2422
Klypin A., Hoffman Y., Kravtsov A. V., Gottlöber S., 2003, ApJ, 596, 19
Komatsu E. et al., 2011, ApJS, 192, 18
Kravtsov A. V., Klypin A., Hoffman Y., 2002, ApJ, 571, 563
Lahav O., 1994, in Balkowski C., Kraan-Korteweg R. C., eds, ASP Conf. Ser. Vol. 67, Unveiling Large-Scale Structures Behind the Milky Way. Astron. Soc. Pac., San Francisco, p. 171
Lavaux G., 2008, Physica D, 237, 2139
Lavaux G., 2010, MNRAS, 406, 1007
Layzer D., 1956, AJ, 61, 383
Libeskind N. I., Yepes G., Knebe A., Gottlöber S., Hoffman Y., Knollmann S. R., 2010, MNRAS, 401, 1889
Lidz A., Zahn O., McQuinn M., Zaldarriaga M., Dutta S., Hernquist L., 2007, ApJ, 659, 865
Linde A., 2008, in Lemoine M., Martin J., Peter P., eds, Lecture Notes in Physics, Vol. 738, Inflationary Cosmology. Springer-Verlag, Berlin, p. 1
Linder E. V., Jenkins A., 2003, MNRAS, 346, 573
Martínez V. J., Saar E., 2002, Statistics of the Galaxy Distribution. Chapman and Hall/CRC Press, Boca Raton, FL
Martinez-Vaquero L. A., Yepes G., Hoffman Y., Gottlöber S., Sivan M., 2009, MNRAS, 397, 2070
Matsumoto M., Nishimura T., 1998, ACM Trans. Model. Comput. Simulation, 8, 3
Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E. T., 1953, J. Chem. Phys., 21, 1087
Mohayaee R., Sobolevskiĭ A., 2008, Physica D, 237, 2145

Moutarde F., Alimi J., Bouchet F. R., Pellat R., Ramani A., 1991, ApJ, 382, 377
Neal R. M., 1993, Technical Report CRG-TR-93-1, Probabilistic Inference Using Markov Chain Monte Carlo Methods. University of Toronto
Neal R. M., 1996, Bayesian Learning for Neural Networks, 1st edn. Springer-Verlag, Berlin
Nusser A., Dekel A., 1992, ApJ, 391, 443
Peacock J., 1999, Cosmological Physics. Cambridge Univ. Press, Cambridge
Peacock J. A., Dodds S. J., 1996, MNRAS, 280, L19
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe, Princeton Univ. Press, Princeton, NJ
Percival W. J., 2005, MNRAS, 356, 1168
Schäfer B. M., 2009, Int. J. Modern Phys. D, 18, 173
Schechter P., 1976, ApJ, 203, 297
Scoccimarro R., 1998, MNRAS, 299, 1097
Scoccimarro R., 2000, ApJ, 544, 597
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Smith R. E. et al., 2003, MNRAS, 341, 1311
Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, MNRAS, 387, 1391
Tatekawa T., Mizuno S., 2007, J. Cosmol. Astropart. Phys., 12, 14
Taylor J. F., Ashdown M. A. J., Hobson M. P., 2008, MNRAS, 389, 1284
Tegmark M. et al., 2004, ApJ, 606, 702
Turner M. S., White M., 1997, Phys. Rev. D, 56, 4439
Wandelt B. D., Larson D. L., Lakshminarayanan A., 2004, Phys. Rev. D, 70, 083511
Wang L., Steinhardt P. J., 1998, ApJ, 508, 483
Zaroubi S., 2002, MNRAS, 331, 901
Zel'Dovich Y. B., 1970, A&A, 5, 84

APPENDIX A: LINEAR STRUCTURE FORMATION

In the linear regime structure formation is governed by a homogeneous growth function $D^+(a)$ acting on the density contrast $\delta(\boldsymbol{x}, a) = D^+(a)\,\delta(\boldsymbol{x}, a = 1)$. For a general cosmology the growth factor $D^+(a)$ can be obtained by numerical solution of the linear growth equation (see e.g. Turner & White 1997; Wang & Steinhardt 1998; Linder & Jenkins 2003):

$$\frac{\mathrm{d}^2 D^+(a)}{\mathrm{d}a^2} + \frac{1}{a}\left(3 + \frac{\mathrm{d}\ln H}{\mathrm{d}\ln a}\right)\frac{\mathrm{d}D^+(a)}{\mathrm{d}a} - \frac{3}{2}\frac{\Omega_\mathrm{m}(a) D^+(a)}{a^2} = 0. \quad \text{(A1)}$$

APPENDIX B: LAGRANGIAN PERTURBATION THEORY

In the following we will give a brief summary of 2LPT to the degree required for the present work. More detailed discussion of Lagrangian perturbation theory in general and its application can be found in the literature (see e.g. Moutarde et al. 1991; Buchert et al. 1994; Bouchet et al. 1995; Scoccimarro 1998, 2000; Bernardeau et al. 2002; Scoccimarro & Sheth 2002). Also see Bernardeau et al. (2002) for a general overview of Eulerian and Lagrangian cosmological perturbation theory.

In an expanding Robertson–Friedman space–time the equations of motion for particles solely interacting through gravity are given as (see e.g. Scoccimarro 2000; Bernardeau et al. 2002)

$$\frac{\mathrm{d}^2 \boldsymbol{x}}{\mathrm{d}\tau^2} + \mathcal{H}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau} - \nabla_x \phi = 0, \quad \text{(B1)}$$

where ϕ is the gravitational potential and ∇_x is the gradient with respect to the Eulerian coordinates $\boldsymbol{x}$, $\mathcal{H} = \mathrm{d}\ln \mathrm{a}/\mathrm{d}\tau$ and the conformal time τ defined by $\mathrm{d}\tau = \mathrm{d}t/a$. In order to solve this set of equations, Lagrangian perturbation theory introduces the following ansatz for a solution:

$$\boldsymbol{x}(\tau) = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}, \tau), \quad \text{(B2)}$$

where $\boldsymbol{\Psi}(\boldsymbol{q}, \tau)$ defines the mapping from the particles initial position $\boldsymbol{q}$ into its final Eulerian position $\boldsymbol{x}$ (see e.g. Scoccimarro 2000; Bernardeau et al. 2002). Equation (B2) together with equation (B1) yields a non-linear equation for the displacement field $\boldsymbol{\Psi}(\boldsymbol{q}, \tau)$ which can be solved perturbatively by expanding around its linear solution (Bernardeau et al. 2002). To linear order, this perturbative approach yields the famous Zel'dovich approximation given as (Doroshkevich 1970; Zel'Dovich 1970; Buchert 1989; Moutarde et al. 1991; Bernardeau et al. 2002)

$$\nabla_q \Psi^{(1)}(\boldsymbol{q}, a) = -D^+(a)\,\delta\,(\boldsymbol{q}, a = 1)\,. \quad \text{(B3)}$$

Adding second-order terms to the perturbative expansion remarkably improves the results of the first order Zel'dovich approximation. In particular, second-order terms account for the fact that gravitational instability is non-local by introducing corrections due to gravitational tidal effects (Bernardeau et al. 2002). The second-order displacement field $\Psi^{(2)}(\boldsymbol{q}, a)$ is then defined by (see e.g. Bernardeau et al. 2002; Scoccimarro & Sheth 2002)

$$\nabla_q \Psi^{(2)}(\boldsymbol{q}, a) = \frac{1}{2} D_2(a) \sum_{i \neq j} \left(\Psi^{(1)}_{i,i}\,\Psi^{(1)}_{j,j} - \Psi^{(1)}_{i,j}\,\Psi^{(1)}_{j,i}\right), \quad \text{(B4)}$$

with $\Psi^{(1)}_{i,j} \equiv \partial \Psi^{(1)}_i / \partial \boldsymbol{q}_j$ and $D_2(a)$ is the second-order growth factor given as

$$D_2(a) \approx -\frac{3}{7}\left(D^+(a)\right)^2\,\Omega_\mathrm{m}^{-(1/143)}, \quad \text{(B5)}$$

which holds for a flat model with non-zero cosmological constant Λ and for $0.01 \leq \Omega_\mathrm{m} \leq 1$ to better than 0.6 per cent accuracy (see e.g. Bouchet et al. 1995; Scoccimarro 1998; Bernardeau et al. 2002, for details).

As has been previously shown, 2LPT recovers correctly the two- and three-point statistics at large scales and further approximates higher order statistics very well (Moutarde et al. 1991; Buchert et al. 1994; Bouchet et al. 1995; Scoccimarro 2000; Scoccimarro & Sheth 2002). Also note that second-order corrections to the Zel'dovich approximation are essential to accurately describe the departure of the large-scale density field from Gaussian initial conditions (Scoccimarro & Sheth 2002; Tatekawa & Mizuno 2007; Jenkins 2010).

Lagrangian solutions up to second order are curl-free, as they follow potential flows (see e.g. Buchert et al. 1994; Scoccimarro 1998; Bernardeau et al. 2002). Therefore, it is convenient to introduce the Lagrangian potentials $\Phi^{(1)}$ and $\Phi^{(2)}$, such that the approximate solution to equation (B1) can be expressed as (see e.g. Buchert et al. 1994; Scoccimarro 1998; Bernardeau et al. 2002)

$$\boldsymbol{x}(\tau) = \boldsymbol{q} - D^+(a)\nabla_q\,\Phi^{(1)} + D_2(a)\nabla_q\,\Phi^{(2)}, \quad \text{(B6)}$$

where the time-independent potentials $\Phi^{(1)}$ and $\Phi^{(2)}$ are solutions to the following Poisson equations (Buchert et al. 1994):

$$\nabla_q^2\,\Phi^{(1)}(\boldsymbol{q}) = \delta\,(\boldsymbol{q}, a = 1) \quad \text{(B7)}$$

and

$$\nabla_q^2\,\Phi^{(2)}(\boldsymbol{q}) = \sum_{i>j}\left[\Phi^{(1)}_{,ii}(\boldsymbol{q})\,\Phi^{(1)}_{,jj}(\boldsymbol{q}) - \left(\Phi^{(1)}_{,ij}(\boldsymbol{q})\right)^2\right]. \quad \text{(B8)}$$

For an excellent guide to the numerical implementation of the 2LPT model the reader is referred to appendix D of Scoccimarro (1998).

APPENDIX C: THE SCHECHTER LUMINOSITY FUNCTION

The Schechter luminosity function is given as (Schechter 1976)

$$\Phi(M)\,\mathrm{d}M = 0.4\,\Phi^{*}\ln(10)\,\left(10^{0.4\,(M^{*}-M)}\right)^{\alpha+1}\,\mathrm{e}^{-10^{0.4\,(M^{*}-M)}}\,\mathrm{d}M. \tag{C1}$$

Note that for the purpose of calculating selection functions the normalization Φ^{*} is not required.

APPENDIX D: HAMILTONIAN FORCES FOR THE LIKELIHOOD TERM

In this section we will discuss the derivation of the Hamiltonian forces for the 2lpt–Poissonian process. To prevent confusion between the variables describing the physical 2LPT model and the variables describing the Hamiltonian inference framework we will re express the 2LPT model in the following form for the purpose of the derivations in this section:

$$\boldsymbol{x}_p = \boldsymbol{x}_p(\delta^i) = \boldsymbol{q}_p - D^1\,\boldsymbol{K}^1_p(\delta^i) + D_2\,\boldsymbol{K}^2_p(\delta^i), \tag{D1}$$

where $\boldsymbol{K}^1_p(\delta^i)$ and $\boldsymbol{K}^2_p(\delta^i)$ are the first- and second-order displacements fields, respectively.

As described in Section 4 the likelihood term of the Hamiltonian potential is given as

$$\Psi_{\mathrm{likelihood}}(\{\delta^i_j\}) = \sum_l R_l\bar{N}_{\mathrm{gal}}\,\left(1+G(a,\delta^i)_l\right) - N_l\ln\left(R_l\bar{N}_{\mathrm{gal}}\,(1+G(a,\delta^i)_l)\right), \tag{D2}$$

with $G(a,s)$ given via the kernel estimate as

$$G(a,s)_l = \sum_p \frac{W(\boldsymbol{x}_p(a,s)-\boldsymbol{x}_l)}{\bar{N}} - 1, \tag{D3}$$

and $\boldsymbol{x}_p(a,s)$ is described by equation (D1) and $W(\boldsymbol{x})$ is a CIC kernel (see e.g. Hockney & Eastwood 1988; Jasche, Kitaura & Ensslin 2009). Furthermore, the Lagrangian displacement vectors are given as

$$K^n_p = \sum_j \boldsymbol{V}'(\boldsymbol{q}_p-\boldsymbol{x}_j)\Phi^n_j, \tag{D4}$$

where $\boldsymbol{V}'(\boldsymbol{x})$ is the gradient of the kernel $W(\boldsymbol{x})$. With these definitions we can write the Hamiltonian forces corresponding to the likelihood term as

$$\frac{\partial\Psi_{\mathrm{likelihood}}(\{\delta^i_j\})}{\partial s_m} = \sum_i\left(1-\frac{1}{R_i\,\bar{N}\,(1+G(a,\delta^i)_l)}\right)\times R_i\,\bar{N}\frac{\partial G(a,\delta^i)_i)}{\partial\delta^i_m}. \tag{D5}$$

The notation can be simplified by introduce the quantity A_i as

$$A_i = \left(1-\frac{1}{R_i\,\bar{N}\,(1+\delta_i(s))}\right)R_i\,\bar{N}. \tag{D6}$$

We can then write

$$\begin{aligned}\frac{\partial\Psi_{\mathrm{likelihood}}(\{\delta^i_j\})}{\partial\delta^i_m} &= \sum_i A_i\frac{\partial G(a,\delta^i)_i}{\partial\delta^i_m}\\ &= \sum_i\frac{A_i}{\bar{N}}\sum_p\frac{\partial W(\boldsymbol{x}_p-\boldsymbol{x}_i)}{\partial\delta^i_m}\\ &= \sum_i\frac{A_i}{\bar{N}}\sum_p\boldsymbol{W}'(\boldsymbol{x}_p-\boldsymbol{x}_i)\frac{\partial\boldsymbol{x}_p}{\partial\delta^i_m}\\ &= \sum_i\frac{A_i}{\bar{N}}\sum_p\boldsymbol{W}'(\boldsymbol{x}_p-\boldsymbol{x}_i)\left(-D^1\,\frac{\partial K^1_p(\delta^i)}{\partial\delta^i_m} + D^2\,\frac{\partial K^2_p(\delta^i)}{\partial\delta^i_m}\right),\end{aligned} \tag{D7}$$

where we made use of equations (D1) and (D4). It can be seen that the Hamiltonian force is the sum of two vectors. In the following we will therefore discuss each term independently. The first term is exactly the Hamiltonian force expected from a pure Zel'dovich approximation without higher order correction terms. We will start by evaluating the Hamiltonian force for the Zel'dovich approximation:

$$\begin{aligned}\frac{\partial\Psi^1_{\mathrm{likelihood}}(\{\delta^i_j\})}{\partial\delta^i_m} &= \sum_i\frac{-D^1\,A_i}{\bar{N}}\sum_p\boldsymbol{W}'(\boldsymbol{x}_p-\boldsymbol{x}_i)\left(\frac{\partial\boldsymbol{K}^1_p(\delta^i)}{\partial\delta^i_m}\right)\\ &= \sum_p\sum_i\frac{-D^1\,A_i}{\bar{N}}\boldsymbol{W}'(\boldsymbol{x}_p-\boldsymbol{x}_i)\\ &\quad\times\sum_j\boldsymbol{V}'(\boldsymbol{q}_p-\boldsymbol{x}_j)\frac{\Phi^1_j}{\partial\delta^i_m}\\ &= \sum_j\sum_p\sum_i\frac{-D^1\,A_i}{\bar{N}}\boldsymbol{W}'(\boldsymbol{x}_p-\boldsymbol{x}_i)\\ &\quad\times\boldsymbol{V}'(\boldsymbol{q}_p-\boldsymbol{x}_j)\frac{\Phi^1_j}{\partial\delta^i_m}.\end{aligned} \tag{D8}$$

The notation can be further simplified by introducing

$$F_j = \sum_p\sum_i\frac{A_i}{\bar{N}}\boldsymbol{W}'(\boldsymbol{x}_p-\boldsymbol{x}_i)\boldsymbol{V}'(\boldsymbol{q}_p-\boldsymbol{x}_j). \tag{D9}$$

We can then write

$$\frac{\partial\Psi^1_{\mathrm{likelihood}}(\{\delta^i_j\})}{\partial\delta^i_m} = -D^1\sum_j F_j\frac{\Phi^1_j}{\partial\delta^i_m}. \tag{D10}$$

The Zel'dovich approximation potential was calculated using the FFT approach, which can be written as

$$\Phi^1_j = \sum_k\frac{-1}{k^2_k}\,\mathrm{e}^{2\pi\,j\,k\frac{\sqrt{-1}}{N}}\sum_n s_n\,\mathrm{e}^{-2\pi n\,k\frac{\sqrt{-1}}{N}}. \tag{D11}$$

Using this expression in equation (D8) we yield

$$\begin{aligned}\frac{\partial\Psi^1_{\mathrm{likelihood}}(\{\delta^i_j\})}{\partial\delta^i_m} &= -D^1\sum_j F_j\sum_k\frac{-1}{k^2_k}\,\mathrm{e}^{2\pi\,j\,k\frac{\sqrt{-1}}{N}}\sum_n\delta^K_{nm}\,\mathrm{e}^{-2\pi n\,k\frac{\sqrt{-1}}{N}}\\ &= -D^1\sum_j F_j\sum_k\frac{-1}{k^2_k}\,\mathrm{e}^{2\pi\,j\,k\frac{\sqrt{-1}}{N}}\,\mathrm{e}^{-2\pi m\,k\frac{\sqrt{-1}}{N}}\\ &= -D^1\sum_k\frac{-1}{k^2_k}\,\mathrm{e}^{-2\pi m\,k\frac{\sqrt{-1}}{N}}\sum_j F_j\,\mathrm{e}^{2\pi\,j\,k\frac{\sqrt{-1}}{N}}.\end{aligned} \tag{D12}$$

This result looks remarkably similar to equation (D18) and at first sight one might be inclined to straightforwardly solve this equation with FFT techniques. However, it is important to note that the signs have changed in the exponents, and hence equation (D8) cannot directly be solved with FFTs. In Appendix E, we show what procedures must be followed in order to apply FFTs to this problem. To further simplify the notation in the following steps we will introduce the quantity J_m, defined as

$$J_m = \sum_k\frac{-1}{k^2_k}\,\mathrm{e}^{-2\pi m\,k\frac{\sqrt{-1}}{N}}\sum_j F_j\,\mathrm{e}^{2\pi\,j\,k\frac{\sqrt{-1}}{N}}. \tag{D13}$$

With this definition the Zel'dovich approximation term of the Hamiltonian force can be written as

$$\frac{\partial \Psi^1_{\rm likelihood}(\{\delta^i_j\})}{\partial \delta^i_m} = -D^1\, J_m. \tag{D14}$$

Next, we will discuss the second-order Lagrangian term in equation (D7):

$$\begin{aligned}\frac{\partial \Psi^2_{\rm likelihood}(\{\delta^i_j\})}{\partial \delta^i_m} &= \sum_i \frac{D^2\, A_i}{\bar{N}} \sum_p \boldsymbol{W}'(\boldsymbol{x}_p - \boldsymbol{x}_i) \left(\frac{\partial \boldsymbol{K}^2_p(\delta^i)}{\partial \delta^i_m} \right) \\ &= \sum_p \sum_i \frac{D^2\, A_i}{\bar{N}} \boldsymbol{W}'(\boldsymbol{x}_p - \boldsymbol{x}_i) \\ &\quad \times \sum_j \boldsymbol{V}'(\boldsymbol{q}_p - \boldsymbol{x}_j) \frac{\Phi^2_j}{\partial \delta^i_m} \\ &= D^2 \sum_j F_j \frac{\Phi^2_j}{\partial \delta^i_m}. \end{aligned} \tag{D15}$$

The second-order Lagrangian potential Φ^2_j can be calculated as

$$\Phi^2_j = \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \phi_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}}, \tag{D16}$$

with ϕ_n given as

$$\phi_n = \sum_{a>b} \phi^{aa}_n\, \phi^{bb}_n - \left(\phi^{ab}_n\right)^2, \tag{D17}$$

where the individual potentials ϕ^{ab}_n are related to the signal δ^i_n via

$$\phi^{ab}_n = \sum_k \frac{k^a_k\, k^b_k}{k^2_k} \mathrm{e}^{2\pi n\, k \frac{\sqrt{-1}}{N}} \sum_l \delta^i_l \mathrm{e}^{-2\pi l\, k \frac{\sqrt{-1}}{N}}. \tag{D18}$$

With these definitions, we can write

$$\begin{aligned}\frac{\partial \psi^2_{\rm LH}(\delta^i)}{\partial \delta^i_m} &= D^2 \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \frac{\partial \phi_n}{\partial \delta^i_m} \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \\ &= D^2 \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \\ &\quad \times \frac{\partial}{\partial \delta^i_m} \left(\sum_{a>b} \phi^{aa}_n\, \phi^{bb}_n - \left(\phi^{ab}_n\right)^2 \right) \\ &= D^2 \sum_{a>b} \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \\ &\quad \times \frac{\partial}{\partial \delta^i_m} \left(\phi^{aa}_n\, \phi^{bb}_n - \left(\phi^{ab}_n\right)^2 \right) \\ &= D^2 \sum_{a>b} \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \\ &\quad \times \left(\frac{\partial \phi^{aa}_n}{\partial \delta^i_m} \phi^{bb}_n + \frac{\partial \phi^{bb}_n}{\partial \delta^i_m} \phi^{aa}_n - 2\phi^{ab}_n \frac{\partial \phi^{ab}_n}{\partial \delta^i_m} \right). \end{aligned} \tag{D19}$$

In the following we will discuss the individual terms. To simplify the notation, we introduce the tensor $\boldsymbol{\tau}^{abcd}$ defined as

$$\begin{aligned}\boldsymbol{\tau}^{abcd}_m &= \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \frac{\partial \phi^{ab}_n}{\partial \delta^i_m} \phi^{cd}_n \\ &= \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \phi^{cd}_n \\ &\quad \times \sum_p \frac{k^a_p\, k^b_p}{k^2_p} \mathrm{e}^{2\pi n\, p \frac{\sqrt{-1}}{N}} \sum_l \delta^K_{lm} \mathrm{e}^{-2\pi l\, p \frac{\sqrt{-1}}{N}} \\ &= \sum_j F_j \sum_k \frac{-1}{k^2_k} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \phi^{cd}_n \\ &\quad \times \sum_p \frac{k^a_p\, k^b_p}{k^2_p} \mathrm{e}^{2\pi n\, p \frac{\sqrt{-1}}{N}} \mathrm{e}^{-2\pi m\, p \frac{\sqrt{-1}}{N}} \\ &= \sum_p \frac{k^a_p\, k^b_p}{k^2_p} \mathrm{e}^{-2\pi m\, p \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{2\pi n\, p \frac{\sqrt{-1}}{N}} \phi^{cd}_n \\ &\quad \times \sum_k \mathrm{e}^{-2\pi n\, k \frac{\sqrt{-1}}{N}} \frac{-1}{k^2_k} \sum_j F_j \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \\ &= \sum_p \frac{k^a_p\, k^b_p}{k^2_p} \mathrm{e}^{-2\pi m\, p \frac{\sqrt{-1}}{N}} \sum_n \mathrm{e}^{2\pi n\, p \frac{\sqrt{-1}}{N}} \phi^{cd}_n\, J_n. \end{aligned} \tag{D20}$$

With these definitions the second-order Lagrangian contribution to the Hamiltonian force can be calculated as

$$\frac{\partial \psi^2_{\rm LH}(s)}{\partial \delta^i_m} = D^2 \sum_{a>b} \left(\boldsymbol{\tau}^{aabb}_m + \boldsymbol{\tau}^{bbaa}_m - 2\boldsymbol{\tau}^{abab}_m \right). \tag{D21}$$

This finally yields the Hamiltonian forces corresponding to the likelihood term

$$\frac{\partial \Psi_{\rm likelihood}(\{\delta^i_j\})}{\partial \delta^i_m} = -D^1\, J_m + D^2 \sum_{a>b} \left(\boldsymbol{\tau}^{aabb} + \boldsymbol{\tau}^{bbaa} - 2\boldsymbol{\tau}^{abab} \right). \tag{D22}$$

APPENDIX E: ADJOINT FFT

The following operation can be performed via FFT methods, when accounting for adjoining the operation:

$$\begin{aligned}\sum_j a_j \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} &= \sum_j \sum_q \hat{a}_q \mathrm{e}^{2\pi j\, q \frac{\sqrt{-1}}{N}} \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \\ &= \sum_q \hat{a}_q \sum_j \mathrm{e}^{2\pi j\, (q+k) \frac{\sqrt{-1}}{N}} \\ &= \sum_q \hat{a}_q\, \delta^K q, -k \\ &= \hat{a}_{-k} \\ &= \hat{a}^*_k, \end{aligned} \tag{E1}$$

where we made use of the fact that a_j is a real quantity, and the * denotes complex conjugation. Therefore, equation (E1) simply describes the application of an FFT followed by a complex conjugation. To solve the adjoint Poisson equation we calculate

$$\begin{aligned}\sum_k \frac{\hat{a}^*_k}{k^2_k} \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} &= \sum_k \hat{b}_k \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} \\ &= \sum_k \sum_j b_j \mathrm{e}^{-2\pi j\, k \frac{\sqrt{-1}}{N}} \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} \\ &= \sum_j b_j \sum_k \mathrm{e}^{-2\pi (j+m)\, k \frac{\sqrt{-1}}{N}} \\ &= \sum_j b_j \delta^K_{j,-m} \\ &= b_{-m} \\ &= b_{N-m}, \end{aligned} \tag{E2}$$

where in the last step we made use of the periodicity of the signal.

APPENDIX F: HAMILTONIAN MASSES

A good guess for the Hamiltonian masses can greatly improve the efficiency of the hybrid Hamiltonian sampler. In order to derive appropriate Hamiltonian masses for the 2LPT–Poissonian system we will follow a similar approach as described in Taylor et al. (2008) and Jasche & Kitaura (2010). Since the efficiency of the Hamiltonian sampler depends on the accuracy of the leapfrog scheme, we will perform an approximated stability analysis of the integrator. The goal of this analysis is to find an expression for the Hamiltonian masses which optimizes the stability of the integration scheme for the 2LPT–Poissonian system.

According to the leapfrog scheme, given in equations (23)–(25), a single application of the leapfrog method can be written in the form

$$p_m(t+\epsilon) = p_m(t) - \frac{\epsilon}{2}\left(\left.\frac{\partial \Psi(\delta^i)}{\partial \delta^i_i}\right|_{\delta^i(t)} + \left.\frac{\partial \Psi(\delta^i)}{\partial \delta^i_m}\right|_{\delta^i(t+\epsilon)} \right), \quad \text{(F1)}$$

$$s_m(t+\epsilon) = s_m(t) + \epsilon \sum_j \mathbf{M}^{-1}_{mj}\, p_j(t) - \frac{\epsilon^2}{2}\sum_j \mathbf{M}^{-1}_{mj}\, \left.\frac{\partial \Psi(\delta^i)}{\partial \delta^i_j}\right|_{\delta^i(t)}. \quad \text{(F2)}$$

We will then expand the Hamiltonian forces given in equation (18) around a fixed value $(\delta^i)^0_m$, which is assumed to be the mean signal around which the sampler will oscillate once it left the burn-in phase. Further, we will only expand up to linear order in the forces, which amounts to second order in the potential and hence to a Gaussian approximation of the 2LPT–Poissonian posterior distribution. For simplicity we will also ignore the second-order Lagrangian term in the forces. Thus, the Hamiltonian forces can be written as

$$\begin{aligned}
\frac{\partial \Psi(\{\delta^i_j\})}{\partial \delta^i_m} &= \frac{\partial \Psi_{\rm prior}(\{\delta^i_i\})}{\partial \delta^i_m} + \frac{\partial \Psi_{\rm likelihood}(\{\delta^i_i\})}{\partial \delta^i_m} \\
&= \sum_j \mathbf{S}^{-1}_{mj}\, \delta^i_j - D^1\, J_m \\
&\approx \sum_j \mathbf{S}^{-1}_{mj}\, \delta^i_j \\
&\quad - D^1 \left(J_m((\delta^i)^0) + \left.\frac{\partial J_m(\delta^i)}{\partial \delta^i_m}\right|_{\delta^i_m=(\delta^i)^0_m} (\delta^i_m - (\delta^i)^0_m) \right) \\
&= \sum_j \left(\mathbf{S}^{-1}_{mj} - \delta^K_{mj}\, D^1\, \left.\frac{\partial J_j(\delta^i)}{\partial \delta^i_j}\right|_{\delta^i_j=(\delta^i)^0_j} \right) \delta^i_j \\
&\quad - D^1 \left(J_m((\delta^i)^0) - \left.\frac{\partial J_m(\delta^i)}{\partial \delta^i_m}\right|_{\delta^i_m=(\delta^i)^0_m} (\delta^i)^0_m \right). \quad \text{(F3)}
\end{aligned}$$

We will simplify the notation by introducing the matrix,

$$\mathbf{A}_{mj} = \mathbf{S}^{-1}_{mj} - \delta^K_{mj}\, D^1\, \left.\frac{\partial J_j(\delta^i)}{\partial \delta^i_j}\right|_{\delta^i_j=(\delta^i)^0_j}, \quad \text{(F4)}$$

and the vector,

$$D_m = -D^1 \left(J_m((\delta^i)^0) - \left.\frac{\partial J_m(\delta^i)}{\partial s_m}\right|_{\delta^i_m=(\delta^i)^0_m} (\delta^i)^0_m \right). \quad \text{(F5)}$$

Equation (F3) can then be written as

$$\frac{\partial \Psi(\{\delta^i_j\})}{\partial \delta^i_m} = \sum_j \mathbf{A}_{mj}\, \delta^i_j + D_m. \quad \text{(F6)}$$

Introducing this approximation into equations (F1) and (F2) yields

$$\begin{aligned}
p_i(t+\epsilon) = &\sum_m \left[\delta^K_{im} - \frac{\epsilon^2}{2}\sum_j \mathbf{A}_{ij}\mathbf{M}^{-1}_{jm} \right] p_m(t) \\
&-\epsilon \sum_j \mathbf{A}_{ij} \sum_p \left[\delta^K_{jp} - \frac{\epsilon^2}{4}\sum_m \mathbf{M}^{-1}_{jm}\mathbf{A}_{mp} \right] r_p(t) \\
&-\frac{\epsilon}{2}\sum_m \left[\delta^K_{im} - \frac{\epsilon^2}{2}\sum_j \mathbf{A}_{ij}\mathbf{M}^{-1}_{jm} \right] D_m \quad \text{(F7)}
\end{aligned}$$

and

$$\begin{aligned}
r_i(t+\epsilon) = &\epsilon \sum_j \mathbf{M}^{-1}_{ij}\, p_j(t) \\
&+\sum_m \left(\delta^K_{im} - \frac{\epsilon^2}{2}\sum_j \mathbf{M}^{-1}_{ij}\mathbf{A}_{jm} \right) r_m(t) \\
&-\frac{\epsilon^2}{2}\sum_j \mathbf{M}^{-1}_{ij}\mathbf{D}_j. \quad \text{(F8)}
\end{aligned}$$

This result can be rewritten in matrix notation as

$$\begin{pmatrix} r(t+\epsilon) \\ p(t+\epsilon) \end{pmatrix} = \mathbf{T} \begin{pmatrix} r(t) \\ p(t) \end{pmatrix} - \frac{\epsilon^2}{2} \begin{pmatrix} \mathbf{M}^{-1}\mathbf{D} \\ \epsilon \left[\mathbf{I} - \frac{\epsilon^2}{2}\mathbf{A}\,\mathbf{M}^{-1} \right] \mathbf{D} \end{pmatrix}, \quad \text{(F9)}$$

where the matrix $\mathbf{T}$ is given as

$$\mathbf{T} = \begin{pmatrix} \left[\mathbf{I} - \frac{\epsilon^2}{2}\mathbf{M}^{-1}\mathbf{A} \right] & \epsilon \mathbf{M}^{-1} \\ -\epsilon\, \mathbf{A} \left[\mathrm{I} - \frac{\epsilon^2}{4}\mathbf{M}^{-1}\,\mathbf{A} \right] & \left[\mathbf{I} - \frac{\epsilon^2}{2}\mathbf{A}\,\mathbf{M}^{-1} \right] \end{pmatrix}, \quad \text{(F10)}$$

with $\mathbf{I}$ being the identity matrix. Successive applications of the leapfrog step yield the following propagation equation:

$$\begin{pmatrix} r^n \\ p^n \end{pmatrix} = \mathbf{T}^n \begin{pmatrix} r^0 \\ p^0 \end{pmatrix} - \frac{\epsilon^2}{2} \left[\sum_{i=0}^{n-1} \mathbf{T}^i \right] \begin{pmatrix} \mathbf{M}^{-1}\mathbf{D} \\ \epsilon \left[\mathbf{I} - \frac{\epsilon^2}{2}\mathbf{A}\,\mathbf{M}^{-1} \right] \mathbf{D} \end{pmatrix}. \quad \text{(F11)}$$

This equation demonstrates that there are two criteria to be fulfilled if the method is to be stable under repeated application of the leapfrog step. First we have to ensure that the first term of equation (F11) does not diverge. This can be fulfilled if the eigenvalues of $\mathbf{T}$ have unit modulus. The eigenvalues λ are found by solving the characteristic equation

$$\det \left[\mathbf{I}\,\lambda^2 - 2\,\lambda \left(\mathrm{I} - \frac{\epsilon^2}{2}\mathbf{A}\,\mathbf{M}^{-1} \right) + \mathbf{I} \right] = 0. \quad \text{(F12)}$$

Note that this is a similar result to what was found in Taylor et al. (2008). Our aim is to explore the parameter space rapidly, and therefore we wish to choose the largest ϵ still compatible with the stability criterion. However, any dependence of equation (F12) also implies that no single value of ϵ will ensure unit modulus for every eigenvalue. For this reason we choose

$$\mathbf{A} = \mathbf{M}. \quad \text{(F13)}$$

We then obtain the characteristic equation:

$$\left[\lambda^2 - 2\,\lambda \left(1 - \frac{\epsilon^2}{2} \right) + 1 \right]^N = 0, \quad \text{(F14)}$$

where N is the number of voxels. This yields the eigenvalues

$$\lambda = \pm i \sqrt{1 - \left[1 - \frac{\epsilon^2}{2}\right]^2} + \left[1 - \frac{\epsilon^2}{2}\right], \tag{F15}$$

which have unit modulus for $\epsilon \leq 2$. The second term in equation (F11) involves evaluation of the geometric series $\sum_{i=0}^{n-1} \mathbf{T}^i$. The geometric series for a matrix converges if and only if $|\lambda_i| < 1$ for each λ_i eigenvalue of $\mathbf{T}$. This clarifies that the non-linearities in the Hamiltonian equations generally do not allow for arbitrary large pseudo-time-steps ϵ. In addition, for practical purposes we usually restrict the mass matrix to the diagonal of equation (F4). In practice we choose the pseudo-time-step ϵ as large as possible while still obtaining a reasonable rejection rate.

Given these assumptions we can assume the mass matrix to be

$$\mathbf{M}_{mj} = \mathbf{S}^{-1}_{mj} - \delta^K_{mj}\, D^1 \left.\frac{\partial J_j(\delta^i)}{\partial \delta^i_j}\right|_{\delta^i_j = (\delta^i)^0_j}, \tag{F16}$$

where

$$\begin{aligned}
\frac{\partial J_m(s)}{\partial s_m} &= \sum_k \frac{-1}{k_k^2}\, \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} \sum_j \frac{\partial F_j}{\partial \delta^i_m}\, \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \\
&= \sum_k \frac{-1}{k_k^2}\, \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} \sum_j \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \\
&\quad \times \sum_p \sum_i \left(\frac{1}{\bar{N}} \boldsymbol{W}'(\boldsymbol{x}_p - \boldsymbol{x}_i) \boldsymbol{V}'(\boldsymbol{q}_p - \boldsymbol{x}_j) \frac{\partial A_i}{\partial \delta^i_m} \right. \\
&\quad \left. \times \frac{1}{\bar{N}} \boldsymbol{W}''(\boldsymbol{x}_p - \boldsymbol{x}_i) \boldsymbol{V}'(\boldsymbol{q}_p - \boldsymbol{x}_j)\, A_i \frac{\partial x_p}{\partial \delta^i_m} \right),
\end{aligned} \tag{F17}$$

where we used of equations (D9) and (D13). According to equation (D6) $\frac{\partial A_i}{\partial \delta^i_m}$ can be expressed as

$$\begin{aligned}
\frac{\partial A_i}{\partial s_m} &= \frac{R_i\, \bar{N}}{\left(R_i\, \bar{N}\, (1 + G(a, \delta^i)_i)\right)^2} \frac{\partial \delta_i(\delta^i)}{\partial \delta^i_m} \\
&= B_i\, \frac{\partial G(a, \delta^i)_i}{\partial \delta^i_m},
\end{aligned} \tag{F18}$$

where we introduced the quantity $B_i = \left(R_i\, \bar{N}\right) / \left(R_i\, \bar{N}\, (1 + G(a, \delta^i)_i)\right)^2$ to simplify notation. We then arrive at the expression

$$\begin{aligned}
\frac{\partial J_m(\delta^i)}{\partial \delta^i_m} &= \sum_k \frac{-1}{k_k^2}\, \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} \sum_j \frac{\partial F_j}{\partial \delta^i_m}\, \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \\
&= \sum_i \sum_k \frac{-1}{k_k^2}\, \mathrm{e}^{-2\pi m\, k \frac{\sqrt{-1}}{N}} \sum_j \mathrm{e}^{2\pi j\, k \frac{\sqrt{-1}}{N}} \\
&\quad \times \sum_p \frac{1}{\bar{N}} \boldsymbol{W}'(\boldsymbol{x}_p - \boldsymbol{x}_i) \\
&\quad \times \boldsymbol{V}'(\boldsymbol{q}_p - \boldsymbol{x}_j)\, B_i\, \frac{\partial G(a, \delta^i)_i}{\partial \delta^i_m}.
\end{aligned} \tag{F19}$$

This paper has been typeset from a TEX/LATEX file prepared by the author.