HAL
open science

Bayesian 3D velocity field reconstruction with virbius

Guilhem Lavaux

To cite this version:

Guilhem Lavaux. Bayesian 3D velocity field reconstruction with virbius. Monthly Notices of the Royal Astronomical Society, 2016, 457 (1), pp.172-197. 10.1093/mnras/stv2915 . hal-01304278

HAL Id: hal-01304278

https://hal.sorbonne-universite.fr/hal-01304278v1

Submitted on 29 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **457,** 172–197 (2016) doi:10.1093/mnras/stv2915

Bayesian 3D velocity field reconstruction with VIRBIUS

Guilhem Lavaux★

Sorbonne Universités, UPMC Univ Paris 6 et CNRS, UMR 7095, Institut dAstrophysique de Paris, 98 bis bd Arago, F-75014 Paris, France

Accepted 2015 December 10. Received 2015 November 30; in original form 2015 February 6

ABSTRACT

I describe a new Bayesian-based algorithm to infer the full three dimensional velocity field from observed distances and spectroscopic galaxy catalogues. In addition to the velocity field itself, the algorithm reconstructs true distances, some cosmological parameters and specific non-linearities in the velocity field. The algorithm takes care of selection effects, miscalibration issues and can be easily extended to handle direct fitting of e.g. the inverse Tully–Fisher relation. I first describe the algorithm in details alongside its performances. This algorithm is implemented in the VIRBIUS (VelocIty Reconstruction using Bayesian Inference Software) software package. I then test it on different mock distance catalogues with a varying complexity of observational issues. The model proved to give robust measurement of velocities for mock catalogues of 3000 galaxies. I expect the core of the algorithm to scale to tens of thousands galaxies. It holds the promises of giving a better handle on future large and deep distance surveys for which individual errors on distance would impede velocity field inference.

Key words: methods: data analysis – methods: statistical – galaxies: statistics – large-scale structure of Universe.

1 INTRODUCTION

Peculiar velocities are deviations of the apparent motion of tracers e.g. galaxies, from the Hubble flow. They are an essential tool to study dynamics of the Local Universe and in particular to probe the underlying gravity field, which is currently assumed to be generated by a Dark Matter density field. At low redshift i.e. $z \lesssim 0.1-0.2$, they are the only practical way of reconstructing the unbiased, true matter density field. The first mention of galaxy peculiar velocities go back to Hubble (Hubble & Humason 1931). When large scale structures data have been first acquired, peculiar velocities have quickly attracted a large attention (Aaronson et al. 1982, 1986; Lynden-Bell et al. 1988a), before fading out due to a lack of large corpus of distance data and robust methods of analysis.

New distance surveys, from which peculiar velocities can be inferred, have emerged in the recent years like SFI++ (Masters et al. 2006; Springob et al. 2007, 2009), 6dFv (Campbell et al. 2014), CosmicFlows-1 (Courtois et al. 2011), CosmicFlows-2 (Tully et al. 2013). More surveys are coming online such as TAIPAN/WALLABY (Beutler et al. 2011; Duffy et al. 2012). These surveys revived peculiar velocities as first class probes of cosmology by providing hundreds of thousands of distances. However, peculiar velocity analysis is notoriously error prone, being sensitive to different bias and systematic effects e.g. homogeneous (Lynden-Bell et al. 1988a) and inhomogeneous (Dekel, Bertschinger & Faber 1990) Malmquist bias, distance indicator calibration uncertainties (Willick 1994) or edge effects. Several attempts have been made at reconstructing the density field directly from distance data. For example, one can note the POTENT method (Bertschinger & Dekel 1989; Dekel et al. 1990, 1999), the Wiener filter approach (Zaroubi et al. 1995) or the Unbiased Minimum Variance algorithm (Zaroubi 2002). Additionally, the procedure to derive the power spectrum of the velocity field is relatively complex and prone to the same aforementioned systematics; though, there have been some early attempts at measuring it (Jaffe & Kaiser 1995; Kolatt & Dekel 1997; Zaroubi et al. 1997; Macaulay et al. 2011, 2012). New methods have also been recently designed to measure more accurately the first moments of cosmic flows from different aspects of distances and luminosities (e.g. Nusser & Davis 2011; Nusser, Branchini & Davis 2011, 2012; Feix, Nusser & Branchini 2014).

A common framework capable of handling all these items at the same time and building a consistent three dimensional (3D) peculiar velocity field is still missing. I am proposing to build such a framework from a full Bayesian joint analysis of the density field (bandwidth restricted Fourier modes of the density field), the cosmological parameters (e.g. Ω_m, the Hubble constant, the amplitude of scalar fluctuations),

★ E-mail: lavaux@iap.fr

© 2016 The Author
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

the observational parameters (e.g. selection function, Tully–Fisher relation) and the limitations of the model (e.g. amount of small scale non-linearities). By incorporating all these issues in a single framework, this model holds the promise of reducing (maybe cancelling) all systematic effects on the estimation of the 3D peculiar velocity field. The software has been named VIRBIUS (VelocIty Reconstruction using Bayesian Inference Software) and will be publicly available later on the author webpage.[1] Parts of the model will resemble VELMOD(Willick 1994; Willick et al. 1997). For example, Willick (1994) modelled the relation between the true distance and the observables for distance indicators the look like Tully–Fisher relations. Also, Willick et al. (1997) modelled the relation between redshift observations and Tully–Fisher observables (i.e. magnitudes and H I linewidth). These elements are parts of VIRBIUS, but they are generalized and included in a wider framework. I note that Johnson et al. (2014) have also pushed the effort of measuring accurately velocity field. In all the aforementioned work, however, a common framework to handle all components self-consistently are not included. Also, the possibility of unseen measurement failure is not accounted for. This will be another major addition (and complexity) to the model. Of course augmenting the model with limited data available comes at a cost: e.g. the power spectrum must be parametrized in terms of a small number of cosmological parameters. Among them I will select a few of particular interests: the overall amplitude of the powerspectrum, which is degenerate with the growth factor and the Hubble constant, which governs the shape. All the other cosmological parameters are kept fixed in this work. I will introduce other parameters that describe the data set itself (e.g. zero-point calibration, noise levels).

The structure of the paper is as follows. In Section 2, I describe both the adopted model and the algorithm that I have developed to explore the parameter space given some distance galaxy catalogue. The model, in Section 2.1, includes description of cosmological expansion, distance uncertainties, and a clean separation between the linear and the non-linear component of the velocity field. The model is fully Bayesian, and priors can be adjusted easily to include more detailed description of selections effects. In Section 2.2, I describe in detail the algorithm that is required to efficiently sample the posterior distribution of all the parameters that enter into the model, including the velocity field itself. In Section 3, I present the results of the test of this algorithm on a variety of mock catalogues: an ideal, though slightly unrealistic, and a mock catalogue generated assuming perfect homogeneity of tracers and Gaussian random fields statistics for velocity fields (Section 3.1), a more realistic mock catalogue based on haloes of an *N*-body simulation either with a trivial or a more complex selection function (Section 3.2). In Section 4, I conclude on the performance of the algorithms and the prospects for its use for existing and future distance surveys.

2 STATISTICAL METHOD

In this Section, I explain the model that I am using to describe self-consistently the velocity field, the cosmology, the redshifts and the distances of the tracers of the velocity field. In Section 2.1, I detail the model and the approximations that I have made. In Section 2.2, I describe the algorithm used to sample the posterior distribution in the huge parameter space.

2.1 Model

2.1.1 The flow model

I propose to solve the general problem of reconstructing in an unbiased way the 3D peculiar velocity field and cosmological parameters from a set of redshifts and distance modulus of tracers. I put $N_{\rm d}$ the number of tracers. I propose a self-consistent approach based on a probabilistic modelling. For the low-redshift Universe, and a given tracer i, it is possible to write a linear relationship between the redshift z_i, the distance d_i, the pseudo-Hubble constant $\widetilde{H}$ at redshift zero and the peculiar velocity $\boldsymbol{v}(\boldsymbol{r})$ as

$$z_i = \widetilde{H} d_i + v_i^r + \epsilon_{z,i}, \tag{1}$$

with $v_i^r = \boldsymbol{v}(d_i \hat{u}_i)\hat{\boldsymbol{u}}_i$ the line of sight component of the peculiar velocity of the ith object, $\hat{\boldsymbol{u}}_i$ the unit vector pointing in the direction of the tracer, $\epsilon_{z,\,i}$ the redshift measurement error. This is the usual Hubble relation, though we have replaced H by $\widetilde{H}$ to take into into account the fact that the calibration of distance indicator may not be absolute. Additionally, we do not have access to a precise probe of the distance. The equation (1) is only valid at extremely low redshift. The aim of this work is to have a self-consistent and accurate reconstruction of velocity field for large and deeper distance survey. Of course, different cosmological distances appear, like the luminosity and the comoving distances. From now on, we will use d_i as the comoving distance of an object i and $d_i^{\rm L}$ its corresponding luminosity distance. The exact relation combining cosmological and peculiar velocity, while they are non-relativistic, induced Doppler effect is the following:

$$1 + z_i = \left(1 + \bar{z}_i\left(d_i^{\rm L}\right)\right)\left(1 + \frac{v_i^r}{c}\right), \tag{2}$$

with $\bar{z}_i$ the cosmological redshift, which depends on the luminosity distance, v_i^r the line-of-sight component of the velocity field, assumed to be small compared to the speed of light c and in the rest frame of large scale structures. The relation, for a flat universe, is explicitly the following (e.g. Weinberg 1972)

$$\bar{d}^L(\bar{z}) = \frac{c(1+\bar{z})}{\widetilde{H}} \int_{z=0}^{\bar{z}} {\rm d}z\, \frac{1}{E(z)}, \tag{3}$$

[1] http://www.iap.fr/users/lavaux/

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

and

$$E(z) = \left(\Omega_{\rm M}(1+z)^3 + \Omega_\Lambda\right)^{1/2}. \tag{4}$$

The comoving distance $\bar{d}(d^L, \bar{z})$ is related to the luminosity distance d^L according to

$$\bar{d}(d^L, \bar{z}) = \frac{\bar{d}_L}{1+\bar{z}}. \tag{5}$$

The equation (3) is numerically invertible which, assuming d^L is known, allows for the derivation of the cosmological redshift. In the text, we will introduce the observational counterpart of the luminosity distance, called distance modulus μ, whose definition is

$$\mu = 5\log_{10}\left(\frac{\bar{d}^L}{10\ {\rm pc}}\right). \tag{6}$$

Finally, the relation (2) can be rewritten as followed:

$$v_i^r = c\frac{z_i - \bar{z}_i\left(\bar{d}_i^{\rm L}\right)}{1+\bar{z}_i\left(d_i^{\rm L}\right)}. \tag{7}$$

Davis & Scrimgeour (2014) recently reminded the community that using the linear approximation instead of equation (7) leads to substantial error even at relatively low redshift ($z \lesssim 0.05$). So, it is fundamental to include the complete treatment in my analysis so that the reconstructed velocity field are unbiased for future peculiar velocity surveys. I am assuming that the measured redshift is without error in the above equation. Of course, that is not the case, and the redshift error will be treated in the next section.

Though this relation between the v_i^r and the observed redshift is more complex than equation (1), it does not introduce any new systematic errors. The only problem that is introduced is the proper tracking of the cosmological redshift $\bar{z}_i$ and the comoving distance d_i when the luminosity distance d_i^L changes. In all this work, all algorithms make use of the equation (7) instead of the linear relation (1). In the above, I am considering that the observation of luminosity distance is perfect. That is not the case in practice as a number of effects are changing the apparent luminosity such as gravitational lensing, Integrated Sachs–Wolfe effect and gravitational redshift and peculiar velocities (Sasaki 1987; Pyne & Birkinshaw 2004; Bonvin, Durrer & Gasparini 2006). For the moment, we will neglect all these effects, keeping in mind that in data they will eventually have to be inserted into the likelihood analysis of luminosity distances. The last of these effects could be important to ensure consistent treatment of peculiar velocities as highlighted by Sasaki (1987), Hui & Greene (2006). To summarize, the distance modulus itself is affected by peculiar velocities at first order because the observed flux is itself sensitive to beaming and Doppler effects. The observed luminosity distance is in fact (Hui & Greene 2006):

$$d_{\rm o}^L = \bar{d}_{\rm LSS}^L\left(1 + \frac{1}{c}(2\boldsymbol{v}_e - \boldsymbol{v}_o).\hat{\boldsymbol{n}}\right) \tag{8}$$

with $d_{\rm o}^L$ the luminosity distance determined in the observer rest frame i.e. from observed flux, $\bar{d}_{\rm LSS}^L$ the actual luminosity distance of the object in an homogeneous universe with a FLRW metric, $\boldsymbol{v}_e$ ($\boldsymbol{v}_o$, respectively) the peculiar velocity of the emitter (observer, respectively) with respect to this homogeneous background and c the speed of light in vacuum. This relation is exact at first order. In this work, we will neglect the impact of this term, while focusing on the peculiar velocity present in the Doppler effect of the observed spectrum (equation 2). I note that the introduction of this correction would not change the algorithm fundamentally but introduce additional complexity in the formulation of the likelihood of the distance modulus (as detailed in equation 16). I also note that most peculiar velocity analysis (except supernovae) neglect the full impact of this term (Johnson et al. 2014).

2.1.2 *The million parameter likelihood analysis*

Historically, direct extrapolation of the velocity field from this relation has lead to a number of biases, like the inhomogeneous and homogeneous Malmquist biases (Dekel et al. 1990). They originate from the reuse of an imprecise distance indicator for both estimating the line of sight peculiar velocity $v_i^r = \boldsymbol{v}.\hat{\boldsymbol{u}}_i$ and using it as an estimate of the true distance $q^h d_i$, with $q^h = \widetilde{H}/H$. We are however not doomed to be limited by this problem. I propose to consider the distance itself as a random variable to generate the velocity field. This is not an entirely new proposal. In Willick et al. (1997), the VELMOD technique was already trying to improve the distance using a likelihood approach. They used a peculiar velocity field predicted using linear theory of gravitational instabilities and galaxy redshift surveys as a prior for the velocity field. The idea of generating random velocity fields in agreement with observation is not new either, it originates back to the constrained realization of Gaussian random fields in cosmological context (Hoffman & Ribak 1991, 1992). There was no published work that has attempted to blend both constrained realizations, distance sampling and parameter estimation. We are not bound to be limited to proceed sequentially for the analysis of peculiar velocity field. Notably, it is in principle possible to adjust both the power spectrum and the field itself, as it is done for the data of the Cosmic Microwave Background (Wandelt, Larson & Lakshminarayanan 2004).

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

Ideally, we would like to have a representation of the joint probability of the velocity field, the distances, the Hubble constant and possibly cosmological parameters p and an additional noise parameter $\sigma_{\rm NL}$, which will characterize the departure from linear theory of the actual velocity field. So, in practice we will fit the following model:

$$1 + z_i = \left(1 + \bar{z}_i\left(d_i^L\right)\right)\left(1 + \frac{\boldsymbol{v}_{\rm linear}(q^h d_i \hat{u}_i).\hat{\boldsymbol{u}}_i + \epsilon_{{\rm NL},i}}{c}\right) + \epsilon_{z,i}, \tag{9}$$

$$= \left(1 + \bar{z}_i\left(d_i^L\right)\right)\left(1 + \frac{Hf\boldsymbol{\Psi}_{\rm linear}\left(q^h d_i \hat{u}_i\right).\hat{\boldsymbol{u}}_i + \epsilon_{{\rm NL},i}}{c}\right) + \epsilon_{z,i}, \tag{10}$$

with

$$\langle \epsilon_{{\rm NL},i}\epsilon_{{\rm NL},j}\rangle = \sigma^2_{{\rm NL,type}(i)}\delta_{i,j}. \tag{11}$$

d_i the comoving distance from the observer of the tracer i (actually a function of d_i^L in this work) and H the Hubble constant at redshift $z = 0$. Each tracer is given an assignment type(i). I note that d_i is the comoving coordinates scaled with $\tilde{H}$. So, it is not the true distance but the distance in the convention of the calibration of the distance indicator. If the distance ladder is correctly built then $q^h = 1$, however, this is not in general guaranteed. In equation (10), we have introduced the displacement field Ψ, which is expressed in comoving coordinates. If we assume that no vorticity is created on large scales, it can be simply described through its divergence, which I will call Θ to follow earlier conventions on velocity fields. In the Lagrangian linear regime, the displacement field is related to the velocity field by a linear relation, which gives the second part of equation (10). Thus, we have the following relations:

$$\nabla_{\boldsymbol{r}}.\Psi = \Theta(\boldsymbol{r}), \tag{12}$$

$$\boldsymbol{v}_{\rm linear}(\boldsymbol{r}) = fH\Psi_{\rm linear}(\boldsymbol{r}). \tag{13}$$

Physically, Θ is related to the density fluctuations at present time in the Universe owing to continuity equation (e.g Peebles 1980). Additionally, I will call $\hat{\Theta} = \{\hat{\Theta}(\boldsymbol{k}_q)\}$ the discrete Fourier basis which represent the full continuous field Θ. Thus, their formal relationship is

$$\Theta(\boldsymbol{r}) = \frac{1}{L^3}\sum_{\boldsymbol{q}}\hat{\Theta}(\boldsymbol{k}_q)\mathrm{e}^{\mathrm{i}\boldsymbol{k}_q.\boldsymbol{r}}, \tag{14}$$

with

$$\boldsymbol{k}_q = \frac{2\pi}{L}\boldsymbol{q} \tag{15}$$

and $\boldsymbol{q} \in \{0, 1, \ldots, N-1\}^3$, N the resolution of the reconstructed field. These amplitudes will serve as free parameters of the field in the rest of the text. The splitting of the velocity into the scaling and displacement components allows us to make a shortcut later on when adjusting the Hubble constant to the data. The most simple model (equation 11) has a single type, but it is possible to have several. This approach can be required if we try to model a set of tracers which could be split into subpopulation such as clustered and non-clustered (i.e. elliptical versus spiral galaxies). It presents the other advantage of isolating potential catastrophic errors in the distance or spectroscopic measurement of a tracer. For example, the assignment of a supernova to a galaxy is sometimes dubious as its observed spectrum can be heavily blue shifted by the explosion processes. In other cases, the adjusted luminosity distance can be a strong outlier in statistical empirical relation such as the Tully–Fisher relation. The introduction of several types of tracers (including outliers) is related to the problem of the Gaussian mixture (Pearson 1894; Dempster, Laird & Rubin 1977). I evaluate the velocity field at the real distance using the scaling factor $q^h = \widetilde{H}/H$. $\boldsymbol{v}_{\rm linear}(\boldsymbol{r})$ has the same statistical properties as the velocity field derived from linear perturbation theory at $z = 0$.

Finally, the data are given by duet for each galaxy i: the distance modulus μ_i and the observed redshift z_i. I will assume that the noise on the observed distance modulus and the redshift measurement are both Gaussian, with standard deviations $\sigma_{\mu,i}$ and $\sigma_{z,i}$, respectively. Because the two data are acquired independently the likelihood, i.e. the probability of observing the data $\{(\mu_i, z_i)\}$, given the model is immediately given by

$$\mathcal{L} = P\left(\{\mu_i, z_i\} \,|\, \left\{d_i^L\right\}, \{\sigma_{z,i}, \sigma_{\mu,i}\}, \{\hat{\Theta}(\boldsymbol{k}_q)\}, H, \widetilde{H}, \Sigma_{\rm NL}, \mathcal{T}, \left\{p_t^{\rm type}\right\}\right) \propto \prod_{i=1}^{N_{\rm d}}\left(\sigma^2_{z,i}(1+\bar{z}_i)^{-2} + \sigma^2_{{\rm NL,type}(i)}\right)^{-1/2}$$
$$\times\exp\left\{-\frac{1}{2}\sum_{i=1}^{N_{\rm d}}\frac{\left[v_i^r(z_i, d_i) - Hf\Psi_{r,i}(q^h)\right]^2}{\left[\left(\sigma^2_{z,i}(1+\bar{z}_i(d_i)\right)^{-2} + \sigma^2_{{\rm NL,type}(i)}\right]} - \frac{(\mu_i - 5\log_{10}(d_i^L/10{\rm pc}))^2}{\sigma^2_{\mu,i}}\right\}, \tag{16}$$

with $\Psi_{r,i}(q^h) = \Psi_{\rm linear}(q^h d_i \hat{\boldsymbol{u}}_i).\hat{\boldsymbol{u}}_i$, $\Sigma_{\rm NL} = \{\sigma_{{\rm NL},q}\}$, $\mathcal{T} = \{{\rm type}(q)\}$ and $N_{\rm d}$ the number of provided tracers (i.e. the size of set $\{\sigma_{\mu,i}\}$). Using Bayes identity, we may now express the posterior probability of the parameters, given the

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

data:

$$P(\mathcal{D}^L = \left\{d_i^L\right\}, \hat{\Theta} = \{\hat{\Theta}(\boldsymbol{k}_q)\}, \widetilde{H}, H, A_{\rm S}, \{\sigma_{{\rm NL},t}\}, \mathcal{T}| \\ \mathcal{M} = \{\mu_i\}, \mathcal{Z} = \{z_i\}, \Sigma_z = \{\sigma_{z,i}\}, \Sigma_\mu = \{\sigma_{\mu,i}\}) \\ = \frac{\mathcal{L} \times \pi(\mathcal{D}^L)\pi(\hat{\Theta})\pi(\Sigma_{\rm NL})\pi(H)\pi(\{{\rm type}(i)\})\pi(A_{\rm S})}{\sum_{\mathcal{T}'} \int {\rm d}H {\rm d}\hat{\Theta} {\rm d}\mathcal{D}^L {\rm d}\Sigma_{\rm NL}\, \mathcal{L} \times \pi(\mathcal{D}^L)\pi(\mathcal{T}')\pi(\hat{\Theta})\pi(\{\sigma_{{\rm NL},t}\})\pi(H)}, \tag{17}$$

where $\mathcal{T}'$ runs over all possible type combinations in the denominator, t is the index of one of the type. The functions π are priors on the specified parameters. I assume that the statistics of $\hat{\Theta}$ is determined by the power spectrum $P(k)$ of primordial density fluctuations scaled to redshift zero. This power spectrum may itself depends on some cosmological parameters. I have chosen to incorporate the Hubble constant H and the amplitude of the power spectrum $A_{\rm S}$ as free parameters but to keep the other parameters at their best-fitting value from other probes, as e.g. *WMAP*9 (Hinshaw et al. 2012) or *Planck* (Planck Collaboration XVI 2013). $A_{\rm S}$ is defined as the pre-factor in the unnormalized power spectrum:

$$P_{\Theta\Theta}(k) = A_{\rm S}\left(\frac{k}{1\,h{\rm Mpc}^{-1}}\right)^{n_{\rm s}} T^2(k), \tag{18}$$

with $T(k)$ the transfer function, normalized to one for $k \to 0$. These free parameters will allow us to run a self-consistent check of the cosmology. We thus have

$$\pi(\hat{\Theta}|H, A_{\rm S}) = \prod_{\boldsymbol{q}} (2\pi P_{\Theta\Theta}(k_{\boldsymbol{q}}; H, A_{\rm S}))^{-1/2} \exp\left(-\frac{|\hat{\Theta}(\boldsymbol{k}_{\boldsymbol{q}})|^2}{2P_{\Theta\Theta}(k_{\boldsymbol{q}}; H, A_{\rm S})}\right). \tag{19}$$

The natural basis of representation of $\hat{\Theta}$ is thus the Fourier basis. Because Ψ is assumed to be without vorticity, it can be solved with the same Green function as the gravity. We can introduce an auxiliary scalar field Φ such that $\Psi = \nabla_x \Phi$ and Φ must satisfy the Poisson equation $\Delta\Phi = \Theta$. Thus, in Fourier space we obtain

$$\Phi(\boldsymbol{r}) = -\sum_{\boldsymbol{q}} \frac{1}{k_{\boldsymbol{q}}^2} {\rm e}^{i\boldsymbol{k}_{\boldsymbol{q}}\cdot\boldsymbol{r}} \hat{\Theta}(\boldsymbol{k}_{\boldsymbol{q}}) \tag{20}$$

The full expression of the displacement Ψ in terms of $\hat{\Theta}$ is thus after taking the gradient of Φ:

$$\Psi(\boldsymbol{r}) = \sum_{\boldsymbol{q}} \frac{i\boldsymbol{k}_{\boldsymbol{q}}}{k_{\boldsymbol{q}}^2} {\rm e}^{i\boldsymbol{k}_{\boldsymbol{q}}\cdot\boldsymbol{r}} \hat{\Theta}(\boldsymbol{k}_{\boldsymbol{q}}) \tag{21}$$

2.1.3 Some notes on priors

The prior on the distance translates our preconception of the localization of the tracers in the volume into a mathematical expression. I chose to consider two possibilities. The first kind of prior that I consider consists in fitting an empirical distribution $f_{\rm g}(d)$ of galaxies assuming isotropy and uniformity in the choice of the tracers. The distribution is self-consistently estimated from the ensemble of reconstructed distances, and thus in unit of Mpc. The empirical distribution is given by

$$\pi_{\rm isotropy}(\mathcal{D}^L|p, d_{\rm cut}, n) \propto D^p \exp\left(-\left(\frac{D}{d_{\rm cut}}\right)^n\right), \tag{22}$$

which is trivially combined in the total distance prior

$$\pi_{\rm isotropy}(\mathcal{D}^L|p, d_{\rm cut}, n) \propto \prod_{i=1}^{N_{\rm d}} f_{\rm g}\left(d_i^L; p, d_{\rm cut}, n\right). \tag{23}$$

Note that the use of this prior expands the parameter space to include $\{p, d_{\rm cut}, n\}$, and we use the luminosity distance of the tracers.

For some mock catalogue, I will consider a second choice i.e. that tracers are homogeneously distributed in a given determined by luminosity distance. While this choice is questionable as the tracers are more expected to be homogeneously distributed in comoving volume, this choice simplifies greatly the tests. It also does not remove any value to the VIRBIUS model as in any practical case the first prior will be used, which automatically absorb differences between luminosity distances and comoving distances in the parametrization. Thus, the prior takes the form:

$$\pi_{\rm homogeneous}(\mathcal{D}^L) \propto \prod_{i=1}^{N_{\rm d}} \left(d_i^L\right)^2. \tag{24}$$

Finally, for the very Local Universe, the comoving and luminosity distances are equal; thus, this prior correspond to a classical problem. In particular, this prior is related to the 'homogeneous Malmquist bias' correction (Lynden-Bell et al. 1988b; Strauss & Willick 1995), which leaves the inhomogeneous part not modelled. Strauss & Willick (1995) indicates that the correction introduced by the inhomogeneous component is subdominant in their simulation. Of course, this statement depends on the statistical distribution of the tracers themselves, as elliptical galaxies will be located more in the centre of the density peaks. We will put the homogeneous approximation of the prior to the test in Section 3.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

Algorithm 1 Blocked sampling algorithm

```
procedure GENERATEMARKOVCHAINELEMENT(s)
    for j = 0 to N_S do
        s_{i,j} ↶ P(s_{-,j} | s̄_{i,j})
    end for
end procedure

procedure GENERATEMARKOVCHAIN
    s ← 0
    loop
        GENERATEMARKOVCHAINELEMENT(s)
        write state s in a file
    end loop
end procedure
```

Each tracer i is given a type type(i). The prior probability of this typing is given by a finite set of values:

$$\pi(\mathcal{T}|\mathcal{P}) = \prod_{i=1}^{N_\mathrm{d}} p^{\mathrm{type}}_{type(i)}, \tag{25}$$

where $\mathcal{P}$ is the ensemble of possible probabilities $\{p^{\mathrm{type}}_j\}$ with j going over the available types, and type(i) maps the ith galaxy to the jth type.

Finally, I assume a uniform prior on the Hubble constant H, the effective Hubble constant $\widetilde{H}$ and on the variances $\sigma^2_{\mathrm{NL},q}$. I acknowledge that the uniform prior on $\widetilde{H}$ is not equivalent to a uniform prior on the zero-point calibration of the distance indicator, which are often linearly derived from magnitudes. Assuming that a uniform prior on the magnitude of the zero-point would yield a prior $\pi(\widetilde{H}) \propto 1/\widetilde{H}$, which is stricter than a pure uniform prior on $\widetilde{H}$.

In the above model, I have not treated the problem of selection effects. There could be some concern that the selection function of catalogues is not modelled here. Indeed, according to Willick (1994), Strauss & Willick (1995), depending on the choice of the used distance indicator (e.g. forward Tully–Fisher versus inverse Tully–Fisher), some systematic bias could be introduced in the velocity/distance reconstruction. I will argue in the following section that they have nearly no effect on the algorithm except in the determination of distances.

2.2 Sampling algorithms

I use the blocked Gibbs sampling method (Geman & Geman 1984; Liu, Wong & Kong 1994; Wandelt et al. 2004) to solve for the problem of having an unbiased estimate of the velocity field, including proper error bars on all parameters of the adopted model. This sampling technique is related to Markov Chains such as the Metropolis–Hasting algorithm (Metropolis & Ulam 1949; Metropolis et al. 1953; Hastings 1970), but in this case we always accept the new proposed move. Blocked Gibbs sampling is converging efficiently in two cases: cosmic variance limited problems and high signal-to-noise (S/N) ratio regimes. Unfortunately, it has potentially long convergence when model parameters are correlated and/or the model has to face intermediate S/N ratio regimes. Gibbs sampling has the advantage of splitting a complicated posterior into pieces that are easier to compute. I note that we have a Markov chain whose state $\mathcal{M}_i$ is described by the vector[2]

$$\mathcal{M}_i = \left(H_i; \widetilde{H}_i; A_{S,i}; \Sigma_{\mathrm{NL},i}; \hat{\Theta}_i; \mathcal{D}^L_i; \mathcal{P}_i; \mathcal{T}_i; d_{\mathrm{cut},i}; p_i, n_i\right), \tag{26}$$

$$= \left(s_{i,j}\right) \tag{27}$$

with $\hat{\Theta}$ the Fourier modes of the velocity field, sampled on a finite grid of modes $\boldsymbol{k}_q$, $\mathcal{T}_i$ the typing of tracers (noted {type(q)} above) and $\mathcal{P}_i$ the probability of each type. I remind the reader that n, p and d_{cut} are the parameters of the model for the selection function given in equation (22). I note N_S the number of variables in $\mathcal{M}_i$. The second equation (equation 27) implicitly defines the ordering of the parameters in the state $\mathcal{M}_i$. I will use the following notation to indicate that I will condition on everything except the indicated variable $s_{i,j}$:

$$\bar{s}_{i,j} = \left(s_{i,0}, \ldots, s_{i,j-1}, s_{i-1,j+1}, \ldots, s_{i-1,N_\mathrm{S}}\right). \tag{28}$$

However, if I have analytically marginalized according to the velocity potential $\hat{\Theta}$ then I note

$$\widetilde{\bar{s}}_{i,j} = \bar{s}_{i,j} \backslash (\hat{\Theta}_i) \tag{29}$$

In general, the algorithm will produce a new chain state using the Algorithm 1, with $s_{-,j}$ indicating that we do not consider the specific value of the parameter s_j but the general posterior of this parameter.

[2] The last three parameters are only involved for the more detailed selection function.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

Algorithm 2 Partially collapsed Gibbs sampling algorithm in `VIRBIUS`.

procedure GENERATEPARTIALLYCOLLAPSEDMARKOVCHAINELEMENT(s)
  for $j = 0$ to N_S^0 **do**
    $s_{i,j} \curvearrowleft P_c(s^0_{-,j} | \widetilde{\bar{s}}^0_{i,j}, \widetilde{\bar{s}}^1_{i,j})$
  end for
  Generate a constrained realization $\hat{\Theta}_i$
  for $j = 0$ to N_S^1 **do**
    $s_{i,j} \curvearrowleft P(s^1_{-,j} | \bar{s}^0_{i,j}, \hat{\Theta}_i, \bar{s}^1_{i,j})$
  end for
end procedure

Table 1. Dictionary for the notation of the parameters used in this work and explicit parameters on to which the conditional probability explicitly depends.

Parameter	Parameter name $s_{-,j}$	Explicit dependency
Physical Hubble constant	H	$\mathcal{D}^L$, $\widetilde{H}$, $\Sigma_{\rm NL}$, $A_{\rm S}$
Distance zero-point calibration	$\widetilde{H}$	$\mathcal{D}^L$, H, $\Sigma_{\rm NL}$
Non-linear/spurious error model	$\Sigma_{\rm NL}$	$\hat{\Theta}$, $\mathcal{P}$, $A_{\rm S}$
Amplitude of scalar fluctuations	$A_{\rm S}$	$\mathcal{D}^L$, H, $\widetilde{H}$, $\Sigma_{\rm NL}$
Velocity field scalar mode	$\hat{\Theta}$	$\mathcal{D}^L$, $\Sigma_{\rm NL}$, H, $\widetilde{H}$, $A_{\rm S}$
Luminosity distances	$\mathcal{D}^L$	$\hat{\Theta}$, $\widetilde{H}$, H
Type probability	$\mathcal{P}$	$\mathcal{T}$
Type	$\mathcal{T}$	$\mathcal{D}^L$, $\mathcal{P}$, $\Sigma_{\rm NL}$
Distance prior effective distance	$d_{\rm cut}$	$\mathcal{D}^L$
Distance prior slope parameter	p	$\mathcal{D}^L$

The above algorithm corresponds to the canonical Gibbs-Sampling algorithm. However, some parameters are strongly correlated to $\hat{\Theta}$, as e.g. $\widetilde{H}$. To improve the convergence, I will use the partially collapsed Gibbs Sampler algorithm (van Dyk & Park 2008). In the context of this work, it is possible to analytically marginalize according to $\hat{\Theta}$ when $s_{-,j}$ is H, $\widetilde{H}$, $\sigma_{{\rm NL},q}$ and $A_{\rm S}$. This adds complexity to each of the conditional posterior but it is still numerically tractable because the concerned posteriors are monodimensional. Defining:

$$\mathcal{M}_i^0 = \left(H_i; \widetilde{H}_i; A_{S,i}; \Sigma_{{\rm NL},i}\right) = \left(s^0_{i,j}\right), \text{ and} \tag{30}$$

$$\mathcal{M}_i^1 = \left(\mathcal{D}_i^L; \mathcal{P}_i; \mathcal{T}_i; d_{{\rm cut},i}; p_i, n_i\right) = \left(s^1_{i,j}\right), \tag{31}$$

the new algorithm is given in Algorithm 2 in which we used N_s^q to specify the number of elements in $\mathcal{M}_i^q$. The probability $P_{\rm c}$ is obtained by analytically marginalizing the conditional posterior $P(s_{-,j}, \hat{\Theta} | \widetilde{\bar{s}}_{i,j})$ according to $\hat{\Theta}$. Each of the used conditional posterior can be deduced from the main posterior (17). I will now detail them one by one. I give in Table 1 the parameters on to which each conditional probability function depends explicitly. The last element of the sampling chain corresponds to the sampling of the parameters of the distance selection prior, which in the program is done slightly separately and generates the three parameters of the prior in a single call.

Before detailing each of the algorithm required to sample the parameters, I note that this approach allows us for alleviating potential systematic biases arising from selection. The formalism that I gave in the previous section can be transformed to allow for a more complete fitting procedure that includes the distance indicator itself. The incorporation of the effects of selection in the distance relation fitting procedure consists then in simply multiplying the likelihood by the selection function $S(m, \eta, d)$ in the notations of Strauss & Willick (1995), m the apparent magnitude, η the luminosity linewidth and d the distance. As the algorithm uses conditional posteriors to explore the parameter space, this selection function will disappear from all expressions except the one that concerns the distance. If the selection is separable between (m, η) and d, the use of the forward Tully–Fisher algorithm would even not be sensitive to any details of the selection in (m, η). Of course, these assumptions are relatively strict and it may be that the parameter space is more entangled, e.g. between m and d, which would not allow this simplification. This remark is related to the distance indicator based on Tully–Fisher, though other indicators could benefit albeit on a different set of variables for the selection function. The entire algorithm presented in this paper is nevertheless completely general.

I now review each of the conditional posterior one by one to derive their expression.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

2.2.1 The Hubble constant H

The conditional posterior of the Hubble constant may be derived from the main posterior expression (17). After marginalization according to $\hat{\Theta}$, it gives:[3]

$$P(H|\mathcal{D}^L, \mathcal{Z}, \Sigma_z, \widetilde{H}, A_{\rm S}, \Sigma_{\rm NL}, \mathcal{T}) \propto |2\pi[\mathbf{C}^w(H)]|^{-1}\exp\left(-\frac{1}{2}\sum_{i,j=1}^{N_{\rm d}} v_i^r v_j^r [\mathbf{C}^w(H)]^{-1}_{i,j}\right) \tag{32}$$

with v_i^r as given by equation (7), the residual velocity once the apparent Hubble flow is subtracted. We have also used the covariance matrix $C^w_{i,j}$, which, following the model of equation (10), is defined as

$$[\mathbf{C}^w]_{i,j} = \left\langle v_i^r v_j^r \right\rangle \tag{33}$$

$$= \left\langle \epsilon^2_{i,NL}\right\rangle + \left\langle \epsilon^2_{i,z}\right\rangle + \sum_{a,b=1}^{3} \hat{u}_a^{(i)}\hat{u}_b^{(j)} \left\langle v_{{\rm linear},a}(d_i q^h \hat{\boldsymbol{u}}^{(i)}) v_{{\rm linear},b}(d_j q^h \hat{\boldsymbol{u}}^{(j)})\right\rangle \tag{34}$$

$$= \left(\sigma^2_{\rm NL,type(i)} + \sigma^2_{z,i}(1+\bar{z}_i)^{-2}\right)\delta_{i,j} + C^{v,r}_{i,j}, \tag{35}$$

where $\hat{u}_{\rm a}^{(i)}$ refers to the ath component of the unit vector pointing in the direction of the ith galaxy,

$$C^{v,r}_{i,j} = \langle v_{r,i} v_{r,j}\rangle = \sum_{\mu,\nu=1}^{3} \hat{u}_\mu^{(i)}\hat{u}_\nu^{(j)} \left[C^v_{\mu,\nu}\right]_{i,j}, \tag{36}$$

$q^h = \widetilde{H}/H$, and

$$\left[\mathbf{C}^v_{\mu,\nu}\right]_{,i,j} = (fH)^2 \int \frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3}\,\frac{k_\mu k_\nu}{|\boldsymbol{k}|^4} {\rm e}^{iq^h \boldsymbol{k}.(d_i\hat{\boldsymbol{u}}^{(i)} - d_j\hat{\boldsymbol{u}}^{(j)})} P_{\Theta\Theta}(|\boldsymbol{k}|) \tag{37}$$

is the covariance matrix of the large scale part of the velocity field. The covariance matrix $\mathbf{C}^v_{\mu,\nu}$ is derived from the prior on $\hat{\Theta}$ given in equation (19). In practice, the matrix $\mathbf{C}^v$ is computed using the Fourier–Taylor algorithm of Appendix B. The conditional probability given in equation (32) is mixing the need of small residual and correlated fluctuations through cosmic flows. It may be highly non-Gaussian and not necessarily with a single maximum. It is however a one-dimensional posterior, which makes it possible to tabulate it. I have used the algorithm of Appendix A to generate a random sample from this density distribution.

2.2.2 The effective Hubble constant $\widetilde{H}$

I have separated the Hubble constant presently linked to autocorrelations of the velocity field from the one corresponding to the redshift–distance relation. In Section 2.2.1, we have obtained the (complicated) posterior of the H. The conditional posterior of $\widetilde{H}$, again marginalized according $\hat{\Theta}$ takes the same form as (32), except that $\widetilde{H}$ is left free and H is kept constant. As $\widetilde{H}$ is involved in more non-linear relations due to cosmological redshift effects, the conditional posterior is non-Gaussian in several aspects. The full expression of this probability density is

$$P(\widetilde{H}|\mathcal{D}^L, \mathcal{Z}, \Sigma_z, H, A_{\rm S}, \Sigma_{\rm NL}, \mathcal{T}) \propto |2\pi[\mathbf{C}^w(\widetilde{H})]|^{-1}\exp\left(-\frac{1}{2}\sum_{i,j=1}^{N_{\rm d}} v_i^r(\widetilde{H}) v_j^r(\widetilde{H}) [\mathbf{C}^w(\widetilde{H})]^{-1}_{i,j}\right), \tag{38}$$

with $\mathbf{C}^w$ defined as in the previous section.

2.2.3 The tracer types

The model includes some freedom on the type of non-linearity (modelled by the extra noise $\sigma_{{\rm NL},\,k}$ as determined in Section 2.2.5) that affects each tracer. The model that I have adopted is the Gaussian mixture where each type is given an unconditional probability and the adopted extra noise depends on the type. The typing mechanism is represented by the projection function type(k). In the Gibbs sampling framework we can assume that we know the value of $\{\sigma_{{\rm NL},\,a}\}$ and infer statistically the unconditional probability of the type type(k). The probability the object k has a type type(k) equal to q is thus proportional to the probability of the type multiplied by the probability that the error term is likely according to the numbers derived in Section 2.2.5. Mathematically, the likelihood (equivalently the probability) that the set of tracers $\{k\}$ has some residuals $\{\epsilon_k\}$ given the type projector $\mathcal{T}$ and the probabilities of typing $\mathcal{P}$ is

$$P(\{\epsilon_k\}|\mathcal{T}, \mathcal{P}) \propto \prod_{k=1}^{N_{\rm t}} p^{\rm type}_{{\rm type}(k)} \frac{1}{\sqrt{\sigma^2_{\rm NL,type(k)} + \sigma^2_{z,k}(1+\bar{z}_k)^{-2}}} \exp\left(-\frac{\epsilon_k^2}{2\left(\sigma^2_{\rm NL,type(k)} + \sigma^2_{z,k}(1+\bar{z}_k)^{-2}\right)}\right) \tag{39}$$

[3] I am not explicitly giving the dependency of all the terms in this expression, which would be too notation heavy. I am keeping all dependencies implicit. For example, the covariance matrix $[\mathbf{C}^w]$ actually depends on H, $\widetilde{H}$, $\sigma_{\rm NL}$, $A_{\rm S}$ and all distances.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

with $\epsilon_a = v_a^r(z_a, d_a^L) - Hf\Psi_a^r(q^h)$. Using Bayes identity, we can derive the conditional probability of the type mapper $\mathcal{T}$ given the residuals $\{\epsilon_k\}$ and the type probability $\mathcal{P}$:

$$P(\mathcal{T} = \mathcal{T}^r|\{\epsilon_k\}, \mathcal{P}) = \frac{P(\{\epsilon_k\}|\mathcal{T}^r, \mathcal{P})\pi(\mathcal{T}^r)}{\sum_{\mathcal{T}'} P(\{\epsilon_k\}|\mathcal{T}', \mathcal{P})\pi(\mathcal{T}')} = \prod_{k=1}^{N_{\rm t}} P\left(\text{type}(k) = q|\Sigma_{\rm NL}, p_q^{\rm type}, \epsilon_k\right). \tag{40}$$

with, for a uniform prior on the type:

$$P\left(\text{type}(k) = q|\Sigma_{\rm NL}, \left\{p_{q'}^{\rm type}\right\}_{\{q'\}}, \epsilon_k\right) = \frac{p_q^{\rm type}\left(\sigma_{{\rm NL},q}^2 + \sigma_{z,k}^2(1+\bar{z})^{-2}\right)^{-1/2}\exp\left(-\frac{\epsilon_k^2}{2(\sigma_{{\rm NL},q}^2+\sigma_{z,k}^2(1+\bar{z}_k)^{-2})}\right)}{\sum_{q'} p_{q'}^{\rm type}\left(\sigma_{{\rm NL},q'}^2 + \sigma_{z,k}^2(1+\bar{z})^{-2}\right)^{-1/2}\exp\left(-\frac{\epsilon_k^2}{2(\sigma_{{\rm NL},q'}^2+\sigma_{z,k}^2(1+\bar{z}_k)^{-2})}\right)}. \tag{41}$$

To sample the adequate type for the tracer k, a brute force approach is largely sufficient: for each tracer we compute the $N_{\rm t}$ probabilities given by equation (41), generate a random number $r \in [0, 1[$ and choose the type b that satisfies

$$b = \max\left\{c \left| \sum_{q=1}^{c} P\left(\text{type}(k) = q|\Sigma_{\rm NL}, \left\{p_{q'}^{\rm type}\right\}_{\{q'\}}, \epsilon_k\right) \le r\right.\right\}. \tag{42}$$

This type b is then assigned to the tracer a for the state i of the Markov Chain.

2.2.4 *The tracer probability*

For this step, we assume that the type of each tracer is known and we want to infer the probability of unconditionally typing a particle to the type q. This can be readily derived from equation (39) using Bayes identity:

$$P(\mathcal{P}|\mathcal{T}) = \frac{\prod_{q=1}^{N^{\rm t}} p_q^{|T_q|}}{\int_{\mathcal{P}'}\prod_{q=1}^{N^{\rm t}} p_q'^{|T_q|}{\rm d}\mathcal{P}'}, \tag{43}$$

with $\mathcal{P}'$ the set of all possible probabilities such that

$$\sum_{q=1}^{N^{\rm t}} p_q = 1, \tag{44}$$

$$0 < p_q < 1 \text{ for all } 1 \le q \le N^{\rm t}. \tag{45}$$

with $T_q = \{k|\text{type}(k) = q\}$, and $|T_q|$ the number of elements in T_q. This probability density is a Dirichlet distribution. I am using the function `gsl_ran_dirichlet` of the GNU Scientific Library (GSL) to generate samples of such a distribution, conditioned on the number of tracers $|T_q|$ in each type q.

2.2.5 *Model error* $\sigma_{{\rm NL},\,k}$

We consider here the amount of extra noise $\sigma_{\rm NL}$ that is not captured by the part of the model that uses linear perturbation theory to derive the velocity field. The conditional posterior distribution, as derived from the main likelihood (17), is

$$P\left(\sigma_{{\rm NL},k}^2|\hat{\Theta}, \widetilde{H}, H, \mathcal{D}^L, \mathcal{Z}, \Sigma_z\right) \propto \prod_{i/\text{type}(i)=k}\left(\sigma_{z,i}^2(1+\bar{z}_i)^{-2} + \sigma_{{\rm NL},k}^2\right)^{-1/2}\exp\left(-\frac{\epsilon_i^2}{2\left(\sigma_{z,i}^2(1+\bar{z}_i)^{-2} + \sigma_{{\rm NL},k}^2\right)}\right). \tag{46}$$

The posterior is written in terms of $\sigma_{{\rm NL},k}^2$, as I have indicated that I am taking a uniform prior on $\sigma_{{\rm NL},k}^2$ and not $\sigma_{{\rm NL},\,k}$ (Section 2.1). This is again a one-dimensional distribution and I use the algorithm of Appendix A.

2.2.6 *Power normalization* $A_{\rm S}$

The normalization of the power spectrum of Θ is left free in the model. This allows to account for the possibility of a different growth rate of perturbations or a different amplitude of scalar perturbations of our Local Universe. Again, we are faced with a mono-dimensional conditional posterior distribution. The problem shares a lot of similarities with Section 2.2.1. We change parameter and write $A_{\rm S} = \alpha A_{S,\,i-1}$. The conditional posterior, marginalized according to $\hat{\Theta}$ becomes

$$P(\alpha|\mathcal{D}^L, \mathcal{Z}, \Sigma_z, \widetilde{H}, H, \sigma_{\rm NL}) \propto |\alpha\mathbf{C}^{v,r} + \mathbf{N}_{\rm NL}|^{-1/2}\exp\left(-\frac{1}{2}\sum_{i,j=1}^{N_{\rm d}} v_i^r(z_i, \bar{z}_i)[\alpha\mathbf{C}^{v,r} + \mathbf{N}_{\rm NL}]_{i,j}^{-1} v_j^r(z_j, \bar{z}_j)\right). \tag{47}$$

with $[\mathbf{N}_{\rm NL}]_{i,j} = (\sigma_{{\rm NL},\text{type}(i)}^2 + \sigma_{z,i}^2(1+\bar{z}_i)^{-2})\delta_{i,j}$. In the above equation, we have reused the covariance matrix $\mathbf{C}^{v,r}$ of the line-of-sight component of the velocity fields sampled at the tracer positions. This matrix is given in equation (36).

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

2.2.7 *The scalar potential of the velocity field*

Now, we consider the conditional probability of the Fourier modes of $\hat{\Theta}$, the opposite of the divergence of the velocity field. After simplification, the conditional posterior takes the following form:

$$P(\hat{\Theta}|\mathcal{Z}, \mathcal{D}^L, H, \Sigma_{\rm NL}) \propto \exp\left(-\sum_{i=1}^{N_{\rm d}} \frac{\left(v_i^r(z_i, \bar{z}_i) - Hf\Psi_{r,i}\right)^2}{2\left(\sigma_{z,i}^2(1+\bar{z}_i)^{-2} + \sigma_{\rm NL,type(i)}^2\right)}\right) \exp\left(-\sum_{q=1}^{N_q} \frac{|\hat{\Theta}(\boldsymbol{k}_q)|^2}{2P_{\Theta\Theta}(|\boldsymbol{k}_q|)}\right). \tag{48}$$

Sampling from this probability consists in generating a Gaussian random field satisfying $P(k)$, but with some integrals constrained by redshift observations. As the errors on constraints are Gaussian, we may use the algorithm proposed by Hoffman & Ribak (1991, 1992). For this, I remind the reader of the algorithm. The constrained random field $\hat{\Theta}^{\rm CR}$ is built from a random part $\hat{\Theta}^{\rm RR}$ and the correlated part:

$$\hat{\Theta}^{\rm CR}(\boldsymbol{k}) = \hat{\Theta}^{\rm RR}(\boldsymbol{k}) + \langle\hat{\Theta}(\boldsymbol{k})c_i\rangle C_{ij}^{w,-1}\left(c_i - \widetilde{c}_i^{\rm RR}\right), \tag{49}$$

with c_i the ith constraint to apply, $\widetilde{c}_i^{\rm RR}$ the mock observation of the same constraint in the pure random realization $\hat{\Theta}^{\rm RR}(\boldsymbol{k})$, $C_{ij}^w = \langle c_i c_j\rangle$, as given in equation (35), is the covariance matrix of the constraints. By construction, in the infinite volume limit, we have

$$\langle\hat{\Theta}^{\rm RR}(\boldsymbol{k})\hat{\Theta}^{\rm RR}(\boldsymbol{q})\rangle = (2\pi)^3\delta_D(\boldsymbol{k}+\boldsymbol{q})P_{\Theta\Theta}(k), \tag{50}$$

with $\delta_{\rm D}$ the Dirac distribution. In the case of this work, the contraints c_j are line-of-sight component of the velocity field. For the jth tracer, the constraint c_j is

$$c_j = \sum_{\mu=1}^{3}\hat{\boldsymbol{r}}_{j,\mu}\int\frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3}\,\frac{ik_\mu}{k^2}{\rm e}^{id_j\boldsymbol{k}.\hat{\boldsymbol{r}}_j}F_{\rm NL}(\boldsymbol{k})\hat{\Theta}(\boldsymbol{k}) + \epsilon_{j,\rm NL} + \epsilon_{j,z}, \tag{51}$$

where $\hat{r}_{j,\mu}$ is the μthcomponent of $\hat{r}_j$ the sky direction of the jth tracer, $\epsilon_{j,\rm NL}$ ($\epsilon_{j,z}$, respectively) is corresponding to the non-linear component not captured by $\hat{\Theta}$ (the redshift measurement error, respectively). I added a filter $F_{\rm NL}(\boldsymbol{k})$ to remove the contribution of modes that are below the scale of non-linearity. Effectively it is a Heaviside function on the norm of $\boldsymbol{k}$:

$$F_{\rm NL}(\boldsymbol{k}) = \begin{cases} 1 \text{ if} |\boldsymbol{k}| < k_{\rm NL}, \\ 0 \text{ otherwise.}\end{cases} \tag{52}$$

The correlation between $\hat{\Theta}$ and c_j is

$$\langle\hat{\Theta}(\boldsymbol{q})c_j\rangle = \sum_{\mu=1}^{3}\hat{r}_{j,\mu}\int\frac{{\rm d}^3\boldsymbol{k}}{(2\pi)^3}\,\frac{ik_\mu}{k^2}{\rm e}^{id_j\boldsymbol{k}.\hat{\boldsymbol{r}}_j}\langle\hat{\Theta}(\boldsymbol{k})\hat{\Theta}(\boldsymbol{q})\rangle = -\frac{i\hat{\boldsymbol{r}}_j.\boldsymbol{q}}{q^2}{\rm e}^{-id_j\boldsymbol{q}.\hat{\boldsymbol{r}}_j}P_{\Theta\Theta}(|\boldsymbol{q}|). \tag{53}$$

The equation above is computed globally using all tracers with the algorithm in Appendix C. If we did not use this algorithm, we would have needed $\mathcal{O}(N_{\rm d}\times N_{\rm g})$, with $N_{\rm g}$ the number of grid elements $\boldsymbol{q}$. On the other hand, the Fourier–Taylor Wiener algorithm allows the same value to be computed in $\mathcal{O}(N_{\rm d}) + \mathcal{O}(N_{\rm g})$. As indicated in Section 2.2.1, the covariance matrix is obtained by applying the Fourier–Taylor algorithm of Appendix B from the Fast Fourier Transform (FFT) of the weighed power spectrum on a regular grid. I do not use the analytically exact expression (as given in e.g. Gorski 1988) because it leads to neglecting finite grid effects and periodic boundary effects, which are dominant for the reasonable physical grid sizes which I consider (with typical a side length of 500 Mpc). We compute the mock observations $\widetilde{c}_j$ on the unconstrained field $\hat{\Theta}^{\rm RR}$ in exactly the same way: we do the FFT on a grid and then do a Fourier–Taylor synthesis to obtain the velocity field values. Finally, we can add a random realization of the noise to the interpolated value to construct $\widetilde{c}_j$. All values are recombined using equation (49) to obtain the final constrained Gaussian random field.

2.2.8 *Galaxy distances*

The conditional posterior of the distances may be derived from the main posterior expression (17) as

$$P(\mathcal{D}^L|\mathcal{Z}, \Sigma_z, \mathcal{M}, \Sigma_\mu, H, \sigma_{\rm NL}) \propto \mathcal{L}\times\pi(\mathcal{D}^L) \propto \prod_{i=1}^{N_{\rm d}} p_i^d\left(d_i^L\right), \tag{54}$$

with

$$p_i^d(d^L) = \frac{1}{\sqrt{2\pi\left(\sigma_{\rm NL}^2 + \sigma_{z,i}^2(1+\bar{z}_i)^{-2}\right)}}\exp\left(-\frac{1}{2(\sigma_{\rm NL}^2+\sigma_{z,i}^2(1+\bar{z}_i)^{-2})}\left(v_i^r(z_i, \bar{z}_i(d^L)) - fH\Psi_r(d(d^L)\hat{\boldsymbol{u}}_i)\right)^2\right)$$
$$\times\exp\left(\frac{(\mu_i - 5\log_{10}(d^L/10{\rm pc}))^2}{2\sigma_{\mu,i}^2}\right). \tag{55}$$

I note that the conditional posterior distribution of the distances of each tracers is separable in $N_{\rm d}$ independent monodimensional conditional posterior. This comes from the assumption that the noise on the redshift measurement is uncorrelated from tracer to tracer, and that all

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

correlations between tracer is accounted for by the velocity field and, possibly, the bias function. Selection function and clustering bias would typically be retained in this expression. Both can be added multiplicatively to p_i^d. For example it can be $S(m_i, \eta_i, d)$ for the selection function and $(1 + b\delta(d\hat{r}_i))$ for the clustering bias as already stated for VELMOD class of models (Willick 1994; Strauss & Willick 1995). The problem of sampling from this posterior reduces here to a sampling problem from $N^{\rm d}$ monodimensional posteriors. To achieve that, I use the classical algorithm of computing the inverse of the cumulative distribution applied on a random realization of a uniform distribution bounded by [0, 1]. The displacement field is computed at the appropriate position using the Fourier–Taylor synthesis algorithm of Appendix B.

2.2.9 *The distance prior parameters*

As the selection function of the galaxies in distance catalogues is poorly known, we have introduced in Section 2.1 a flexible selection function that depends on three parameters $\{p, d_{\rm cut}, n\}$. The conditional likelihood of the distances is exactly equal to the prior (22) in this case. Using Bayes identity and a uniform prior on the aforementioned parameters, the sampling of this parameters is achieved by another block sampling step. The probability of one of the three parameters, e.g. p, given the others is, using Bayes identity,

$$P(p|d_{\rm cut}, n, \mathcal{D}^L) = \frac{P(p, \ldots)}{\int {\rm d}p\ P(p, \ldots)} = \frac{\pi(\mathcal{D}^L|p, d_{\rm cut}, n)}{\int {\rm d}p\ \pi(\mathcal{D}^L|p, d_{\rm cut}, n)}, \tag{56}$$

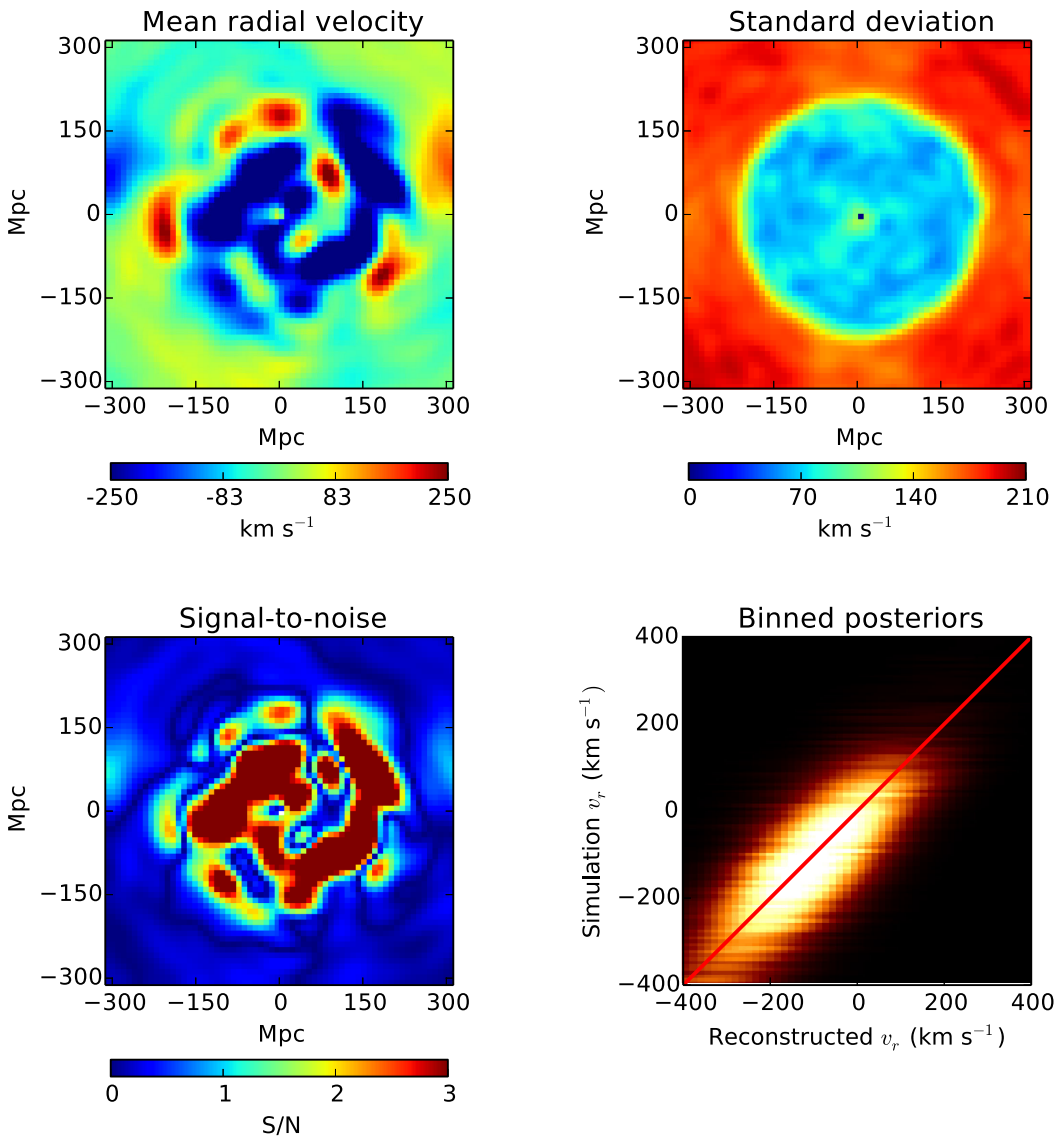

Figure 1. Results of the test on the mock catalogue based on Gaussian random field: central slice of the ensemble averages for the line of sight of the velocity field (top-left panel), the variance expressed as a standard deviation (top-right panel), the resulting S/N (bottom-left panel). In the bottom-right panel, I present a Bayesian comparison between the reconstructed velocities and the actual true velocity field of the simulation using the entire set of posterior distributions. The details are given in Section 3.1. Only the voxels whose centres are within 150 Mpc from the observer are considered in this panel. In all panels, the selection is isotropic, thus the absence of note on the axis.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

where $P(p\ldots)$ is the posterior probability of equation (17), with explicit dependency on the parameters of the prior $\pi(\mathcal{D}^L)$. All functions simplifies except the explicit dependence linking the distances to the distance selection, despite the possible existence of a selection function on galaxy properties (like the apparent/absolute magnitude or the H I linewidth for Tully–Fisher relation). Such a sampling step is achieved using the algorithm of Appendix A. To ensure a proper decorrelation, we loop over this block a number of times (typically 10).

3 TESTS ON MOCK CATALOGUES

The method that I have described in Section 2 is relatively complex and involves two major component: the model of the peculiar velocity field as traced by the galaxies or the clusters of the galaxies and the algorithm itself to adjust the data to the model. This involves two separate sets of tests. First, I will focus on test of the algorithm itself and the performance of the fit on data that were produced to correspond exactly to the model. This is the objective of the Section 3.1. Second, I will look into the capabilities and the limits of the model at reconstructing the velocity field and distances from noisy more realistic mock data sets, typically halo catalogues from *N*-body simulation, which is the objective of Section 3.2.

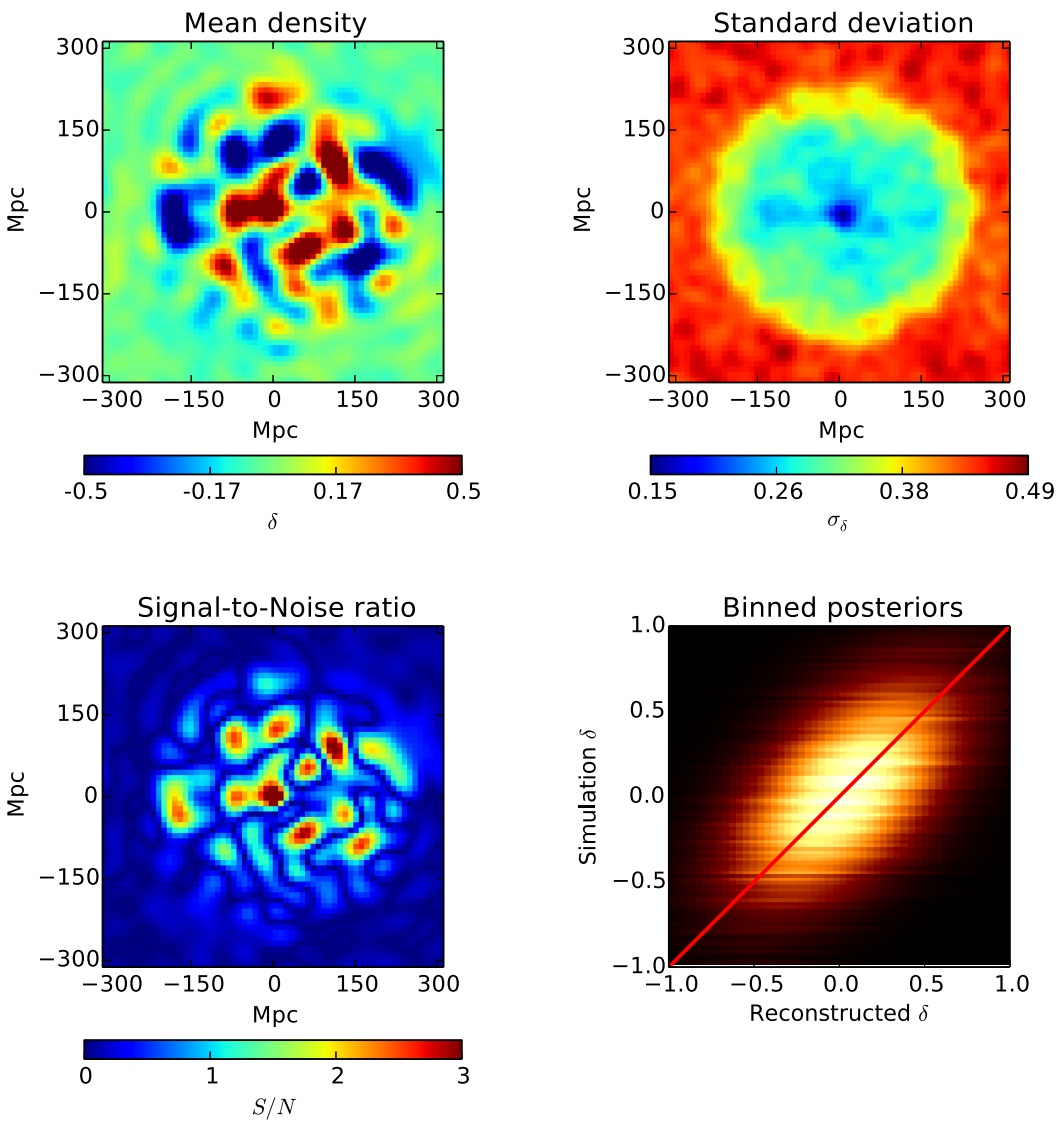

Figure 2. Results of the test on the mock catalogue based on Gaussian random field, density field distribution: central slice of the ensemble average (top-left panel), the variance expressed as a standard deviation (top-right panel), the resulting S/N (bottom-left panel). In the bottom-right panel, I present a Bayesian comparison between the reconstructed densities and the actual true density field of the simulation using the entire set of posterior distributions. In all panels, the selection is isotropic, thus the absence of note on the axis.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

3.1 Gaussian random field based mock catalogues

In this Section, I present the generation and the results of the test of the code against idealized mock tracer catalogues. These catalogues are generated in such a way that the statistics and properties of the tracers follow exactly the model presented in Section 2.1. However, they are slightly unrealistic by removing aspects not captured by the model, such as non-linearities or correlations between velocity field and tracer positions. These aspects may lead to biases in the results. It is none the less an interesting exercise to evaluate the performance of the algorithm.

I generate the mock catalogue of tracers, which should be galaxies, as follows.

(i) I generate a random realization of a 'density field', with power spectrum given by linear theory linearly extrapolated to $z = 0$ using the expression of Eisenstein & Hu (1998) for the power spectrum of density fluctuations, without the wiggles. The power spectrum is truncated at $k_{\max} = 0.1\ \mathrm{Mpc}^{-1}$.

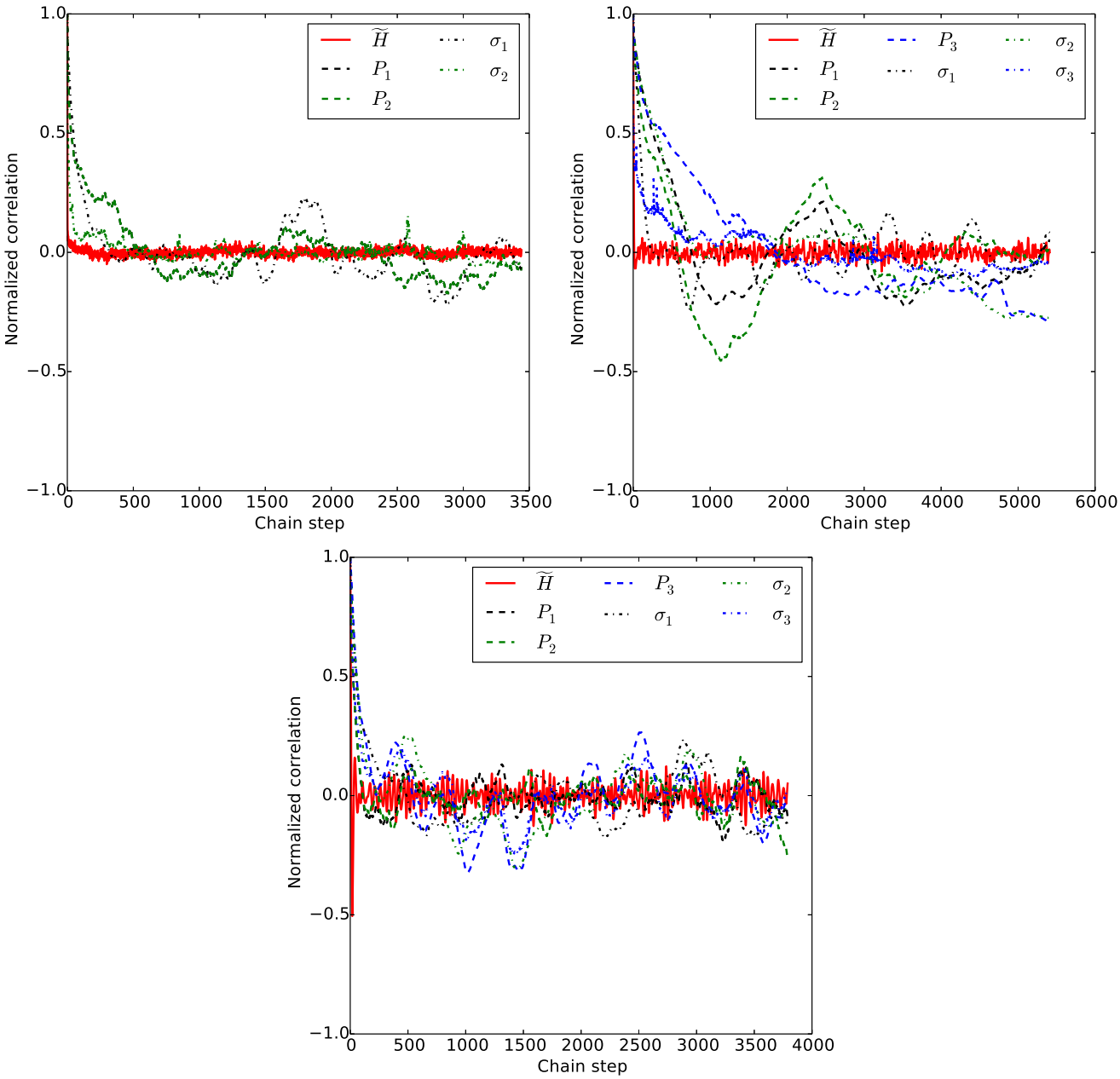

Figure 3. Autocorrelation of the value taken by metaparameters for the three chains considered in this work: $\widetilde{H}$ (thick solid red), the mixture probability P_i ($i \in \{1, 2, 3\}$), dashed lines, respectively, in black, green and blue, σ_i ($i \in \{1, 2, 3\}$), the error standard deviation for each tracer type, dash–dotted lines, respectively, in black, green and blue. The different panels correspond, respectively, to the homogeneous Gaussian random field case (topleft), the H3000 catalogue (topright), the Hcomplex catalogue (bottom middle). The horizontal axis tracks the step number along each of the chain.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

(ii) I pick 3000 randomly located tracers within a sphere of 200 Mpc, assuming that the tracers are homogeneous in luminosity distance coordinates. The luminosity distance is then converted in comoving distance and cosmological redshift.

(iii) I compute the velocity field at a resolution of $N = 512$, and evaluate its value at the position of tracers using a trilinear interpolation.

(iv) I compute the line-of-sight component of the peculiar velocity and the distance modulus. The mock observables are then generated taking a homogeneous noise of $\sigma_\mu = 0.2$, $\sigma_z = 20$ km s^{-1} and $\sigma_{\rm NL} = 200$ km s^{-1}.

I have chosen a fiducial ΛCDM cosmology with $\Omega_{\rm M} = 0.30$, $\Omega_{\rm b} = 0.04$, $\sigma_8(z = 0) = 0.84$, $H = 80$ km s^{-1} Mpc^{-1}, $n_{\rm S} = 1$. VIRBIUS is run assuming the same cosmology, leaving free both the amplitude of the power spectrum, assuming that there are either two or one specie(s) of tracers, the zero-point calibration $\widetilde{H}$ and the parameters of the selection function. The grid used to compute the Fourier–Taylor algorithms has $N = 64$ elements per dimension, which is sufficient to capture the details of $k = 0.1$ Mpc^{-1} and ensure a correct fast interpolation using the algorithm of Appendix B.

Before considering the whole analysis, I am showing in Figs 1 and 2 the result of reconstructing the velocity field from an exactly known cosmology, distance and small-scale non-linearities. I used a Gaussian noise with an amplitude of 200 km s^{-1} to model the small scale non-linearities as indicated above. In Fig. 1 (Fig. 2, respectively), I am showing the reconstruction of the line-of-sight component of the velocity field (density field, respectively). The figures were generated with 1055 Monte Carlo samples. In Fig. 1, the top-left panel (top right, respectively) shows the ensemble mean radial component (the standard deviation, respectively) of the velocity field. The bottom-left panel shows the S/N obtained by dividing the field of the top left panel by the one shown in the top right panel. Finally, the bottom-right panel is

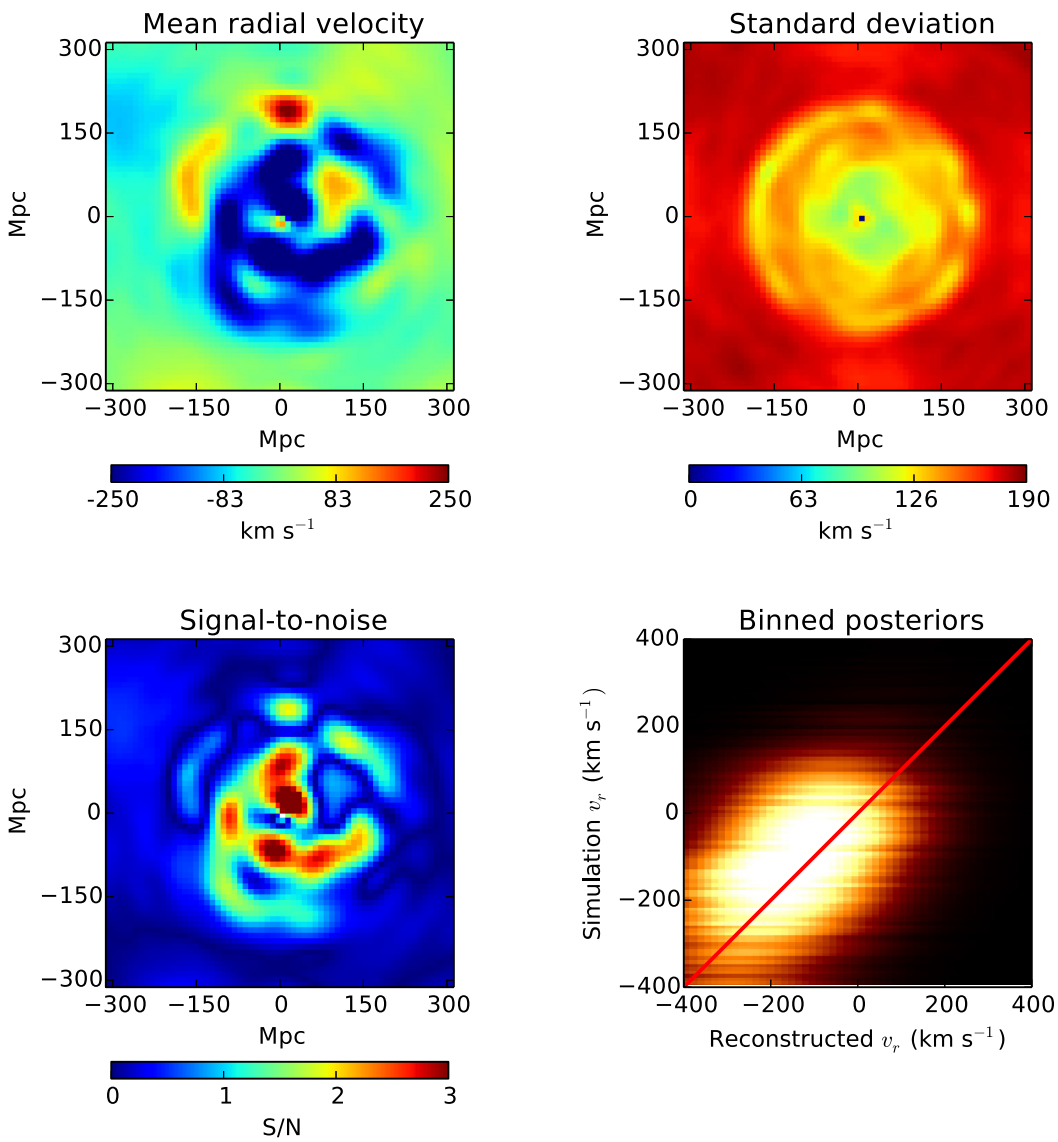

Figure 4. Results of the test on the mock catalogue based on Gaussian random field: central slice of the ensemble averages for the line of sight of the velocity field (top-left panel), the variance expressed as a standard deviation (top-right panel), the resulting S/N (bottom-left panel). In the bottom-right panel, I present a Bayesian comparison between the mean reconstructed velocities, and the actual true velocities of the mock tracers using the entire set of posterior distributions.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

obtained by showing all the individual posterior distribution for the velocities. These posterior distribution are binned on a grid, and the total intensity is the sum of the posterior intensities for each considered grid element. The amplitude of the distribution is coded in colour with respect to the x-axis, while the line-of-sight component of the velocity of the corresponding object as given by the simulation is shown on the y-axis. If the algorithm works correctly, all the distributions should overlap with the thick red diagonal, though not necessarily centred as the distributions are correlated. The correlation actually removes the effective number of independent samples compared to the number of elements that are plotted. The unbiased aspect of the method is exhibited by the plots of the normalized residuals in Fig. 8, where I have represented the histograms of the quantity

$$\epsilon_{\rm r} = \frac{\bar{v}_{\rm r}(\boldsymbol{x}) - v_{\rm r}^{\rm sim}(\boldsymbol{x}))}{\sigma(\boldsymbol{x})}. \tag{57}$$

The same histogram for a pure Gaussian distribution is represented with dashed green line. It can be immediately seen that the overall distribution of residuals is close to Gaussian with unit variance. Additionally, there is no obviously strong shift in the mean.

The panels are similar in Fig. 2 but this time for the density field. We see that in this optimistic configuration the velocity field is very well reconstructed, on the other hand the density field is already significantly noisy. These results are the benchmarks to which we will compare the performance of other reconstructions.

The results for more complete tests are given in Figs 3, 4, 5, 6 and 7. The figures were generated with 6894 (3037, respectively) Monte Carlo samples for the two types (one type, respectively) scenario. Fig. 3 shows the convergence rate of some metaparameters along the Markov chain (top-left panel for this mock catalogues). The convergence for $\widetilde{H}$ is typically fast, decorrelating in a few iterations. That is the result of the partially collapsed Gibbs sampler. The other shown parameters, related to the Gaussian mixture of Section 2.2.5, take substantially more

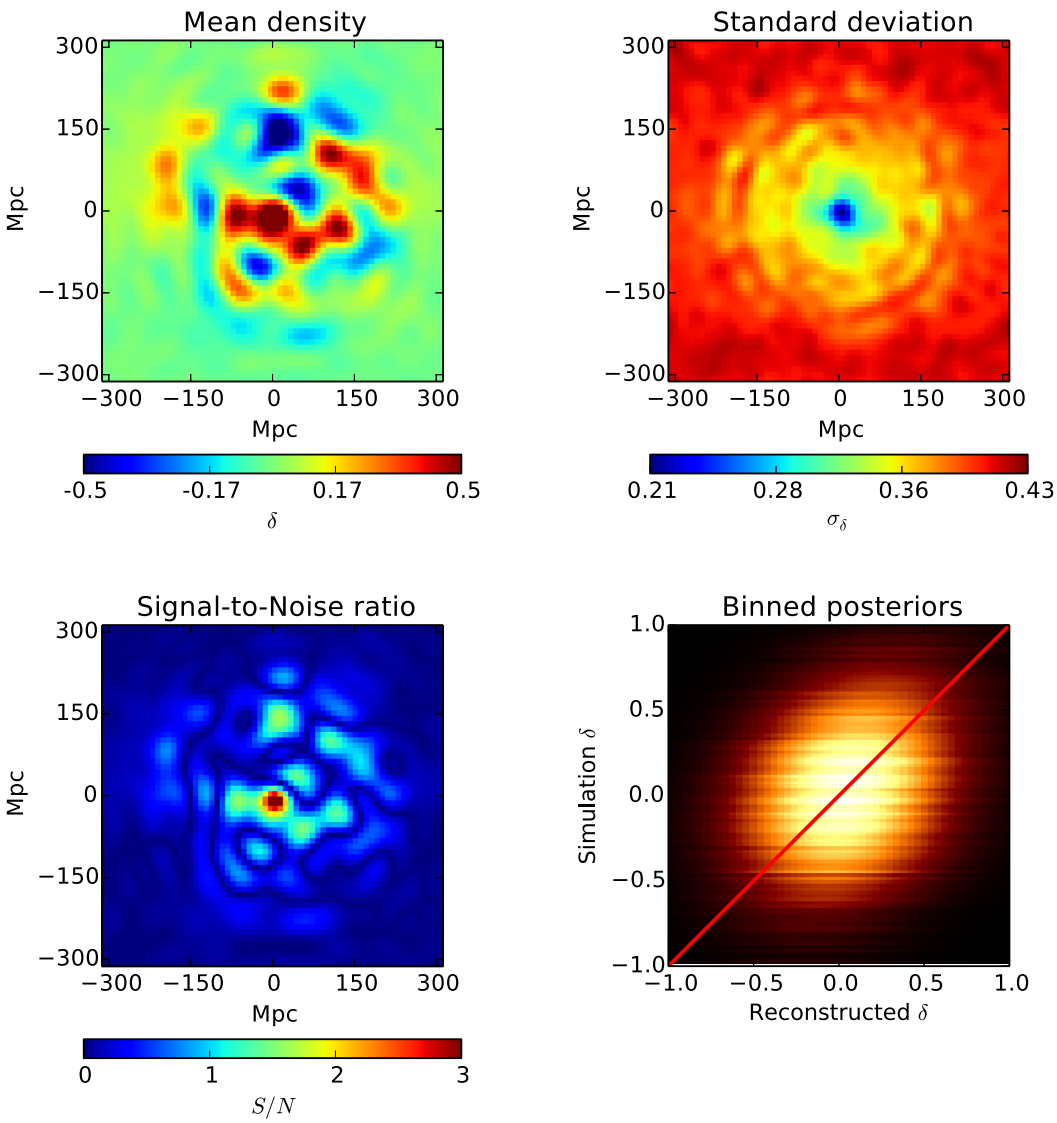

Figure 5. Results of the test on the mock catalogue based on Gaussian random field. All panels are related to density field distribution: central slice of the ensemble average (top-left panel), the variance expressed as a standard deviation (top-right panel), the resulting S/N (bottom-left panel).

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

steps to decorrelate, of the order of a few hundreds. This is expected as they are tied to a particular realization of the velocity field, contrary to other metaparameters which are obtained through direct marginalization over the velocity field. We can expect the convergence length to increase when the number of tracer is low or the noise is high as the uncertainty over the velocity field will increase and thus it takes more steps to explore the parameter space. We will see how it is the case with the other mock catalogues in the next section.

Fig. 4 gives a synthetic view of the velocity field component of the posterior distribution. The details of the panels have been given previously for Fig. 1. Comparing Figs 1–4 shows how much the relaxation of the assumption of fixed cosmology, noise and distances degrades the quality of the reconstructed velocities.

Fig. 5 gives a similar view as for the velocity field, but this time for the density i.e. the divergence of the velocity field. The view is the same as the one in Fig. 2, but this time we have all the parameters being sampled. This field is expected to be much noisier as it is indeed the case when looking at ensemble average quantities and S/N. In the centre, where S/N is the highest we are only at maximum at 3σ, while for peculiar velocities it was possible to go at higher S/N. The resulting comparison of the individual posterior distribution in the bottom-right panel shows this high uncertainty. Compared to Fig. 2, there has been a strong loss of information on the density field. It reaches a point where the binned posterior in the bottom-right panel shows no constraint at all from data.

In Fig. 6, I show the individual distributions of the metaparameters of the chain, i.e. the effective Hubble constant $\widetilde{H}$, the model error amplitude, the probability for each type of model error and the overall power spectrum normalization $A_{\rm S}$. For $\widetilde{H}$, $A_{\rm S}$ and the model error, a thick vertical red line shows the value used to generate the mock catalogue. I am also showing in shaded yellow the range of values compatible at 90 per cent with the data according to the model. As expected all distributions are compatible with the thick vertical line at their peak positions, i.e. always well within the 90 per cent region of the distribution. In the case of the model error amplitude (upper-right panel), we

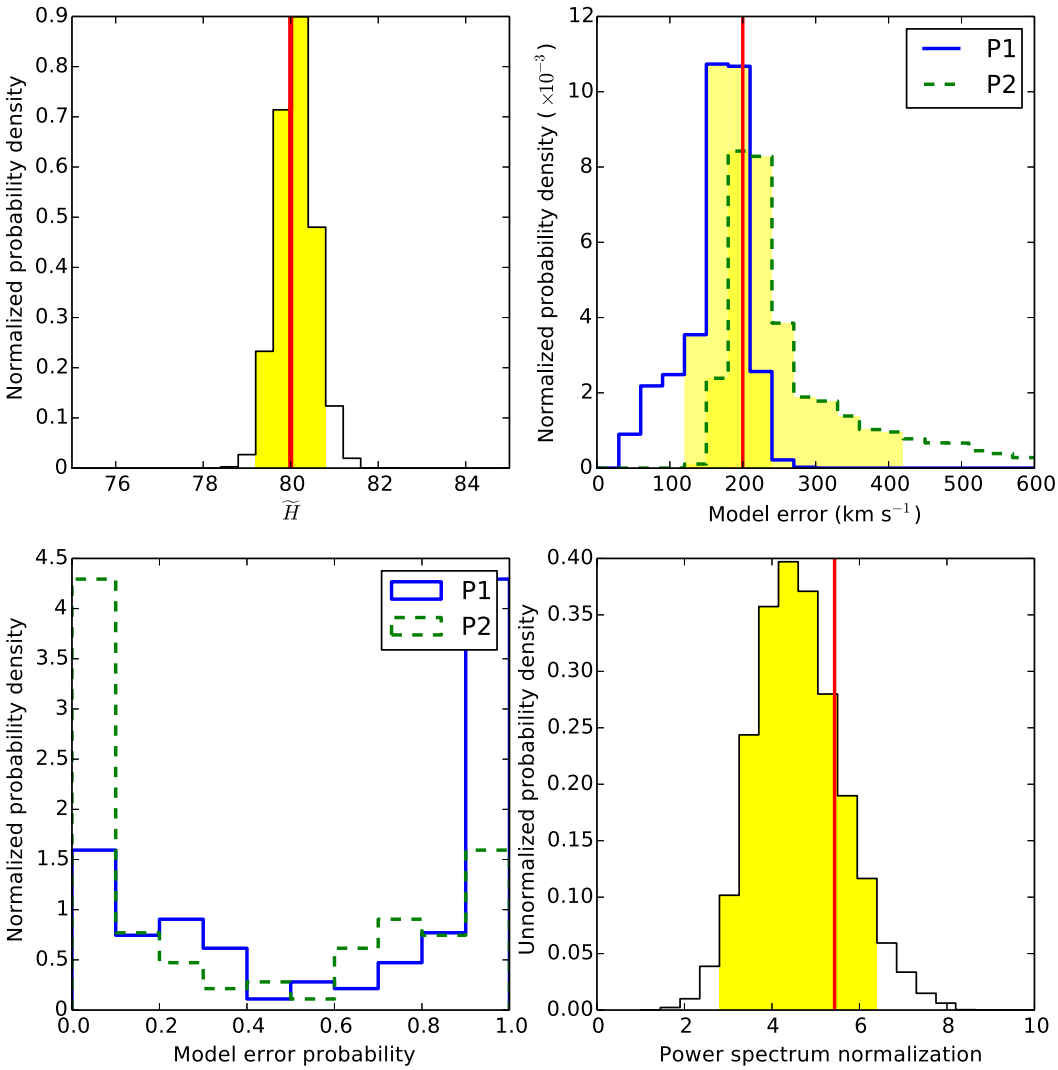

Figure 6. Results of the test on the mock catalogue based on Gaussian random field. All panels show metaparameters posterior distributions: the zero-point calibration $\widetilde{H}$ (top-left panel), the extra small scale non-linearities $\{\sigma_{{\rm NL},\,k}\}$ (top-right panel), the probability of tracer type $\mathcal{P}$ (bottom-left panel), and the power spectrum normalization in units of 10^6 Mpc^3 (bottom-right panel). The yellow shaded regions highlights the range of parameters for which the probability of containing the actual value is ∼90 per cent.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

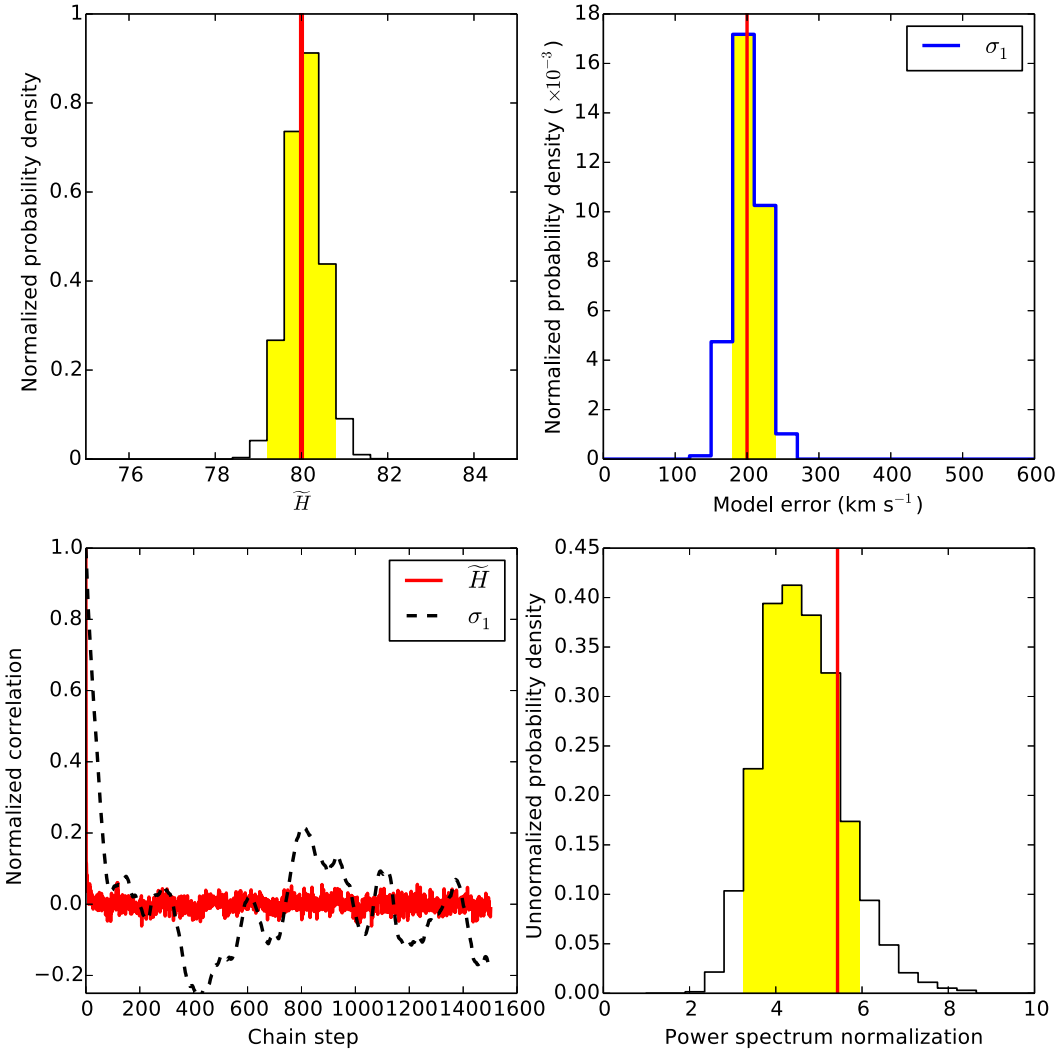

Figure 7. Analysis of the same mock catalogue based on Gaussian random field as presented in Fig. 6, but this time restricted to a single type of tracer. The panels present similar quantities: the zero-point calibration $\widetilde{H}$ (top-left panel), the extra small scale non-linearities $\sigma_{\rm NL}$ (top-right panel), the power spectrum normalization $A_{\rm S}$ in units of 10^6 Mpc^3 (bottom-left panel). Additionally, the autocorrelation of the chains is given in the bottom-left panels. The yellow shading also represents the range of parameters accepted at 90 per cent probability.

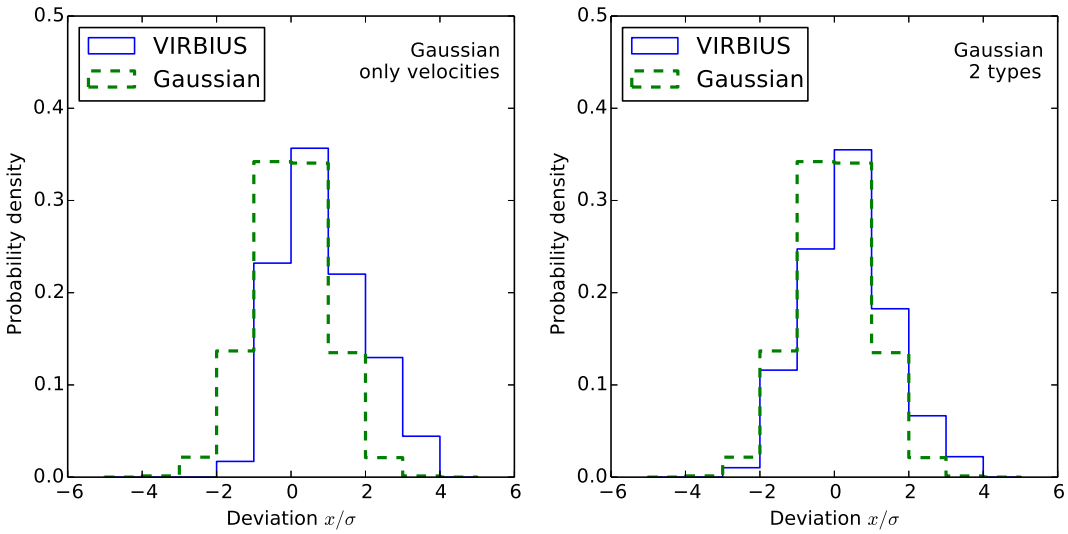

Figure 8. Distribution of the velocity residuals $\epsilon_{\rm r}$ between the simulation and the reconstructed mean component of the peculiar velocities for the different experiments on mock catalogues generated from Gaussian random fields. Left-hand panel refer to the same experiment as Fig. 1 and right-hand panel to Fig. 4.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

are faced with two distributions overlapping with the fiducial value. Finally, the posterior distribution shown in the bottom-left panel do not indicate a strong imbalance between the different error model, which exactly corresponds to the significant overlap of the distributions in the top right panel. The results of the distributions obtained assuming a single type are given in Fig. 7. All the metadistributions are of course narrower than for the test with two types, especially for the distribution of $\sigma_{\rm NL}$. The width of the distribution for $A_{\rm S}$ is also visibly narrower. The peculiar velocity field, not represented here, is however mostly unaffected. In the test with two assumed types of tracers, the algorithm separated the data in two pieces: one including more than 93 per cent of the tracers 50 per cent of the time, and the others. This separation results in a small number of objects having high-velocity dispersion, which both skew the P1 distribution towards low value of $\sigma_{\rm NL}$ and allow a the P2 distribution to venture to very high values of this same $\sigma_{\rm NL}$. No obvious bias is visible in all the distributions. The autocorrelation of the chain is shorter by a factor ∼2. These tests indicate that the algorithm and the software works as expected on an idealistic test case.

3.2 Halo based mock catalogue

In this Section, I consider more realistic, but more complicated mock catalogue to which I apply the methodology developed in this article. Two mock catalogues are considered, both based on the haloes of a cosmological pure dark matter *N*-body simulation. This simulation have been computed using the following cosmological parameters: $\Omega_{\rm M} = 0.30$, $\Omega_{\rm b} = 0.045$, $\Omega_\Lambda = 0.70$, $H = 80$ km s^{-1} Mpc^{-1}, $\sigma_8 = 0.80$. The volume covered by the simulation is a cube with a side of 500 h^{-1} Mpc with 512^3 particles. From this simulation, haloes were extracted using the ROCKSTAR software (Behroozi, Wechsler & Wu-Y. 2013). The minimum halo size have been kept to its default value of 10, checking that particles are effectively bound. The total number of haloes of the simulation is thus 238 520. The two catalogues are created from the same simulation but choosing different selection properties. The first mock catalogue, nicknamed H3000, is generated by extracting randomly 3000

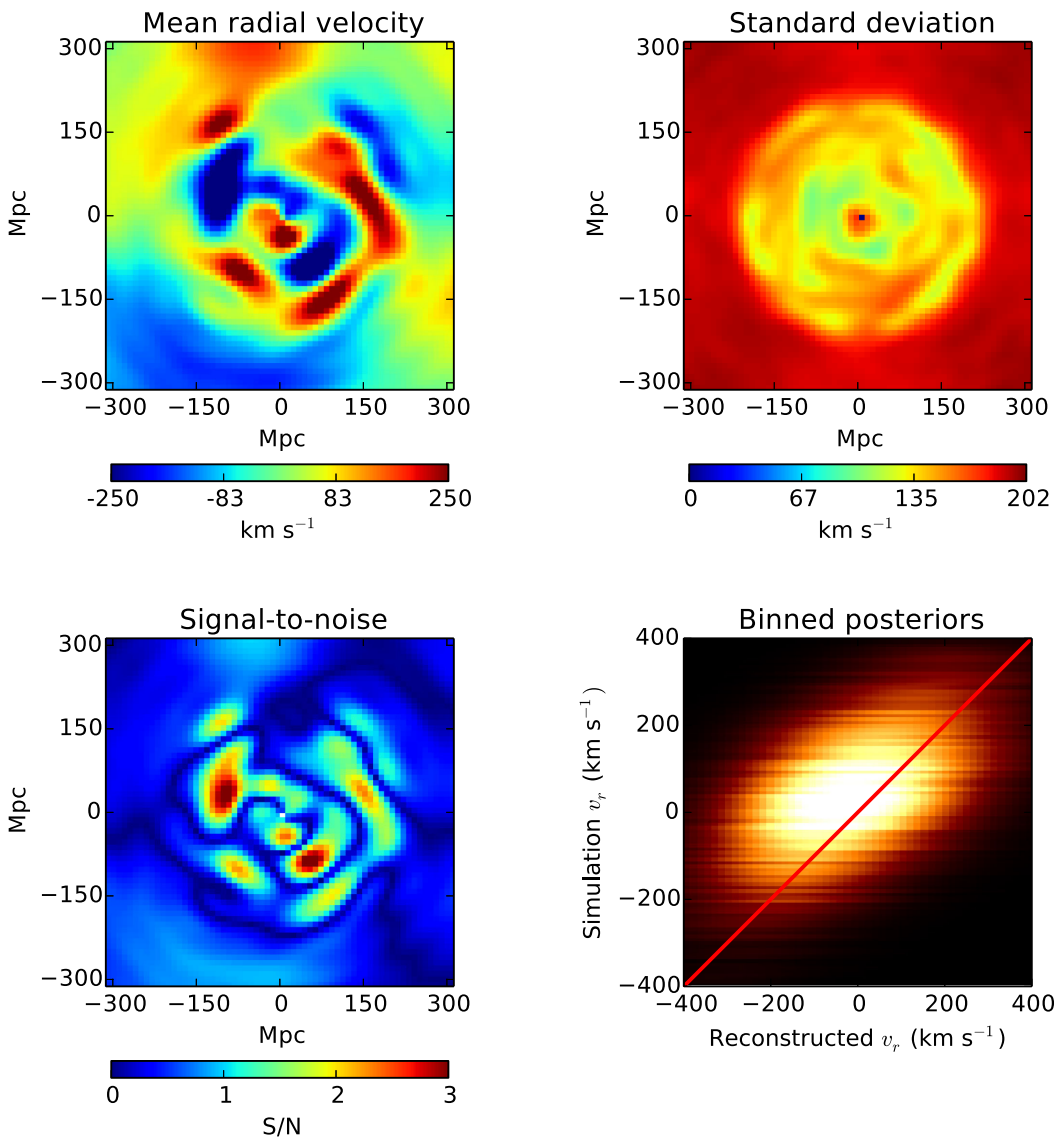

Figure 9. Same as Fig. 4 but for the H3000 mock catalogue.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

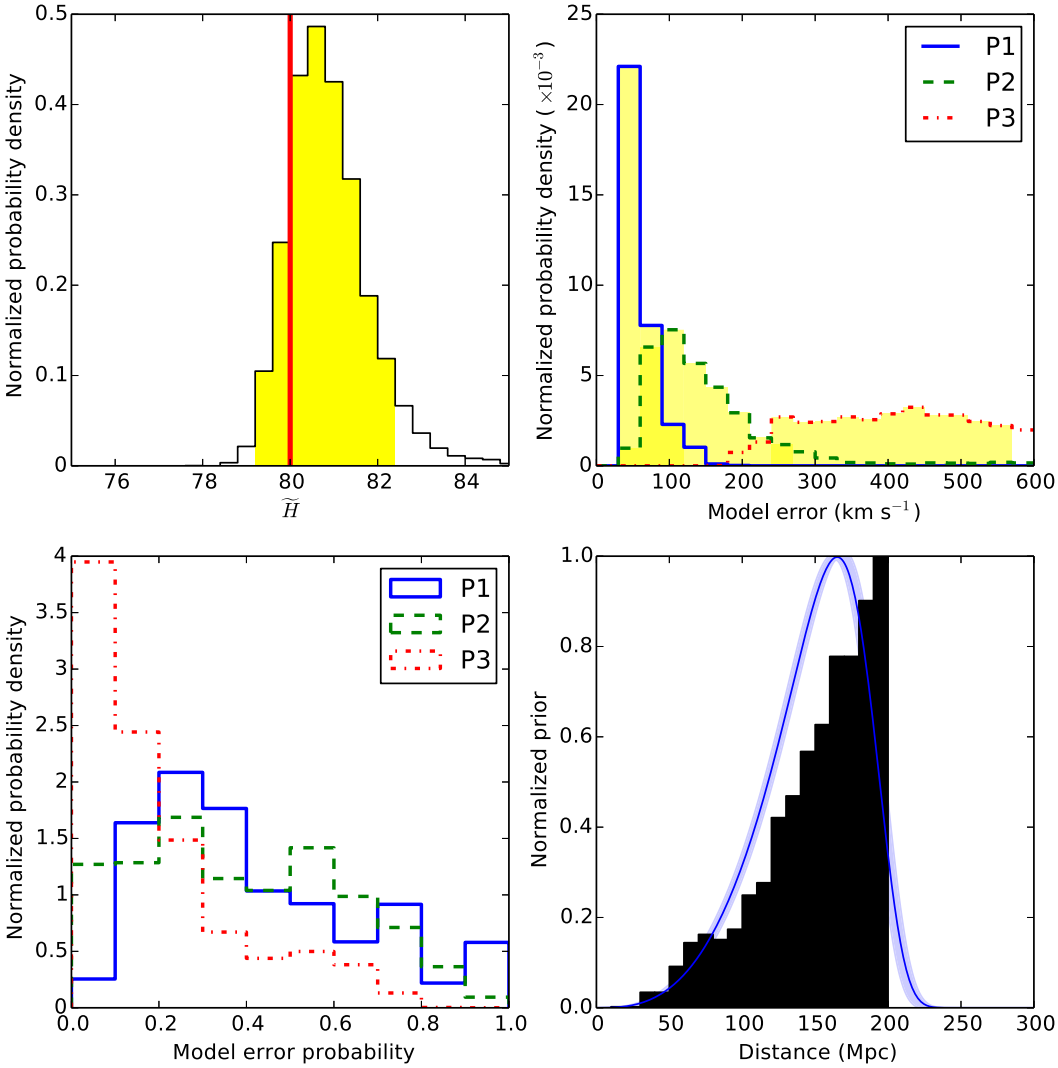

Figure 10. Results of the test on the the H3000 mock catalogue: metaparameters distributions: $\widetilde{H}$ (top left panel), $\{\sigma_{\mathrm{NL},k}\}$ (top right panel), $\mathcal{P}$ (bottom-left panel), selection function (bottom-right panel). Some bins are still clearly dominated by the noise due to the long correlation length of the chain.

haloes from the rockstar catalogue at a distance less than 200 h^{-1} Mpc of the centre of the simulation box. The chosen distance modulus error is $\sigma_\mu = 0.2$, typical for a Tully–Fisher relation. The second mock catalogue, nicknamed Hcomplex, is built from a more complex selection function, derived from the M/L relation used in Lavaux et al. (2008), equation (7). The principal condition for acceptance in the catalogue is to have $L/(d^L)^3 \leq 2.78 \times 10^4$, with L in solar luminosities and d the comoving distance in Mpc/h^{-1}. The selection is not supposed to be strictly realistic, but to mimic a realistic abundance for distance catalogue. This mock catalogue has 2000 tracers, with a distance modulus error $\sigma_\mu = 0.1$, typical of higher quality distance indicator like Supernovae (SNe) or Tip of the Red Giant branch (TRGB).

For these mock catalogues, I restrained from fitting the amplitude of the power-spectrum A_S at the same time as the other parameters. The problem comes from the degeneracy between A_S and the distribution of tracers of the velocity field. In halo catalogues, these two quantities are not independent as tracers are typically located in the peak of the density distribution. This tends to credit the velocity field with more power to try to homogenize the distribution of tracers. This problem arises when we are faced with data which needs a sufficiently precise prior on the density of tracers. I have run a chain for which the selection is exactly known i.e. homogeneous in luminosity distance space, but for which A_S is left free. The mean recovered value is 2.9 ± 0.5 times higher that the value used to make initial conditions. Thus, I postpone the resolution of this problem to future work. Here, I will limit myself to fix A_S to its fiducial value and investigate the rest of the parameter space.

The results are given in Figs 3, 9 and 10 for the H3000 halo mock catalogue, and in Figs 11 and 12 for the Hcomplex halo mock catalogue. The length of the chain is 10 824 for H3000 and 3905 for Hcomplex. The convergence test is shown in Fig. 3 (top right for H3000, bottom middle for Hcomplex). The convergence of the parameters of the Gaussian mixture is quite slow for H3000 due to the substantial uncertainty on the velocity field potential. This triggers a large correlated exploration of distances and model error parameters. The chain concerning H3000 has 10825 samples, whereas the one concerning Hcomplex has 1796 and is clearly converged. However, for Hcomplex, this convergence is much faster, confirming the previous scenario. In that case, the tracer density is sufficiently high to largely reduce velocity

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

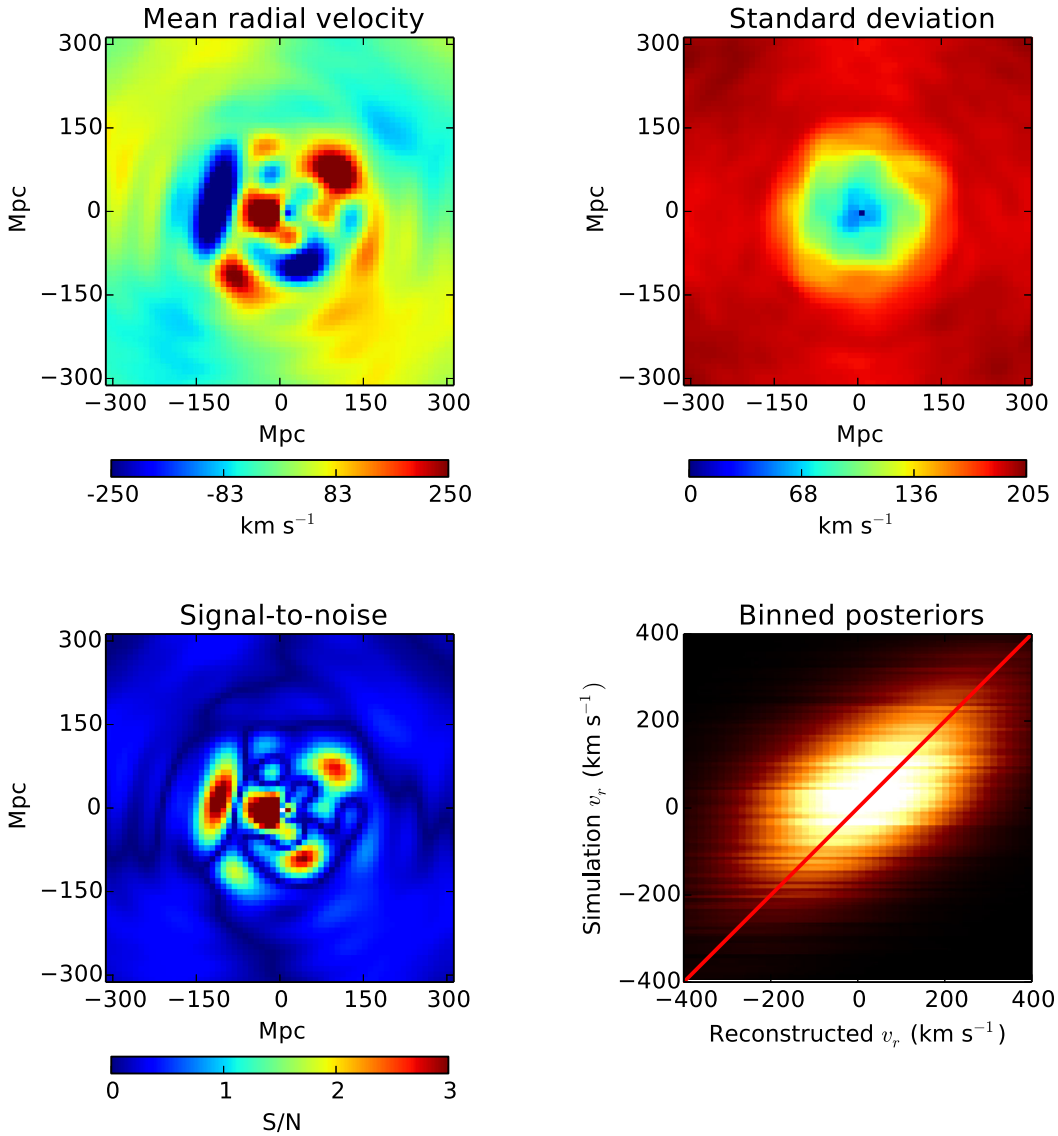

Figure 11. Same as Fig. 4 but for the Hcomplex mock catalogue.

field uncertainties in the effective volume covered by the catalogue. That statement can actually be generalized. If the catalogues of tracer is dense, then the velocity potential will be quite tightly constrained which does not leave much freedom for the auxiliary metaparameter like $\Sigma_{\rm NL}$, which in turn reduce possibilities for the reconstructed true distances.

The panels of the other figures are showing the same quantities as the ones in Figs 4 and 6, with the exception of the bottom-right panel of Figs 10 and 12 which shows the selection function of galaxies in the catalogue. In these same panels, I show the actual distribution of the galaxies for H3000 (a simple d^2 law) and Hcomplex as a function of luminosity distance. I note that the selection function for H3000 has been fitted by VIRBIuS according to prescription. However, as there is a sharp cut at 200 Mpc in the mock catalogues it is not able to reproduce this fairly. The parameter most affected by this discrepancy is in principle $\widetilde{H}$. We see that, within error bars, it does introduce significant bias. On the other hand, the galaxy population of Hcomplex is fairly represented by the selection function fitted by VIRBIUS (Fig. 12), and $\widetilde{H}$ is measured without any systematic effect.

The application of this Bayesian methodology on the two mock catalogues is satisfactory. All the posterior distributions are in agreement compared to the velocities of the simulations and the metaparameters are in agreement with the input value, such as $\widetilde{H}$ and the selection function. The analysis also yields that the haloes can be classified into two main populations: the results on the measurement of $\Sigma_{\rm NL}$, assuming three populations, is quite clear in particularly for the top right panel of Fig. 12. There is a population of haloes with a model error of $\sim$400 km s^{-1}and another at $\sim$100 km s^{-1}. The probability assigned to each model is not clear, all models getting a share of $\sim$30 per cent. Which means a probability of $\sim$66 per cent for the low-velocity dispersion model and $\sim$33 per cent for the high-velocity dispersion one. It is possible to run a meta-analysis for which the chain is run assuming different maximum number of populations. The likelihood of the data, marginalized according to sampled parameters, given the population model can then be computed alongside the Bayes factor between the different models.

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

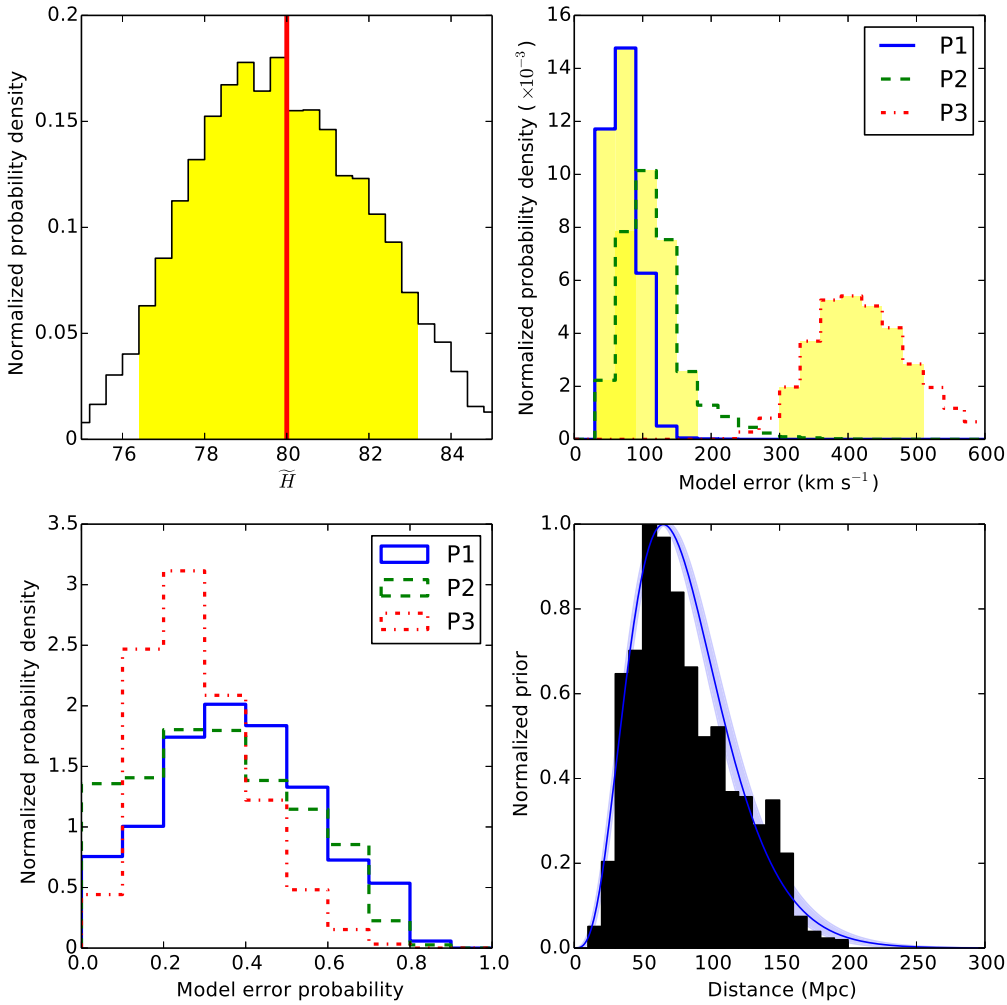

Figure 12. Same as Fig. 10 for the Hcomplex mock catalogue.

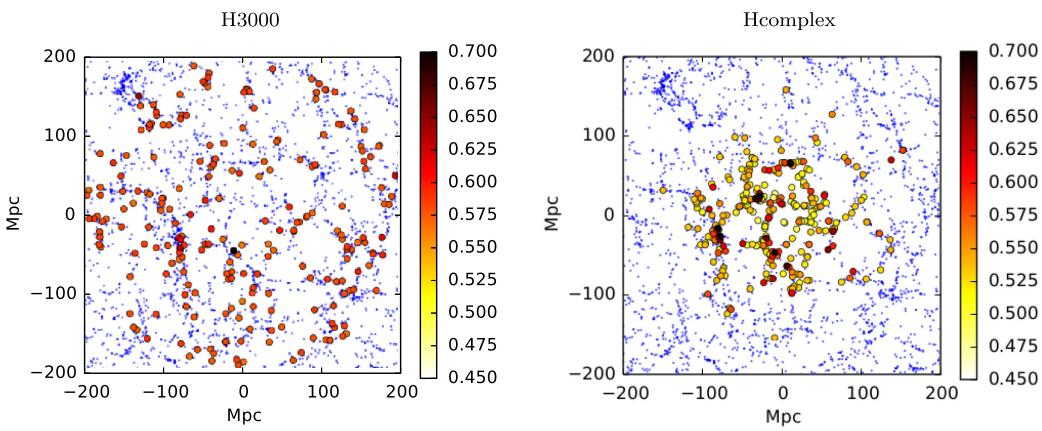

Figure 13. Results of the automatic classification for the two halo mock catalogues (H3000 in the left-hand panel, Hcomplex in the right-hand panel). I am showing a thin slice of both the mock catalogues and the simulations. The blue points correspond to the haloes of the original simulation, while the discs correspond to the mock catalogues. The disc are coloured according to the probability that a tracer has a velocity dispersion greater than 100 km s^{-1}.

This shows the power of this Bayesian methodology: in addition to allowing a consistent reconstruction of peculiar velocities, it is now possible to quantitatively measure the probability that a halo (or a galaxy in real distance catalogues) belongs to a high-velocity dispersion population or not in observations. This can be the case if the halo hosting the galaxy is a substructure for example. It could also be a supernova deeply embedded in the galaxy potential well. In Fig. 13, I show an attempt at classifying the haloes depending on their probability of belonging to one or the other population. In this figure, a slice of the two catalogues are shown: in the left-hand panel H3000, and in the

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

right-hand panel Hcomplex. The haloes of the simulation are represented as small blue points. The haloes selected to be part of the mock catalogues are shown as coloured discs. The colour of the discs is assigned depending on whether the probability that it has a high-velocity dispersion i.e. greater than 100 km s^{-1}. While in the H3000 catalogue, it is difficult to separate two populations, in the Hcomplex catalogue two populations clearly emerge spatially. The darkly coloured discs are qualitatively more likely to be at knots of the cosmic web, compared to the lightly coloured discs. Of course, as indicated early on in Section 2, the other main interest of the classifier consists in the natural removal of potential outliers from the observational data sets. Typically, the outliers will be put in separate type with a huge velocity dispersions which will remove any effect that they could have on the fitting of the velocity field.

4 CONCLUSION

The advent of new large distance catalogues, like 6dFv (Campbell et al. 2014) or Cosmic Flows-2 (Tully et al. 2013) is opening a new era in cosmic velocity field analysis. This era will continue with TAIPAN (Beutler et al. 2011) and WALLABY (Duffy et al. 2012). To face this avalanche of new data and get the most of them, new methods of analysing redshift and distance surveys are required.

For this work, I develop, implement and test a Bayesian algorithm to reconstruct peculiar velocity field from distance catalogues. This method is based on a Bayesian formulation of the reconstruction problem, assuming that the data follows the model given in equation (10). Additionally, all cosmological expansion corrections are already built in. I have shown that in the case of mock catalogues following perfectly the data model, then the algorithm manages to recover all quantities in an unbiased way. However, for more realistic mock catalogues taken from *N*-body simulations, some bias is introduced due to approximation either in the model, in the used priors for distances and velocities. This bias is the most prominent for the amplitude of the power spectrum when the noise on distances is sufficiently large to have to rely on velocity field correlations to infer the position of tracers. Finally, this algorithm includes an automatic tracer classifier that makes it interesting to classify object: either as observational outliers, or as object belonging to collapse structures and thus exhibiting higher velocity dispersions.

While the results obtained here are promising, they are not the end of the development of this method. One of the next step would be to harden the fitting of the amplitude A_S of the density fluctuation power spectrum by adding a bias field in the distance prior. This bias field would need to be stochastically related to $\Delta\Theta \sim -\delta_m$ as it is more likely to have haloes in locations where the flow is converging. However, the sampling of this bias field would be substantially more complex, and I am postponing this to later work. Simpler improvements would consist in allowing several samples with different radial selection function and fitting the inverse Tully–Fisher relation at the same time as the other parameters. Finally, the density field sampled using this algorithm could be used as a prior for either ARES (Jasche & Wandelt 2013b) or BORG (Jasche & Wandelt 2013a; Jasche, Leclercq & Wandelt 2015) algorithms of reconstruction of density field from galaxy spectroscopic surveys.

ACKNOWLEDGEMENTS

I would like to thank Michael J. Hudson for many useful discussions and encouragements to pursue this project when it was in its early stage. I also would like to thank Jens Jasche for many useful discussions on Bayesian statistics. I thank Stephen Turnbull for reading an early version of this manuscript. I thank the anonymous referee for the detailed reading, and the suggested corrections which have greatly improved the manuscript.

I acknowledge support from CITA National Fellowship and financial support from the Government of Canada Post-Doctoral Research Fellowship. Research at Perimeter Institute is supported by the Government of Canada through Industry Canada and by the Province of Ontario through the Ministry of Research and Innovation.

This work was granted access to the HPC resources of The Institute for scientific Computing and Simulation financed by Region Île-de-France and the project Equip@Meso (reference ANR-10-EQPX-29-01) overseen by the French National Research Agency (ANR) as part of the 'Investissements d'Avenir' program.

This work has been done within the Labex ILP (reference ANR-10-LABX-63) part of the Idex SUPER, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme Investissements d'avenir under the reference ANR-11-IDEX-0004-02.

Special thanks go to Stéphane Rouberol for his support during the course of this work, in particular for guaranteeing flawless use of all required computational resources.

I acknowledge financial support from 'Programme National de Cosmologie and Galaxies' (PNCG) of CNRS/INSU, France.

REFERENCES

Aaronson M., Huchra J., Mould J., Schechter P. L., Tully R. B., 1982, ApJ, 258, 64
Aaronson M., Bothun G., Mould J., Huchra J., Schommer R. A., Cornell M. E., 1986, ApJ, 302, 536
Anderson C., Dahleh M., 1996, SIAM J. Sci. Comput., 17, 913
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Bertschinger E., Dekel A., 1989, ApJ, 336, L5
Beutler F. et al., 2011, MNRAS, 416, 3017
Bonvin C., Durrer R., Gasparini M. A., 2006, Phys. Rev. D, 73, 023523
Campbell L. et al., 2014, MNRAS, 443, 1231

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

Courtois H. M., Tully R. B., Makarov D. I., Mitronova S., Koribalski B., Karachentsev I. D., Fisher J. R., 2011, MNRAS, 414, 2005
Davis T. M., Scrimgeour M. I., 2014, MNRAS, 442, 1117
Dekel A., Bertschinger E., Faber S. M., 1990, ApJ, 364, 349
Dekel A., Eldar A., Kolatt T., Yahil A., Willick J. A., Faber S. M., Courteau S., Burstein D., 1999, ApJ, 522, 1
Dempster A. P., Laird N. M., Rubin D. B., 1977, J. R. Stat. Soc. B, 39, 1
Duffy A. R., Meyer M. J., Staveley-Smith L., Bernyk M., Croton D. J., Koribalski B. S., Gerstmann D., Westerlund S., 2012, MNRAS, 426, 3385
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Feix M., Nusser A., Branchini E., 2014, JCAP, 9, 19
Geman S., Geman D., 1984, IEEE Trans. Pattern Anal. Mach. Intell., 6, 721
Gorski K., 1988, ApJ, 332, L7
Hastings W. K., 1970, Biometrika, 57, 97
Hinshaw G. et al., 2012, 208, 19
Hoffman Y., Ribak E., 1991, ApJ, 380, L5
Hoffman Y., Ribak E., 1992, ApJ, 384, 448
Hubble E., Humason M. L., 1931, ApJ, 74, 43
Hui L., Greene P. B., 2006, Phys. Rev. D, 73, 123526
Jaffe A. H., Kaiser N., 1995, ApJ, 455, 26
Jasche J., Wandelt B. D., 2013a, MNRAS, 432, 894
Jasche J., Wandelt B. D., 2013b, ApJ, 779, 15
Jasche J., Leclercq F., Wandelt B. D., 2015, JCAP, 1, 36
Johnson A. et al., 2014, MNRAS, 444, 3926
Kolatt T., Dekel A., 1997, ApJ, 479, 592
Lavaux G., Mohayaee R., Colombi S., Tully R. B., Bernardeau F., Silk J., 2008, MNRAS, 383, 1292
Liu J. S., Wong W. H., Kong A., 1994, Biometrika, 81, 27
Lynden-Bell D., Faber S. M., Burstein D., Davies R. L., Dressler A., Terlevich R. J., Wegner G., 1988a, ApJ, 326, 19
Lynden-Bell D., Faber S. M., Burstein D., Davies R. L., Dressler A., Terlevich R. J., Wegner G., 1988b, ApJ, 326, 19
Macaulay E., Feldman H., Ferreira P. G., Hudson M. J., Watkins R., 2011, MNRAS, 414, 621
Macaulay E., Feldman H. A., Ferreira P. G., Jaffe A. H., Agarwal S., Hudson M. J., Watkins R., 2012, MNRAS, 425, 1709
Masters K. L., Springob C. M., Haynes M. P., Giovanelli R., 2006, ApJ, 653, 861
Metropolis N., Ulam S., 1949, J. Am. Stat. Assoc., 44, 335
Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, J Chem. Phys., 21, 1087
Nusser A., Davis M., 2011, ApJ, 736, 93
Nusser A., Branchini E., Davis M., 2011, ApJ, 735, 77
Nusser A., Branchini E., Davis M., 2012, ApJ, 744, 193
Pearson K., 1894, Phil. Trans. R. Soc. 185, 71
Peebles P. J. E., 1980, The Large-scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Planck Collaboration XVI, 2013, A&A, 571, A16
Pyne T., Birkinshaw M., 2004, MNRAS, 348, 581
Sasaki M., 1987, MNRAS, 228, 653
Springob C. M., Masters K. L., Haynes M. P., Giovanelli R., Marinoni C., 2007, ApJS, 172, 599
Springob C. M., Masters K. L., Haynes M. P., Giovanelli R., Marinoni C., 2009, ApJS, 182, 474
Strauss M. A., Willick J. A., 1995, Phys. Rep., 261, 271
Tully R. B. et al., 2013, AJ, 146, 86
van Dyk D. A., Park T., 2008, J. Am. Stat. Assoc., 103, 790
Wandelt B. D., Larson D. L., Lakshminarayanan A., 2004, Phys. Rev. D, 70, 083511
Weinberg S., 1972, Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity. Wiley, New York
Willick J. A., 1994, ApJS, 92, 1
Willick J. A., Strauss M. A., Dekel A., Kolatt T., 1997, ApJ, 486, 629
Zaroubi S., 2002, MNRAS, 331, 901
Zaroubi S., Hoffman Y., Fisher K. B., Lahav O., 1995, ApJ, 449, 446
Zaroubi S., Zehavi I., Dekel A., Hoffman Y., Kolatt T., 1997, ApJ, 486, 21

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

APPENDIX A: ADAPTIVE SAMPLING FROM A NON-TRIVIAL ONE-DIMENSIONAL DISTRIBUTION

The canonical algorithm to sample from an arbitrary probability distribution is based on the cumulative probability function:

$$P(x \leq X) = \int_{-\infty}^{X} \mathrm{d}x \; p(x). \tag{A1}$$

If y is sampled from a uniform distribution bounded by [0, 1], then

$$x = P^{-1}(y), \tag{A2}$$

where P^{-1} is the inverse function of P, is distributed according to p. Unfortunately, in the cases of this work, none of the distributions are analytical, and even less integrable and invertible. We must rely on a numerical scheme for this. I consider different limiting cases to try to optimize the random number generation.

A1 Single modal sharply peaked or weakly multimodal

The first case that I consider is in the approximation that $p(x)$ has mostly a single sharp peak, with maybe some smaller insignificant subpeaks around this one. This is typically the case when sampling the cosmological hyper parameters like the Hubble constant H, the distance calibration $\widetilde{H}$ or the spurious noise on velocities $\sigma_{\rm NL}$. The evaluation of the posterior in these cases is costly, involving the inversions of large matrices with a size of the order the number of tracers in the original catalogue. It is thus required to make as few evaluations as possible. The position and sharpness of the peak is not clearly a priori known, and we have to rely on adaptivity. The algorithm proceeds as follow.

(i) It looks for the maximum $x_{\rm max}$ of the 1D posterior distribution using the Brent minimization algorithm specified in the GSL. The evaluate of the posterior are cached as the minimization proceeds.

(ii) From the cached values, it estimates the curvature by computing the average of $(x - x_{\rm max})^2$ from the cached values, using a linear interpolation of the logarithm of the posterior.

(iii) This curvature gives the typical sampling size required to capture the details of the posterior, which is now sampled regularly relative to $x_{\rm max}$. The evaluation is executed till the ratio between the edge value of the posterior and its peak value is less than some threshold. This procedure is run in parallel on shared memory computers.

(iv) The final sample is then generated using a rejection sampler, assuming the evaluated discrete distribution as proposal distribution.

This algorithm is robust and manages to generate sample in cases the distribution extends over several orders of magnitudes, while retaining accuracy of the target distribution. This part is specifically critical to ensure the stability of the final Markov Chain.

A2 Multimodal smoothly varying

The second case corresponds to the sampling of distances assuming a fixed large scale velocity field. Many distance posterior will be simple. However, some will be truncated (for example if we are on the edge of the survey), or severely multimodal in the case of a collapsing structure. For that reason, we cannot entirely apply the previous algorithm and we in place simplify it by removing the steps 1 and 2. In place, I just evaluate regularly the distance posterior in a sufficiently fine-grained way to resolve the evolution of the velocity field and the typical error bar on distance inferred from both $\sigma_{\rm NL}$ and the distance modulus error. The rejection sampler is then used again to clean the discreteness effects. This algorithm can only be used here because the evaluation of the posterior is very fast, contrary to the most cases considered in the previous section.

APPENDIX B: FOURIER–TAYLOR INTERPOLATION ON A NON-REGULAR MESH

The statistical algorithm described in this work makes use of tracers put at positions not located on a mesh. It is notably necessary to compute $\Psi(\boldsymbol{r})$ at any location in the volume of reconstruction. Of course $\Psi(\boldsymbol{r})$ is obtained through a Fourier synthesis operation described in equation (21) that we take now as a definition, in dimension d:

$$\Psi(\boldsymbol{r}) = \frac{1}{L^d}\sum_{q=1}^{N_q}\frac{i\boldsymbol{k}_q}{|\boldsymbol{k}_q|^2}\mathrm{e}^{i\boldsymbol{k}_q\cdot\boldsymbol{r}}\hat{\Theta}(\boldsymbol{k}_q). \tag{B1}$$

We make use of the idea originally developed by Anderson & Dahleh (1996) to have a fast interpolation of Fourier series. The Taylor expansion of trigonometric functions converges very rapidly, it is thus tempting to do a Taylor expansion of $\Psi(\boldsymbol{r})$ against the nearest grid point $\boldsymbol{g}$. The general expansion is the following:

$$\Psi_s(\boldsymbol{r}) = \frac{1}{L^d}\sum_{q=1}^{N_q}\frac{ik_s}{|\boldsymbol{k}_q|^2}\mathrm{e}^{i\boldsymbol{k}_q\cdot(\boldsymbol{r}-\boldsymbol{g})}\hat{\Theta}(\boldsymbol{k}_q)\mathrm{e}^{i\boldsymbol{k}_q\cdot\boldsymbol{g}} \tag{B2}$$

$$= \frac{1}{L^d}\sum_{q=1}^{N_q}\sum_{n=0}^{+\infty}\frac{i^n}{n!}\left(\sum_{j=1}^{d}(\boldsymbol{k}_q)_j(r_j-g_j)\right)^n\frac{ik_s}{|\boldsymbol{k}_q|^2}\hat{\Theta}(\boldsymbol{k}_q)\mathrm{e}^{i\boldsymbol{k}_q\cdot\boldsymbol{g}} \tag{B3}$$

$$= \frac{1}{L^d}\sum_{q=1}^{N_q}\sum_{n=0}^{+\infty}\frac{i^n}{n!}\left(\sum_{\{a_p\}}\prod_{p=1}^{d}\left(r_j-g_j\right)^{a_p}F(\boldsymbol{k}_q,n,\{a_p\})\right)\frac{ik_s}{|\boldsymbol{k}_q|^2}\hat{\Theta}(\boldsymbol{k}_q)\mathrm{e}^{i\boldsymbol{k}_q\cdot\boldsymbol{g}} \tag{B4}$$

$$= \sum_{n=0}^{+\infty}\frac{1}{n!}\sum_{\{a_p\}}\left(\prod_{p=1}^{d}\left(r_p-g_p\right)^{a_p}\right)Q_{s,n}(\{a_p\},\boldsymbol{g}), \tag{B5}$$

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

Table B1. This table gives the expression of $F(\boldsymbol{k}, n, \{a_p\})$ for the first three orders and for $d = 3$.

Order n	$\{a_p\}$	$F(\boldsymbol{k}, n, \{a_p\})$
$n = 0$	$a_1 = a_2 = a_3 = 0$	1
$n = 1$	$a_1 = 1, a_2 = a_3 = 0$	k_1
	$a_1 = 0, a_2 = 1, a_3 = 0$	k_2
	$a_1 = a_2 = 0, a_3 = 1$	k_3
$n = 2$	$a_1 = 2, a_2 = a_3 = 0$	k_1^2
	$a_1 = 1, a_2 = 1, a_3 = 0$	$2k_1k_2$
	$a_1 = 0, a_2 = 1, a_3 = 1$	$2k_2k_3$
	$a_1 = 1, a_2 = 0, a_3 = 1$	$2k_1k_3$
	$a_1 = 0, a_2 = 2, a_3 = 0$	k_2^2
	$a_1 = 0, a_2 = 0, a_3 = 2$	k_3^2

with

$$F(\boldsymbol{k}, n, \{a_p\}) = \binom{n}{a_1} \left(\prod_{p=2}^{d-1} \binom{a_p}{a_{p-1}} \right) \left(\prod_{p=1}^{d} k_p^{a_p} \right), \tag{B6}$$

$$\hat{Q}_{s,n}(\{a_p\}, \boldsymbol{k}) = i^n F(\boldsymbol{k}_q, n, \{a_p\}) \frac{ik_s}{|\boldsymbol{k}_q|^2} \hat{\Theta}(\boldsymbol{k}_q) \tag{B7}$$

$$Q_{s,n}(\{a_p\}, \boldsymbol{g}) = \frac{1}{L^d} \sum_{q=1}^{N_q} \hat{Q}_{s,n}(\{a_p\}, \boldsymbol{k}_q) \mathrm{e}^{i\boldsymbol{k}_q \cdot \boldsymbol{g}}. \tag{B8}$$

In the above, we have used the notation $\binom{n}{p} = n!/(p!(n-p)!)$. The set of indices $\{a_p\}$ are under the constraint $\sum_{p=1} a_p = n$. The multidimensional array Q_s is only the Fast Fourier synthesis of $\hat{Q}_{s,n}$ on a regular mesh. $\hat{Q}_{s,n}$ can be easily computed from the modes $\hat{\Theta}(\boldsymbol{k}_q)$. Finally, $\Psi_s(\boldsymbol{r})$ is obtained by taking all $Q_{s,n}$ and summing after multiplication by the adequate polynomial in the distance to the nearest grid point to $\boldsymbol{r}$. As an example, I give the expression of the cases $n = 0$, 1 and 2 and $d = 3$ in Table B1.

I note that the number of required FFTs goes quickly up as the number of elements in $\{a_p\}$ for a given n. This number is exactly equal to $\binom{n+d-1}{n}$. Fortunately, the series converges very quickly and for the problem studied in this work; it is sufficient to stop at $n = 3$. The time complexity of this algorithm becomes $\mathcal{O}(N_q) + \mathcal{O}(N_\mathrm{g} \log N_\mathrm{g})$ instead of $\mathcal{O}(N_q N_\mathrm{g} \log N_\mathrm{g})$ if the sum in equation (21) was carried out directly. The expected gain is huge when N_q becomes high, typically for large distance surveys of several thousands of objects.

APPENDIX C: FOURIER–TAYLOR WIENER FILTER

A similar problem as the one tacked in Appendix B is met when one computes the mean velocity field compatible with the velocity of a set of tracers as given by equation (53). The problem is the following, given a set of weights $\{w_j\}$ and positions $\{\boldsymbol{r}_j\}$, we need to determine the value of the modes $\hat{f}_m(\boldsymbol{k})$:

$$\hat{f}_m(\boldsymbol{k}) = P(|\boldsymbol{k}|) \frac{-ik_m}{k^2} \sum_{j=1}^{N_\mathrm{d}} w_j \hat{r}_{j,m} \mathrm{e}^{-i\boldsymbol{k}.\boldsymbol{r}_j} \tag{C1}$$

A priori, it is relatively simple and it can be achieved by direct summation for a time complexity of $\mathcal{O}(N_\mathrm{d} N_q \log N_q)$ where N_q is the number of modes $\boldsymbol{k}$. It turns out that this is costly because of the big number of trigonometric functions to evaluate. Additionally, if the number of tracers becomes significant, e.g. a few thousands, the filling of this array becomes quite costly. So I propose to use another Fourier–Taylor expansion to reduce the cost.

The equation (C1) may be rewritten as followed:

$$\hat{f}_m(\boldsymbol{k}) = P(|\boldsymbol{k}|) \frac{-ik_m}{k^2} \sum_{\boldsymbol{g}} \left(\sum_{j=1}^{N_\mathrm{d}} w_j \hat{r}_{j,m} \mathrm{e}^{-i\boldsymbol{k}.(\boldsymbol{r}_j - \boldsymbol{g})} \delta_\mathrm{K}(\boldsymbol{g} - \boldsymbol{r}_j^{(g)}) \right) \mathrm{e}^{-i\boldsymbol{k}.\boldsymbol{g}}, \tag{C2}$$

where $\delta_\mathrm{K}(\boldsymbol{x}) = \prod_j \delta_\mathrm{K}(x_j)$ is the Kronecker-delta $\delta_\mathrm{K}(a) = 1$ if and only if $a = 0$, $\boldsymbol{r}_j^{(g)}$ is the position of the grid point nearest to $\boldsymbol{r}_j$. The symbol $\sum_{\boldsymbol{g}}$ means that the summation is run on all the grid points. The Kronecker-delta is well

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022

defined as we are only considering a countable set of grid point position. Now we can expand the argument of the inner exponential:

$$\hat{f}_m(\boldsymbol{k}) = P(|\boldsymbol{k}|)\frac{-ik_m}{k^2}\sum_{\boldsymbol{g}}\sum_{n=0}^{+\infty}\left(\sum_{j=1}^{N_\mathrm{d}} w_j \hat{r}_{j,m}\frac{(-i)^n}{n!}\left(\boldsymbol{k}.(\boldsymbol{r}_j - \boldsymbol{r}_j^{(g)})\right)^n \delta_\mathrm{K}(\boldsymbol{g} - \boldsymbol{r}_j^{(g)})\right)\mathrm{e}^{-i\boldsymbol{k}.\boldsymbol{g}} \tag{C3}$$

$$= P(|\boldsymbol{k}|)\frac{-ik_m}{k^2}\sum_{n=0}^{+\infty}\frac{(-i)^n}{n!}\sum_{\{a_p\}} F(\boldsymbol{k}, n, \{a_p\})\hat{G}_m(\boldsymbol{k}, n, \{a_p\}) \tag{C4}$$

with

$$\hat{G}_m(\boldsymbol{k}, n, \{a_p\}) = \sum_{\boldsymbol{g}} G_m(\boldsymbol{g}, n, \{a_p\})\mathrm{e}^{-i\boldsymbol{k}.\boldsymbol{g}} \tag{C5}$$

$$G_m(\boldsymbol{g}, n, \{a_p\}) = \sum_{j=1}^{N_\mathrm{d}} w_j \hat{r}_{j,m}\left(\prod_{p=1}^{d}\left(r_j - r_j^{(g)}\right)^{a_p}\right)\delta_\mathrm{K}(\boldsymbol{g} - \boldsymbol{r}_j^{(g)}), \tag{C6}$$

with $F(\boldsymbol{k}, n, \{a_p\})$ as defined in equation (B6). Again, we can recognize a Fast Fourier analysis step in equation (C5). This analysis is run on a multidimensional array which has non-zero values only at grid points which are next to a tracer. The values of the field depends on moments of the distribution of tracers assigned to each grid elements. Finally, the time complexity is again now $\mathcal{O}(N_\mathrm{t}) + \mathcal{O}(N_q \log N_q)$. I note that the transforms defined in this appendix have more application to compute the Fourier transform of a function sampled on a non-regular grid, provided one is ready to remove the highest modes which are the most imprecise due to the Taylor expansion.

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

Downloaded from https://academic.oup.com/mnras/article/457/1/172/988703 by guest on 29 August 2022