Primordial power spectrum and cosmology from black-box galaxy surveys

Florent Leclercq,[1, a)] Wolfgang Enzi,[2, b)] Jens Jasche,[3, 4, c)] and Alan Heavens[1, d)]
[1)] *Imperial Centre for Inference and Cosmology (ICIC) & Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, United Kingdom*
[2)] *Max-Planck Institute for Astrophysics, Karl-Schwarzschild Strasse 1, D-85748 Garching, Germany*
[3)] *The Oskar Klein Centre, Department of Physics, Stockholm University, Albanova University Center, SE 106 91 Stockholm, Sweden*
[4)] *Excellence Cluster Universe, Technische Universität München, Boltzmannstrasse 2, D-85748 Garching, Germany*

(Dated: 24 September 2019)

We propose a new, likelihood-free approach to inferring the primordial matter power spectrum and cosmological parameters from arbitrarily complex forward models of galaxy surveys where all relevant statistics can be determined from numerical simulations, i.e. *black-boxes*. Our approach, which we call simulator expansion for likelihood-free inference (SELFI), builds upon approximate Bayesian computation using a novel effective likelihood, and upon the linearisation of black-box models around an expansion point. Consequently, we obtain simple "filter equations" for an effective posterior of the primordial power spectrum, and a straightforward scheme for cosmological parameter inference. We demonstrate that the workload is computationally tractable, fixed *a priori*, and perfectly parallel. As a proof of concept, we apply our framework to a realistic synthetic galaxy survey, with a data model accounting for physical structure formation and incomplete and noisy galaxy observations. In doing so, we show that the use of non-linear numerical models allows the galaxy power spectrum to be safely fitted up to at least $k_{\rm max} = 0.5\ h/{\rm Mpc}$, outperforming state-of-the-art backward-modelling techniques by a factor of ~ 5 in the number of modes used. The result is an unbiased inference of the primordial matter power spectrum across the entire range of scales considered, including a high-fidelity reconstruction of baryon acoustic oscillations. It translates into an unbiased and robust inference of cosmological parameters. Our results pave the path towards easy applications of likelihood-free simulation-based inference in cosmology. We have made our code pySELFI and our data products publicly available at http://pyselfi.florent-leclercq.eu.

I. INTRODUCTION

The cosmic large-scale structure constitutes one of the major sources of information for modern cosmology. According to the current paradigm, all observable structures originate from tiny primordial fluctuations, which evolved via gravitational amplification into the presently-observed cosmic web (see e.g. Peebles, 1980; Peacock, 1999). The statistics of the initial density field are measured to be extremely close to Gaussian-distributed (see e.g. Planck Collaboration, 2018a). As a consequence, the primordial matter power spectrum is a very powerful cosmological probe: it is a sufficient statistical summary under the assumption that fluctuations are Gaussian, and – even if this assumption is violated – it remains close to capturing all of the information for all models allowed by observations. A particularly important cosmological signature, imprinted on the primordial matter power spectrum at the time of recombination, is baryon acoustic oscillations (BAOs). It constitutes a fixed comoving length scale (a "standard ruler"), which, when measured at different cosmic times, gives information on the expansion history of the Universe, including the late-time era of accelerated expansion. BAOs are thus one of the main probes to determine the equation of state of a possible dark energy component (see e.g. Eisenstein, 2005; Albrecht *et al.*, 2006; Percival *et al.*, 2007). A large variety of early-Universe models exhibit a deterministic relation between physical parameters of interest and the primordial matter power spectrum. The latter is therefore an interesting intermediate product for cosmological analyses, allowing parameter inference and model selection to be performed *a posteriori* without (or with minimal) loss of information. It can be seen as a largely agnostic and model-independent parametrisation of cosmological theories, which relies only on weak assumptions (isotropy and gaussianity).

For a long time, measuring the cosmological matter power spectrum has been one of the main goals of galaxy survey data analysis. However, inferring its shape accurately is a challenging task. Various systematic effects such as redshift uncertainties, complex survey geometries, selection effects, missing observations and foreground contamination can greatly hinder the measurement and analysis (see e.g. Ross *et al.*, 2012; Jasche & Lavaux, 2017). This problem is particularly important for the next generation of optical surveys, such as pro-

[a)] Electronic mail: florent.leclercq@polytechnique.org; http://www.florent-leclercq.eu/
[b)] Electronic mail: enzi@mpa-garching.mpg.de
[c)] Electronic mail: jens.jasche@fysik.su.se
[d)] Electronic mail: a.heavens@imperial.ac.uk

vided by ESA's Euclid space mission or the Large Synoptic Survey Telescope (LSST), which are expected to be dominated by systematic rather than statistical uncertainty (see e.g. Laureijs *et al.*, 2011; LSST Science Collaboration, 2012). Even if systematic effects arising from the survey strategy were fully understood and controlled, many theoretical challenges would still be present: galaxy biasing, anisotropic clustering in redshift space, and non-linear structure growth at late times, which reduce the detectability of cosmological signatures such as BAOs (see e.g. Meiksin, White & Peacock, 1999; Eisenstein, Seo & White, 2007). Because of the limited reliability of data models, fits of the galaxy power spectrum focus on linear and mildly non-linear scales, using typically a largest wavenumber of $k_{\mathrm{max}} = 0.3\ h/\mathrm{Mpc}$ (e.g. Ross *et al.*, 2015). However, the number of modes used in the analysis scales as k^3_{max}, meaning that any improvement of data models at small scales (such as what can be achieved via numerical simulations instead of perturbation theory) gives access to much more cosmological information.

As a response to theoretical and observational challenges, many approaches to measure the power spectrum have been proposed. They can be divided into two broad categories: backward-modelling approaches (often associated with frequentist statistics, counting the frequencies of measurements in mock catalogues, and associated covariance matrices) and likelihood-based forward-modelling approaches (often associated with Bayesian statistics). Backward-modelling approaches suggest to directly account and correct for relevant effects in observational data as a pre-processing step. For example, the BAO "reconstruction" technique removes redshift-space distortions and corrects the density field from bulk motions using the inverse Zel'dovich approximation (e.g. Eisenstein *et al.*, 2007; Padmanabhan *et al.*, 2012; Doumler *et al.*, 2013; Burden, Percival & Howlett, 2015; White, 2015). The end product is an estimator for the primordial matter power spectrum, which can be close to optimal if all relevant effects are modelled (Smith & Marian, 2015; Seljak *et al.*, 2017). As backward-modelling approaches require substantial expert knowledge input, results are often largely model-dependent and with difficult propagation of uncertainties. Importantly, these approaches often rely on fiducial values for the parameters that are the target of the analysis. More recently, several thorough forward-modelling Bayesian approaches have been proposed to jointly infer the three-dimensional matter density field and its power spectrum from galaxy observations, while properly accounting for uncertainties and systematics (Jasche *et al.*, 2010; Jasche & Wandelt, 2013b; Jasche & Lavaux, 2015; Granett *et al.*, 2015; Jasche & Lavaux, 2017). Considerable effort is also put into reconstructing the primordial density field from present observations (Jasche & Wandelt, 2013a; Wang *et al.*, 2013, 2014; Jasche, Leclercq & Wandelt, 2015; Lavaux & Jasche, 2016; Jasche & Lavaux, 2019; Bos, Kitaura & van de Weygaert, 2019). These methods are likelihood-based, meaning that they solve the exact inference problem by sampling from the target distribution via sophisticated Markov Chain Monte Carlo (MCMC) methods (Gibbs sampling and/or Hamiltonian Monte Carlo). In order to make the likelihood tractable, they have to involve approximations of the data model.

This paper has similar scientific aims but follows a different spirit: the presented method uses likelihood-free forward-modelling. It introduces one variant of approximate Bayesian computation (ABC) and treats the data model as a *black-box* simulator, i.e. performs inference without necessity to incorporate any knowledge of the data-generating processes into the analysis. This feature renders ABC ideal to infer the primordial matter power spectrum with arbitrarily complex models of galaxy surveys, including a physical treatment of structure formation and the details of observational processes, which cannot be trivially accounted for in likelihood-based statistical approaches. A popular ABC algorithm is likelihood-free rejection sampling, often coupled to Population Monte Carlo (e.g. Ishida *et al.*, 2015; Akeret *et al.*, 2015; Jennings & Madigan, 2017). More sophisticated approaches known as DELFI (Alsing, Wandelt & Feeney, 2018) and BOLFI (Leclercq, 2018) have also been recently introduced in cosmology. The ABC approach introduced in this work differs from all of the above in two aspects: (*i*) it allows the treatment of a much larger number of parameters (one hundred in this work), which correspond to primordial power spectrum amplitudes at different wavenumbers. To do so, (*ii*) it assumes the availability of a reasonably good guess of the target parameters, based on previous observations. This situation is fairly typical in cosmology, where the allowed space for parameters is already strongly constrained by previous experiments such as the Planck satellite (Planck Collaboration, 2016, 2018b). Under these assumptions, we derive an effective likelihood for the problem. We use an expansion point in parameter space and linearise the black-box around it. The use of finite differencing to compute the gradient of the black-box makes evaluations of the effective likelihood computationally feasible. When further assuming that the prior is Gaussian, we find that the effective posterior distribution for the primordial power spectrum is a Gaussian with mean and covariance matrix given by two simple "filter equations" (equations (25) and (26)), which constitute the main result of this work. Finally, we show how to infer parameters of specific cosmological models using the linearised black-box. We propose to call the algorithm "simulator expansion for likelihood-free inference" (SELFI).

In order to illustrate the performance of the method, we apply it to a black-box which emulates realistic cosmological data. This black-box is built using SIMBELMYNË (Leclercq, Jasche & Wandelt, 2015), a hierarchical probabilistic simulator to generate synthetic galaxy survey

data.[1] The data model involves a full cosmological N-body simulation (performed using our implementation of COLA, Tassev, Zaldarriaga & Eisenstein, 2013) to evolve the three-dimensional initial density field. It includes a treatment of galaxy bias, redshift-space distortions, survey geometry, selection effects, and instrumental noise. The statistical summary chosen is the estimated power spectrum of the galaxy number count field, as is standard in large-scale structure data analysis, but can be readily extended to include more information. As a result, the inferred primordial matter power spectrum is unbiased across the entire range of Fourier modes considered, and includes in particular BAOs, which were not included in the expansion point. Our analysis demonstrates that by using a fully numerical data model in conjunction with our statistical approach, one can safely fit the galaxy power spectrum even far in the non-linear regime, up to at least $k_{\rm max} \approx 0.5\ h/\mathrm{Mpc}$, which provides a factor of ~ 5 increase in the number of modes used, with respect to state-of-the-art backward-modelling techniques. We stress that any possible refinement of the data model used in this work does not change the statistical method, and therefore does not affect the validity of the previous statement.

This paper is organised as follows. In section II, we discuss the statistical method and derive the equations for black-box simulation-based inference of the primordial matter power spectrum and cosmological parameters. In section III, we describe the data-generating model used to test our method. The results obtained by combining the two are discussed in IV. We discuss the application of our method and prospects for cosmological data analysis, and provide our conclusions in section V. Details of the statistical derivations are given in the appendices.

II. METHOD

This section describes our method for simulation-based inference of the primordial matter power spectrum and cosmological parameters from black-box galaxy surveys. In section II A, we design an effective likelihood for black-box models. In section II B, we exploit previous knowledge, as could have been obtained by earlier cosmological probes, in order to linearise the black-box. We discuss the parametrisation of the primordial matter power spectrum and its prior distribution in section II C. The equations for the effective posterior distribution are given in section II D. In section II E, we describe how to optimally choose the hyperparameters appearing in the prior. The inference of cosmological parameters from the linearised black-box is discussed in section II F.

[1] SIMBELMYNË is publicly available at https://bitbucket.org/florent-leclercq/simbelmyne.

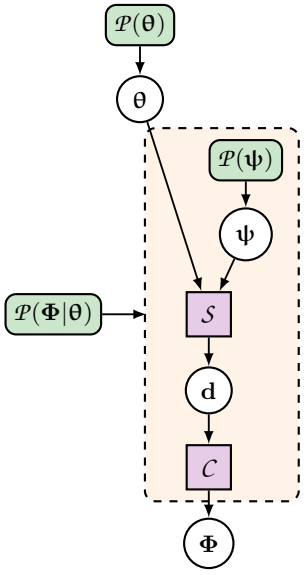

FIG. 1. Hierarchical representation of the black-box model used in this work. The rounded green boxes represent probability distributions and the purple square represent deterministic functions. The variables are $\boldsymbol{\theta}$ (the target parameters), $\boldsymbol{\psi}$ (the nuisance parameters), $\mathbf{d}$ (the full data), $\boldsymbol{\Phi}$ (the summary statistics). The orange dashed rectangle represents the data-generating process, it gives the true (unknown) likelihood $L(\boldsymbol{\theta})$ when $\boldsymbol{\Phi} = \boldsymbol{\Phi}_{\rm O}$ (the summary statistics of the observations).

A. Design of an effective likelihood for black-box models

Table I provides an overview of the different variables appearing in this section and their interpretation in the context of galaxy survey data analysis.

1. The data model

We assume given a black-box model that provides realistic predictions for artificial observations when provided with all necessary input parameters. These consist of the target vector $\boldsymbol{\theta} \in \mathbb{R}^S$ parametrising the primordial power spectrum, and of nuisance parameters $\boldsymbol{\psi} \in \mathbb{R}^T$, independent of $\boldsymbol{\theta}$. Nuisance parameters account for the entire stochasticity of the data model, such as initial phases, noise realisations, sample variance, etc. In our case, nuisance parameters will be all the random numbers generated by the galaxy survey simulator for initial conditions and instrumental noise; there are typically $\mathcal{O}(10^7)$ of those. Once realisations of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are specified, the output of the simulation $\mathbf{d} \in \mathbb{R}^D$ is a deterministic numerical function $\mathcal{S}$, i.e. $\mathcal{P}(\mathbf{d}|\boldsymbol{\theta},\boldsymbol{\psi}) = \delta_{\rm D}(\mathbf{d} - \mathcal{S}(\boldsymbol{\theta},\boldsymbol{\psi}))$, where the symbol $\mathcal{P}$ denotes a probability distribution function (pdf) and $\delta_{\rm D}$ a Dirac delta distribution. We refer to such realisations as mock observations. As usual in ABC approaches, the (often high-dimensional) raw pre-

Symbol	Meaning	Interpretation
$\boldsymbol{\theta} \in \mathbb{R}^S$	Target parameters	Parametrisation of the primordial matter power spectrum
$\boldsymbol{\psi} \in \mathbb{R}^T$	Nuisance parameters	Random numbers involved in the initial phase realisation, instrumental noise, etc.
$\mathbf{d} \in \mathbb{R}^D$	Raw data	Galaxy number counts in a three-dimensional map of the survey volume
$\boldsymbol{\Phi} \in \mathbb{R}^P$	Summary statistics of the data (observed or simulated)	Summaries of the galaxy number count field, such as its estimated power spectrum
$\boldsymbol{\Phi}_\mathrm{O} \in \mathbb{R}^P$	Summary statistics of the observations	Summaries of the observed galaxy number count field
$\boldsymbol{\Phi}_{\boldsymbol{\theta}} \in \mathbb{R}^P$	Summary statistics of simulated data	Summaries of a galaxy number count field, simulated with primordial matter power spectrum given by $\boldsymbol{\theta}$
$\mathbf{s} \in \mathbb{R}^P$	Virtual signal	True summaries of the galaxy number count field, if they were not degraded by nuisances

TABLE I. The statistical variables appearing in section II A and their interpretation in the context of galaxy survey data analysis.

diction $\mathbf{d}$ can be compressed to a set of summary statistics $\boldsymbol{\Phi} \in \mathbb{R}^P$. We assume that this compression is a deterministic function $\mathcal{C}$ of $\mathbf{d}$, i.e. $\mathcal{P}(\boldsymbol{\Phi}|\mathbf{d}) = \delta_\mathrm{D}(\boldsymbol{\Phi} - \mathcal{C}(\mathbf{d}))$. It can be included in the model, so that the black-box is $\mathcal{B} \equiv \mathcal{C} \circ \mathcal{S}$ and

$$\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \delta_\mathrm{D}(\boldsymbol{\Phi} - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\psi})). \tag{1}$$

A graphical representation of the Bayesian hierarchical data model is presented in figure 1.

2. The exact Bayesian problem

Denoting by $\boldsymbol{\Phi}_\mathrm{O}$ the summary statistics of the observations, the inference problem considered is

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}} = L(\boldsymbol{\theta})\frac{\mathcal{P}(\boldsymbol{\theta})}{Z_{\boldsymbol{\Phi}}}, \tag{2}$$

where the likelihood is

$$L(\boldsymbol{\theta}) \equiv \mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}} \tag{3}$$

and the normalisation constant is $Z_{\boldsymbol{\Phi}} \equiv \mathcal{P}(\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}}$. By marginalising over $\boldsymbol{\psi}$ and using equation (1), we have

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int \mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta}, \boldsymbol{\psi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}} \mathcal{P}(\boldsymbol{\psi})\, \mathrm{d}\boldsymbol{\psi} \\ &= \int \delta_\mathrm{D}(\boldsymbol{\Phi}_\mathrm{O} - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\psi}))\mathcal{P}(\boldsymbol{\psi})\, \mathrm{d}\boldsymbol{\psi}. \end{aligned} \tag{4}$$

From equation (4), it is clear that the likelihood involves an intractable integral, the computation of which would require exactly hitting the observed summaries $\boldsymbol{\Phi}_\mathrm{O}$ with the black-box. We are therefore not able to explicitly formulate the true likelihood distribution for the considered problem.

3. The effective likelihood

To overcome this difficulty, in this section we derive an effective likelihood that allows us to perform inference by requiring black-box model evaluations only. For every $\boldsymbol{\theta}$, we can generate an ensemble of N mock data realisations $\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)} = \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\psi}^{(i)})$ with $i \in \{1, ..., N\}$, by drawing independent and identically-distributed realisations of nuisance parameters $\boldsymbol{\psi}$ from the probability distribution $\mathcal{P}(\boldsymbol{\psi})$ and evaluating the black-box. The approach then consists in explicitly conditioning all probabilities on $\{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\}$. Using Bayes' theorem, we have:

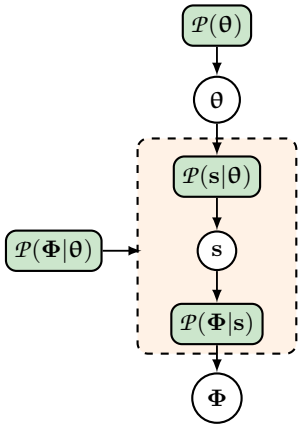

FIG. 2. Hierarchical representation of the approximate model used for inference of the black-box model of figure 1: a virtual signal $\mathbf{s}$ has been introduced as a latent variable. The orange dashed rectangle represents the assumption made about the data-generating process, it gives the effective likelihood $\widehat{L}^N(\boldsymbol{\theta})$ when conditioning on a set of simulations $\{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\}$ and using $\boldsymbol{\Phi} = \boldsymbol{\Phi}_\mathrm{O}$.

$$\begin{aligned} \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\}) &= \frac{\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\}|\boldsymbol{\theta})\, \mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\})} \\ &= \frac{\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\}|\boldsymbol{\theta})}{\mathcal{P}(\{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\})} \frac{\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{\Phi})}. \end{aligned} \tag{5}$$

Using for $\mathbf{\Phi}$ the observed data $\mathbf{\Phi}_\mathrm{O}$, we thus have the new inference problem

$$\begin{aligned}\mathcal{P}(\boldsymbol{\theta}|\mathbf{\Phi})_{|\mathbf{\Phi}=\mathbf{\Phi}_\mathrm{O}} &\approx \mathcal{P}(\boldsymbol{\theta}|\mathbf{\Phi},\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\})_{|\mathbf{\Phi}=\mathbf{\Phi}_\mathrm{O}}\\ &= \widehat{L}^N(\boldsymbol{\theta})\frac{\mathcal{P}(\boldsymbol{\theta})}{Z_{\mathbf{\Phi}}}\end{aligned} \tag{6}$$

where we have defined the first factor on the right-hand side of equation (5) evaluated at $\mathbf{\Phi}_\mathrm{O}$ to be the effective likelihood (a computable approximation of the true likelihood):

$$\widehat{L}^N(\boldsymbol{\theta}) \equiv \frac{\mathcal{P}(\mathbf{\Phi},\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\}|\boldsymbol{\theta})_{|\mathbf{\Phi}=\mathbf{\Phi}_\mathrm{O}}}{\mathcal{P}(\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\})}. \tag{7}$$

To arrive at a more explicit expression for the effective likelihood, we assume that observed data $\mathbf{\Phi}_\mathrm{O}$ and mock realisations $\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}$ are drawn from a common (but unknown) virtual signal $\mathbf{s} \in \mathbb{R}^P$. The hierarchical representation of this effective data model is presented in figure 2. We assume that $\mathbf{s}$ carries the deterministic information on the target parameters $\boldsymbol{\theta}$, which is common to the data and to the mock observations; therefore, $\mathbf{\Phi}_\mathrm{O}$ and $\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}$ only differ by stochastic uncertainties described by the nuisance parameters, which carry no information of interest. Intuitively, $\mathbf{s}$ represents the "true" version of the summaries $\mathbf{\Phi}$, which is degraded by nuisances $\boldsymbol{\psi}$. Under this assumption, we introduce the probability distribution $\mathcal{P}(\mathbf{\Phi}|\mathbf{s})$, from which data realisations are drawn independently once the virtual signal $\mathbf{s}$ is given. Since we do not want to infer $\mathbf{s}$ explicitly, marginalisation yields:

$$\begin{aligned}\mathcal{P}(\mathbf{\Phi},\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\}|\boldsymbol{\theta}) &= \int \mathcal{P}(\mathbf{\Phi},\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\},\mathbf{s}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{s}\\ &= \int \mathcal{P}(\mathbf{\Phi}|\mathbf{s})\mathcal{P}(\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\}|\mathbf{s})\mathcal{P}(\mathbf{s}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{s}.\end{aligned} \tag{8}$$

In some models, it may be possible to derive $\mathbf{s}$ from the model parameters $\boldsymbol{\theta}$. However, this is generally not true for numerical simulators, where the expected summaries have to be estimated through averaging mock realisations. In this work, in absence of prior information on $\mathbf{s}$, we set $\mathcal{P}(\mathbf{s}|\boldsymbol{\theta})$ to a constant. The joint distribution of mock observations for a given virtual signal factorises, so that one can write

$$\mathcal{P}(\mathbf{\Phi},\{\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}\}|\boldsymbol{\theta}) \propto \int \mathcal{P}(\mathbf{\Phi}|\mathbf{s})\left[\prod_{n=1}^{N}\mathcal{P}(\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}|\mathbf{s})\right]\mathrm{d}\mathbf{s}. \tag{9}$$

In order to marginalise over $\mathbf{s}$, we need to postulate a parametric form for the pdf $\mathcal{P}(\mathbf{\Phi}|\mathbf{s})$. In this work, we assume a Gaussian distribution centred on the virtual signal $\mathbf{s}$ with a covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ quantifying the stochastic uncertainties inherent to the observations,[2]

$$-2\log\mathcal{P}(\mathbf{\Phi}|\mathbf{s}) = \log|2\pi\mathbf{\Sigma}_{\boldsymbol{\theta}}| + (\mathbf{\Phi}-\mathbf{s})^\intercal\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{\Phi}-\mathbf{s}). \tag{10}$$

[2] The investigation of different choices for $\mathcal{P}(\mathbf{\Phi}|\mathbf{s})$ is left to future investigations.

The virtual signal can therefore be interpreted as the expectation E of $\mathbf{\Phi}_{\boldsymbol{\theta}}$ once $\boldsymbol{\theta}$ is specified: $\mathbf{s} = \mathrm{E}[\mathbf{\Phi}_{\boldsymbol{\theta}}]$. We also have $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \mathrm{E}[(\mathbf{\Phi}_{\boldsymbol{\theta}}-\mathbf{s})(\mathbf{\Phi}_{\boldsymbol{\theta}}-\mathbf{s})^\intercal]$.

Under these assumptions, marginalisation over $\mathbf{s}$ gives the effective likelihood as $\widehat{L}^N(\boldsymbol{\theta}) = \exp\left[\hat{\ell}^N(\boldsymbol{\theta})\right]$ (the details of the computation are given in appendix A), with:

$$-2\hat{\ell}^N(\boldsymbol{\theta}) = \log\left|2\pi\hat{\mathbf{\Sigma}}'_{\boldsymbol{\theta}}\right| + (\mathbf{\Phi}_\mathrm{O}-\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}})^\intercal\hat{\mathbf{\Sigma}}'^{-1}_{\boldsymbol{\theta}}(\mathbf{\Phi}_\mathrm{O}-\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}}), \tag{11}$$

where

$$\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}} \equiv \mathrm{E}^N[\mathbf{\Phi}_{\boldsymbol{\theta}}] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)} \tag{12}$$

is the ensemble mean of mock observations and E^N stands for the empirical average over the set. The covariance matrix of $\widehat{L}^N(\boldsymbol{\theta})$ and its inverse are defined by

$$\hat{\mathbf{\Sigma}}'_{\boldsymbol{\theta}} \equiv \frac{N+1}{N}\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}, \quad \hat{\mathbf{\Sigma}}'^{-1}_{\boldsymbol{\theta}} \equiv \left(\frac{N+1}{N}\right)^{-1}\hat{\mathbf{\Sigma}}^{-1}_{\boldsymbol{\theta}}, \tag{13}$$

where given a sufficiently large number of mock observations, a computable approximation of $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is estimated in the following way:

$$\begin{aligned}\mathbf{\Sigma}_{\boldsymbol{\theta}} \approx \hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}} &\equiv \mathrm{E}^N\left[(\mathbf{\Phi}_{\boldsymbol{\theta}}-\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}})(\mathbf{\Phi}_{\boldsymbol{\theta}}-\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}})^\intercal\right]\\ &= \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}-\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}})(\mathbf{\Phi}_{\boldsymbol{\theta}}^{(i)}-\hat{\mathbf{\Phi}}_{\boldsymbol{\theta}})^\intercal.\end{aligned} \tag{14}$$

The effective likelihood also requires an estimator $\hat{\mathbf{\Sigma}}^{-1}_{\boldsymbol{\theta}}$ of the inverse covariance matrix $\mathbf{\Sigma}^{-1}_{\boldsymbol{\theta}}$ (see equation (13)). As argued by Sellentin & Heavens (2016) and Jeffrey & Abdalla (2018), the proper Bayesian treatment would in fact consist in replacing the Gaussian likelihood by an alternative, corrected distribution. Nevertheless, keeping a Gaussian effective likelihood is an essential requirement of the present ABC technique; we will therefore be content with the Hartlap, Simon & Schneider (2007) correction, which consists of replacing the true inverse covariance matrix by a scaled inverse sample covariance matrix:

$$\mathbf{\Sigma}^{-1}_{\boldsymbol{\theta}} \approx \hat{\mathbf{\Sigma}}^{-1}_{\boldsymbol{\theta}} \equiv \alpha\left(\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}\right)^{-1}, \quad \alpha \equiv \frac{N-P-2}{N-1}, \tag{15}$$

where P is the number of summary statistics. This correction debiases the expectation value $\left\langle\left(\hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}\right)^{-1}\right\rangle = \alpha^{-1}\mathbf{\Sigma}^{-1}_{\boldsymbol{\theta}}$ of the estimator, under the assumption that $\mathbf{\Sigma}^{-1}_{\boldsymbol{\theta}}$ is inverse-Wishart distributed.

The limiting approximation of $\widehat{L}^N(\boldsymbol{\theta})$ when $N \to \infty$ is $\widetilde{L}(\boldsymbol{\theta}) = \mathcal{P}(\mathbf{\Phi}|\mathbf{s})_{|\mathbf{\Phi}=\mathbf{\Phi}_\mathrm{O}} = \exp\left[\tilde{\ell}(\boldsymbol{\theta})\right]$, with (as intended)

$$-2\tilde{\ell}(\boldsymbol{\theta}) \equiv \log|2\pi\mathbf{\Sigma}_{\boldsymbol{\theta}}| + (\mathbf{\Phi}_\mathrm{O}-\mathbf{s})^\intercal\mathbf{\Sigma}^{-1}_{\boldsymbol{\theta}}(\mathbf{\Phi}_\mathrm{O}-\mathbf{s}). \tag{16}$$

There are a number of interesting similarities and differences between the effective likelihood (equation (11))

and the virtual signal pdf (equation (16)). First, both are Gaussian distributions with respect to the observed data $\boldsymbol{\Phi}_\mathrm{O}$. Second, the mean of the effective likelihood is given by the empirical average of simulated summaries, $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}} = \mathrm{E}^N[\boldsymbol{\Phi}_{\boldsymbol{\theta}}]$. This is an unbiased estimator of the virtual signal $\mathbf{s} = \mathrm{E}[\boldsymbol{\Phi}_{\boldsymbol{\theta}}]$ for a sufficiently large number of simulations. Importantly, the replacement of the expectation E by an empirical average E^N was not an assumption (contrary to the synthetic likelihood, Wood, 2010; Price *et al.*, 2018), but naturally appeared in the derivation (see appendix A). Finally, the covariance of the effective likelihood is $\frac{N+1}{N}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, a multiple of the estimated covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. The numerical prefactor, similar to Bessel's correction for the empirical sample variance, can be understood as follows. For a small number of simulations, the latent space associated to the virtual signal increases the observed scatter. For example, for $N = 1$, the covariance to be used in equation (11) is $2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, which reflects the fact that observed and simulated data $\boldsymbol{\Phi}_\mathrm{O}$ and $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ can be drawn from opposite ends of the scatter around $\mathbf{s}$ of $\mathcal{P}(\boldsymbol{\Phi}|\mathbf{s})$. However, when $N \to \infty$, $\frac{N+1}{N}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \longrightarrow \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, the intrinsic covariance of $\mathcal{P}(\boldsymbol{\Phi}|\mathbf{s})$. This result is reasonable, since for large sample sizes the ensemble mean $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}}$ converges to $\mathbf{s}$, the mean of $\mathcal{P}(\boldsymbol{\Phi}|\mathbf{s})$; in the same limit, we expect the covariance of the effective likelihood to approach $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, the covariance of $\mathcal{P}(\boldsymbol{\Phi}|\mathbf{s})$.

Note that the scaling factor $\frac{N+1}{N}$ directly arises from the presence of the virtual signal, as shown in appendix A. Alternatively, we could have directly assumed a parametric form for the pdf $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$ without introducing the latent variable $\mathbf{s}$. For instance, the synthetic likelihood directly amounts to postulating equation (11), but without the scaling factor $\frac{N+1}{N}$ for the estimated covariance matrix. This means that the end result for $\hat{\ell}^N(\boldsymbol{\theta})$ has little sensitivity to our treatment and assumption for $\mathcal{P}(\mathbf{s}|\boldsymbol{\theta})$. Furthermore, since $\frac{N+1}{N} \longrightarrow 1$ when $N \to \infty$, our result with the virtual signal and the synthetic likelihood are equivalent, provided that the number of simulations is large enough.

Determining $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ requires N model evaluations per target parameters $\boldsymbol{\theta}$. In the next section, we show how the evaluation of this effective likelihood can be made more efficient when exploiting prior information on $\boldsymbol{\theta}$.

B. Linearisation of black-box models

The numerical cost of the computation of the effective likelihood described in section II A may be prohibitively large when the full parameter space has to be explored. However, such an extensive exploration is not always required: often, one has sufficient information to be only interested in a small region of parameter space around a specific prediction. This requirement is fulfilled for inferences of the primordial power spectrum from galaxy surveys. The target parameters $\boldsymbol{\theta}$ typically consist of power spectrum amplitudes in about a hundred different bands of wavevectors; thus the corresponding parameter space is very large. However, the CMB already provides exquisite measurements of the primordial cosmological power spectrum (e.g. Planck Collaboration, 2018a). Any large deviations from these previous measurements would most likely be rejected on methodological grounds (sample size, uncontrolled systematics, etc.) and not be attributed to new physics. It therefore seems reasonable to focus the inference of the primordial power spectrum from galaxy surveys within a narrow region in parameter space around a previous estimate obtained from CMB results.

Following this reasoning, we focus on searching for solutions corresponding to small deviations $\Delta\boldsymbol{\theta}$ around an expansion point $\boldsymbol{\theta}_0$. Consequently, the target parameters are given by $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}$. Assuming that $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}}$ is differentiable with respect to $\boldsymbol{\theta}$, we perform a first-order Taylor expansion in $\Delta\boldsymbol{\theta}$ around $\boldsymbol{\theta}_0$:

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}} \approx \mathbf{f}_0 + \nabla\mathbf{f}_0 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \equiv \mathbf{f}(\boldsymbol{\theta}), \tag{17}$$

where the defined function $\mathbf{f}$ is a linearised version of the averaged black-box. The first term corresponds to the mean mock observations at the expansion point $\boldsymbol{\theta}_0$, i.e. $\mathbf{f}_0 \equiv \hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}_0}$, while the second term involves $\nabla\mathbf{f}_0$, a $P \times S$ matrix corresponding to the gradient of mean mock observations at $\boldsymbol{\theta}_0$, whose components are $(\nabla\mathbf{f}_0)_{ps} \equiv \frac{\partial\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}_0 p}}{\partial\boldsymbol{\theta}_s}$. In this work, we estimate the gradient $\nabla\mathbf{f}_0$ via finite differencing: given a small step size h, each column is approximated by

$$(\nabla\mathbf{f}_0)^\intercal_s \approx \frac{\mathbf{f}(\boldsymbol{\theta}_s) - \mathbf{f}_0}{h}, \quad \text{with} \quad \boldsymbol{\theta}_s \equiv \boldsymbol{\theta}_0 + h\,(\delta_\mathrm{K}^{ss'})_{s'\in[\![1,S]\!]}, \tag{18}$$

where δ_K is the Kronecker delta. To avoid obtaining a noisy gradient, nuisance parameters $\boldsymbol{\psi}$ are kept at the same values in the simulations used to compute the $\mathbf{f}(\boldsymbol{\theta}_s)$ and $\mathbf{f}_0$. Further, we neglect the dependence of the covariance matrix on $\boldsymbol{\theta}$ and we estimate it at the expansion point according to equation (14), i.e. $\hat{\boldsymbol{\Sigma}}'_{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\Sigma}}'_{\boldsymbol{\theta}_0} \equiv \mathbf{C}_0$. Importantly, fully characterising $\mathbf{f}$ under these assumptions only requires evaluations of the simulator, thus ensuring that the data model remains a black-box.

Using the linearised data model $\mathbf{f}$ as a proxy for the data model described in section II A 3 simplifies the expression of the effective likelihood (equation (11)) to

$$-2\hat{\ell}^N(\boldsymbol{\theta}) \approx \log|2\pi\mathbf{C}_0| + [\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}(\boldsymbol{\theta})]^\intercal\,\mathbf{C}_0^{-1}\,[\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}(\boldsymbol{\theta})]\,. \tag{19}$$

Evaluating $\mathbf{f}_0$ and $\mathbf{C}_0$ requires N_0 model evaluations at the expansion point $\boldsymbol{\theta}_0$. The computation of the gradient further requires $N_s \times S$ model evaluations. Therefore, the linearised data model is fully characterised by a fixed total of $N_0 + N_s \times S$ model evaluations. N_0 and N_s are a user choice, but should be of the order of the dimensionality of the data space P. Particularly at the expansion point $\boldsymbol{\theta}_0$, a minimum would be $N_0 \geq P + 3$ but it can be

worth investing more simulations, in order for the estimated covariance matrix $\mathbf{C}_0$ and its inverse to be precise. Since all nuisance parameters are kept fixed in the computation of the $\mathbf{f}(\boldsymbol{\theta}_s)$, N_s can in principle be smaller than P (and as small as 1); however $N_s \gtrsim P$ will yield a safer evaluation of $\nabla\mathbf{f}_0$.

It is important to note that all required model evaluations are done once and for all. Once the linearised data model is known, it is not necessary to perform additional black-box evaluations in order to perform inference from new data. This feature makes the present approach similar to supervised machine learning algorithms, which are only trained once before being applied to multiple data sets. Furthermore, all data model evaluations can be done in parallel, or even on different machines, making the approach very suitable for grid computing.

C. The power spectrum prior distribution

In this paper, we aim at inferring the primordial matter power spectrum $P(k)$, which is a continuous function of wavenumber k. We parametrise it by its amplitudes at a sufficient number S of support wavenumbers k_s. As we are particularly interested in BAOs, we fix a "wiggle-less" power spectrum $P_0(k)$ and we work with the "wiggle function" $\theta(k) \equiv P(k)/P_0(k)$ as target function. Formally, with $\mathbf{P}_0 \in \mathbb{R}^S$ the vector of components $(\mathbf{P}_0)_s \equiv P_0(k_s)$, the inference variable $\boldsymbol{\theta}$ is defined as the S-dimensional vector of components $P(k_s)/(\mathbf{P}_0)_s$.

In order to set up the Bayesian problem, we have to formulate the prior probability distribution of $\boldsymbol{\theta}$. In this work, we include into our prior the following assumptions:

1. the power spectrum is Gaussian-distributed,
2. it is strongly constrained to live close to P_0,
3. it is a smooth function of wavenumber,
4. and the power spectrum P_0 is subject to cosmic variance.

It follows from assumptions 1 and 2 that $\mathcal{P}(\boldsymbol{\theta})$ shall be a Gaussian distribution with mean $\boldsymbol{\theta}_0 \equiv \mathbf{1}_{\mathbb{R}^S}$, which is also to be used as the expansion point.

We now discuss assumptions 3 and 4, in order to build the covariance matrix $\mathbf{S}$ of $\mathcal{P}(\boldsymbol{\theta})$. Let us define the matrix $\mathbf{K}$ as a radial basis function, i.e. by its coefficients

$$(\mathbf{K})_{ss'} \equiv \exp\left[-\frac{1}{2}\left(\frac{k_s - k_{s'}}{k_{\mathrm{corr}}}\right)^2\right]. \tag{20}$$

The hyperparameter k_{corr} determines the length scale on which power spectrum amplitudes of different wavenumber correlate with each other. A large value of k_{corr} corresponds to a strong correlation between $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{s'}$, even if their corresponding wavenumbers are far from each other. On the other hand, for $k_{\mathrm{corr}} \ll k_s, k_{s'}$, $\mathbf{K}$ becomes the identity matrix and $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{s'}$ do not correlate with each other at all. The previous discussion implies that the *a priori* smoothness of the wiggle function can be changed by tuning k_{corr}. In a realistic scenario, the prior covariance is not scale-independent as is $\mathbf{K}$ ($(\mathbf{K})_{ss} = 1$ for all s). We rather want the standard deviations $(\mathbf{S})_{ss}^{1/2}$ to account for the cosmic variance affecting the power spectrum P_0 (and hence the mean $\boldsymbol{\theta}_0$). In terms of power spectrum amplitudes, cosmic variance at a scale k is given by $P(k)^2/N_k$, where $N_k \propto k^3$ is the number of modes of wavenumber k in the considered cosmological volume. Thus, in order to account for cosmic variance in terms of $\boldsymbol{\theta}$, the coefficients $\mathbf{K}$ shall be multiplied by the coefficients of $\mathbf{u}\mathbf{u}^\intercal$, where

$$(\mathbf{u})_s \equiv 1 + \sigma_s = 1 + \frac{\alpha_{\mathrm{cv}}}{k_s^{3/2}} \tag{21}$$

and α_{cv} is a hyperparameter characterising the "strength" of cosmic variance given the considered volume. Finally, the amplitude of the covariance matrix $\mathbf{S}$ can be captured by an overall scaling θ_{norm}^2. The final expression for the prior covariance matrix is therefore

$$\mathbf{S} \equiv \theta_{\mathrm{norm}}^2\, \mathbf{u}\mathbf{u}^\intercal \circ \mathbf{K}, \tag{22}$$

where $\circ$ is the Hadamard product. The standard deviations on the diagonal are $(\mathbf{S})_{ss}^{1/2} = \theta_{\mathrm{norm}}(1 + \sigma_s)$, as intended.

The resulting prior on $\boldsymbol{\theta}$ is characterised by a set of three hyperparameters $\{k_{\mathrm{corr}}, \alpha_{\mathrm{cv}}, \theta_{\mathrm{norm}}\}$, and given as

$$-2\log\mathcal{P}(\boldsymbol{\theta}) \equiv \log|2\pi\mathbf{S}| + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \mathbf{S}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{23}$$

As demonstrated in section IV, a prior of this form results in a smooth posterior mean for the primordial matter power spectrum, while incorporating reasonable uncertainties around the expansion point.

D. The power spectrum effective posterior distribution

Using the effective likelihood with the linearised black-box data model, given in equation (19), and the prior given in equation (23), we arrive at the final expression for the effective posterior distribution (see equation (6)). It is a Gaussian distribution,

$$-2\log\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}} = \log|2\pi\boldsymbol{\Gamma}| + (\boldsymbol{\theta} - \boldsymbol{\gamma})^\intercal \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\gamma}), \tag{24}$$

with mean

$$\boxed{\boldsymbol{\gamma} \equiv \boldsymbol{\theta}_0 + \boldsymbol{\Gamma}\,(\nabla\mathbf{f}_0)^\intercal\, \mathbf{C}_0^{-1}(\boldsymbol{\Phi}_{\mathrm{O}} - \mathbf{f}_0),} \tag{25}$$

and covariance matrix

$$\boxed{\boldsymbol{\Gamma} \equiv \left[(\nabla\mathbf{f}_0)^\intercal\, \mathbf{C}_0^{-1}\nabla\mathbf{f}_0 + \mathbf{S}^{-1}\right]^{-1}.} \tag{26}$$

The proof of this result (detailed in appendix B) uses the same algebra as the derivation of the Wiener filter

Symbol	Meaning	Interpretation
$\boldsymbol{\theta}_0 \in \mathbb{R}^S$	Expansion point of the simulator in parameter space and prior mean	Fiducial primordial matter power spectrum
$\mathbf{S} \in \mathbb{R}^{S\times S}$	Prior covariance matrix	Prior covariance matrix of the primordial matter power spectrum
$\boldsymbol{\Phi}_\mathrm{O} \in \mathbb{R}^P$	Summary statistics of the observations	Summaries of the observed galaxy number count field
$\mathbf{f}_0 \in \mathbb{R}^P$	Estimated average black-box at the expansion point	Estimated average of summaries of simulated galaxy fields with fiducial primordial matter power spectrum
$\mathbf{C}_0 \in \mathbb{R}^{P\times P}$	Estimated covariance matrix of the black-box at the expansion point	Estimated covariance of summaries of simulated galaxy fields with fiducial primordial matter power spectrum
$\nabla\mathbf{f}_0 \in \mathbb{R}^{P\times S}$	Estimated gradient of the average black-box at the expansion point	Estimated gradient of summaries of simulated galaxy fields around the fiducial primordial matter power spectrum
$\boldsymbol{\gamma} \in \mathbb{R}^S$	Posterior mean	Reconstructed primordial matter power spectrum, given the observed summaries $\boldsymbol{\Phi}_\mathrm{O}$
$\boldsymbol{\Gamma} \in \mathbb{R}^{S\times S}$	Posterior covariance matrix	Uncertainties on the reconstruction of the primordial matter power spectrum, given the observed summaries $\boldsymbol{\Phi}_\mathrm{O}$

TABLE II. The statistical variables appearing in SELFI and their interpretation in the context of galaxy survey data analysis.

(Wiener, 1964). Equations (25) and (26) are the main result of this work. Equation (25) provides a simple "filter equation" to infer the primordial matter power spectrum from galaxy observations via complex black-box simulations. Corresponding uncertainties are quantified by the covariance matrix given in equation (26). Table II summarises the variables appearing in equations (25) and (26) and their interpretation in the context of galaxy survey data analysis.

E. Optimisation of the prior hyperparameters

As discussed in section II C, the chosen prior distribution involves three hyperparameters $\{k_\mathrm{corr}, \alpha_\mathrm{cv}, \theta_\mathrm{norm}\}$. By definition, α_cv characterises the strength of cosmic variance in the considered cosmological volume, such that the number of modes at a given scale k is $N_k = k^3/\alpha_\mathrm{cv}^2$. α_cv can therefore simply be measured for a given simulator setup (i.e. box size and mesh). On the contrary, k_corr and θ_norm are free hyperparameters and the power spectrum reconstruction ($\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$) generally depends on their values. In this section, we propose a procedure to find optimal values for k_corr and θ_norm.

k_corr indicates the *a priori* smoothness of reconstructed wiggle functions $\theta(k)$ and θ_norm how much they can deviate from the expansion point. Together, these two parameters characterise the functional shapes of allowed target functions (much like kernels and their hyperparameters in techniques such as Gaussian process regression, see e.g. Rasmussen & Williams, 2006). In analogy with hyperparameter optimisation in machine learning, we propose to optimise k_corr and θ_norm to reproduce the shape of a fiducial wiggle function $\boldsymbol{\theta}_\mathrm{fid}$, using as likelihood the effective posterior distribution derived in section II D. More precisely, the likelihood for k_corr and θ_norm is defined by

$$-2\log\mathcal{P}(k_\mathrm{corr}, \theta_\mathrm{norm}|\boldsymbol{\theta}_\mathrm{fid}) \equiv \log|2\pi\boldsymbol{\Gamma}| + (\boldsymbol{\theta}_\mathrm{fid}-\boldsymbol{\gamma})^\intercal\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}_\mathrm{fid}-\boldsymbol{\gamma}), \tag{27}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$, defined by equations (25) and (26), are functions of k_corr and θ_norm (through $\mathbf{S}$). At this point, it is of course possible to include a hyperprior on $(k_\mathrm{corr}, \theta_\mathrm{norm})$, if desired. The maximum likelihood estimator (or maximum *a posteriori* estimator) then provides the optimal values of k_corr and θ_norm to be used to infer $\boldsymbol{\theta}$, its functional shape being assumed to be that of $\boldsymbol{\theta}_\mathrm{fid}$.

Note that evaluating the likelihood given in equation (27) is cheap once the linearised black-box $\mathbf{f}$ has been computed: no additional data model evaluation is required, only low-dimensional matrix operations. This allows the optimal values of k_corr and θ_norm to be found using standard optimisers.

F. From the power spectrum to cosmological parameters

As argued in the introduction, the primordial matter power spectrum can be seen as a largely model-independent parametrisation of the underlying theory. The goal of this section is to go from the power spectrum to parameters of specific cosmological models. This last step in the analysis can be seen as adding a layer to the Bayesian hierarchical model (see figure 3): $\boldsymbol{\omega}$ is a vector of cosmological parameters which generates the primordial power spectrum coefficients $\boldsymbol{\theta}$. This generative process is usually deterministic, i.e.

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\omega}) = \delta_\mathrm{D}\left(\boldsymbol{\theta} - \mathcal{T}(\boldsymbol{\omega})\right), \tag{28}$$

where $\mathcal{T}$ is a deterministic function of cosmological parameters, typically a Boltzmann solver or a fitting

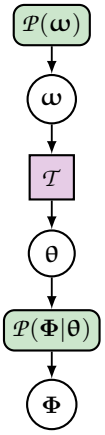

FIG. 3. Hierarchical representation of the Bayesian model for inference of cosmological parameters $\boldsymbol{\omega}$: a layer has been added above the primordial power spectrum $\boldsymbol{\theta}$. The purple square represents the deterministic process generating $\boldsymbol{\theta}$ given $\boldsymbol{\omega}$.

function. Given this assumption, we have $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\omega}) = \mathcal{P}(\boldsymbol{\Phi}|\mathcal{T}(\boldsymbol{\omega}))$, and the inference of cosmological parameters gives

$$\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}} = \widehat{L}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega})\frac{\mathcal{P}(\boldsymbol{\omega})}{Z_{\boldsymbol{\Phi}}}, \tag{29}$$

where the likelihood for cosmological parameters is defined as

$$\begin{aligned}\widehat{L}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega}) &\equiv \mathcal{P}(\boldsymbol{\Phi}|\mathcal{T}(\boldsymbol{\omega}))_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}} \\ &= \widehat{L}^N(\mathcal{T}(\boldsymbol{\omega})).\end{aligned} \tag{30}$$

Noting $\widehat{L}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega}) \equiv \exp\left[\hat{\ell}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega})\right]$ and using the linearised data model, we have (see equation (19))

$$\begin{aligned}-2\hat{\ell}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega}) \approx &\log|2\pi\mathbf{C}_0| \\ &+ \left[\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}(\mathcal{T}(\boldsymbol{\omega}))\right]^\intercal \mathbf{C}_0^{-1}\left[\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}(\mathcal{T}(\boldsymbol{\omega}))\right].\end{aligned} \tag{31}$$

Going from the primordial power spectrum to cosmological parameters is conceptually easy, but often hard in practice, when one only has samples of the power spectrum likelihood (or posterior). In this case, the likelihood to be used for cosmological parameter inference is naively represented by a sum of Dirac delta distributions. As it usually lives in a parameter space containing hundreds to thousands of dimensions, the number of samples is always a limiting factor. For this reason, techniques that effectively broaden the obtained samples, such as Blackwell-Rao estimators (Wandelt, Larson & Lakshminarayanan, 2004) or kernel density estimates have been introduced. We note that our ABC technique does not suffer from this complication: the likelihood for cosmological parameters (equation (31)) is a Gaussian centered on the data $\boldsymbol{\Phi}_\mathrm{O}$ with a fixed covariance matrix $\mathbf{C}_0$. Furthermore, getting predictions for $\boldsymbol{\Phi}$ amounts to evaluating $\mathbf{f}(\mathcal{T}(\boldsymbol{\omega}))$, which does not require additional black-box evaluations once $\mathbf{f}$ is known. The Bayesian problem of inferring cosmological parameters $\boldsymbol{\omega}$ (equation (29)) can therefore easily be solved by standard Markov Chain Monte Carlo techniques.

III. DATA MODEL

In this section, we describe the generation of mock observations used to test the performance of our method. We emphasise that the statistical method presented in section II is applicable to any black-box model, which can feature arbitrarily complex processes. Therefore, the details of the simulator presented in this section are of no relevance to the performance of the statistical method. It is only used in section IV as a showcase, to highlight the performance of our method to handle a complex simulator.

A. Primordial power spectrum parametrisation

Throughout this paper, we work with a cubic equidistant grid with comoving side length of 1 Gpc/h and 256^3 voxels, spanning scales between $k_{s,\mathrm{min}} = 6.28 \times 10^{-3}$ h/Mpc and $k_{s,\mathrm{max}} = 1.4$ h/Mpc. We use $S = 100$ support wavenumbers k_s. The first eight are fixed to the values required by the Fourier grid given our setup. The remaining support wavenumbers are logarithmically spaced up to k_max. Any vector $\boldsymbol{\theta}$ in parameter space is defined at the scales of these support wavenumbers.

Between two consecutive support wavenumbers, we interpolate power spectra $P(k)$ using a one-dimensional spline fit, using $n = 5$ as the degree of the smoothing spline. We checked that this setup yields vanishing differences in the representation of cosmological power spectra, at all wavenumbers of the Fourier grid used in this work.

B. Galaxy surveys

The data model used in this work is a non-linear process meant to approximate the large variety of physical and observational phenomena at play in galaxy surveys. To do so, it uses the SIMBELMYNË cosmological code, an end-to-end generative process for galaxy survey data given a specified primordial power spectrum $P(k)$. The flat ΛCDM model is assumed, and fiducial cosmological parameters used are the Planck 2015 values (Planck Collaboration, 2016, table 4, last column), given in table III. These are used whenever the distance-redshift relation is

needed, as well as for the gravitational evolution.[3]

h	Ω_b	Ω_m	n_S	σ_8
0.6674	0.0486	0.3089	0.9667	0.8159

TABLE III. The cosmological parameters used in this work.

SIMBELMYNË first generates a realisation of the initial density contrast via the convolution approach (see e.g. Peacock & Heavens, 1985). Specifically, the code generates a white noise field $\mathbf{w}$, such that in each cell x the value w_x is drawn from the zero-mean unit-variance Gaussian distribution. The white noise field is multiplied by the square root of the desired cosmological power spectrum in Fourier space to give

$$\delta_k^\mathrm{i} \equiv \sqrt{P(k)}\, w_k. \tag{32}$$

Transformation back to configuration space yields a initial density contrast field $\boldsymbol{\delta}^\mathrm{i}$. One realisation of such an initial density field is shown in the left panel of figure 4.

Generated initial density fields then act as inputs to numerical structure formation simulations. The initial grid of 256^3 voxels is populated by 512^3 dark matter particles placed on a regular lattice. These particles are evolved to the redshift of $z = 19$ via second order Lagrangian Perturbation theory (2LPT) (see e.g. Moutarde *et al.*, 1991; Bouchet *et al.*, 1995; Bouchet, 1996), then with an efficient implementation of COLA (COmoving Lagrangian Acceleration, Tassev, Zaldarriaga & Eisenstein, 2013) from $z = 19$ to $z = 0$. A particle-mesh grid of 1024^3 voxels and 20 timesteps linearly-spaced in the scale factor are used for the evolution with COLA. We checked that, at the scales of interest, this setup yields negligible difference in the representation of final density fields with respect to the prediction of the fully non-linear code GADGET-2 (Springel, 2005). In particular, final density fields contain the additional power expected from non-linear structure formation, at 1% precision up to $k = 0.5$ h/Mpc and 5% up to $k = 1$ h/Mpc. We place the observer at the centre of the box. The maximal distance to the observer is 866 Mpc/h, which we consider sufficiently small to neglect light-cone effects. Therefore, in this paper, we only use the final snapshot of our simulations at redshift zero, and ignore the evolution of matter within the survey volume.

Using their final peculiar velocities with respect to the observer v_r, dark matter particles are placed in redshift space according to the non-linear mapping

$$1+z_\mathrm{obs} = (1+z_\mathrm{cosmo})(1+z_\mathrm{pec}), \quad \text{with} \quad z_\mathrm{pec} \equiv -\frac{v_r}{\mathrm{c}}, \tag{33}$$

where z_cosmo is the true cosmological redshift, z_obs is the "observed" redshift and c is the speed of light. Note that we do not work in the plane-parallel approximation. The particles are then binned to a 256^3-voxels grid with the cloud-in-cell scheme (Hockney & Eastwood, 1981) to give the final density contrast field $\boldsymbol{\delta}^\mathrm{f}$. One realisation of the final redshift-space density is shown in the middle panel of figure 4.

The galaxy density $\boldsymbol{\rho}^\mathrm{g}$ is predicted using a linear bias model, used in various previous studies (e.g. Verde *et al.*, 2002; Ross *et al.*, 2015), such that in any cell x,

$$\rho_x^\mathrm{g} = \bar{N}\,(1 + b\,\delta_x^\mathrm{f}). \tag{34}$$

In this work, we use $b = 1.2$ and $\bar{N} = 0.119$ (corresponding to an observed galaxy number density $\bar{n} = 2\times10^{-3}$ $(h/\mathrm{Mpc})^3$, achievable for instance with the Euclid spectroscopic survey, see e.g. Majerotto *et al.*, 2012). For the sake of simplicity, b and $\bar{N}$ are fixed, but they could be straightforwardly treated as additional nuisance parameters and marginalised over.

The last step corresponds to a virtual observation of the galaxy field, accounting for observational effects expected in actual surveys. To do so, we use the three-dimensional survey response operator (or window) $\mathbf{W}$, consisting of the product of the radial selection function $R(r)$ and the angular survey mask and completeness function $C(\hat{\mathbf{n}})$ for any line-of-sight $\hat{\mathbf{n}}$, i.e. $\mathbf{W}(\hat{\mathbf{n}}, r) \equiv R(r)\, C(\hat{\mathbf{n}})$. This operator accounts for the fact that we are only looking at certain parts of the sky and that we have different detection probabilities of galaxies depending on their distance. We obtained a simple model for $\mathbf{W}$ on a grid matching our simulations (256^3 voxels covering a volume of $(1\ \mathrm{Gpc}/h)^3$) as follows. For the angular completeness, we mask the galactic plane by setting $C(\hat{\mathbf{n}})$ to 0 for galactic latitudes $-10° \leq b \leq 10°$ and 1 otherwise, the system of coordinates being defined such that the observer is at the origin and the plane of equation $z = 0$ is the galactic plane. For the radial selection function $R(r)$, we use a Schechter luminosity function (Schechter, 1976) with previously-published parameters for the r-band: $\alpha = -1.05$ and $M_* = -20.44$ (Blanton *et al.*, 2003), a limiting apparent magnitude of $m = 18.5$ and absolute magnitude cuts $-25 \leq M \leq -21$. This choice makes our synthetic survey complete up to a luminosity distance of $D_\mathrm{L} = 794$ Mpc/h, corresponding to $r = 684$ Mpc/h (see figure 5).

We emulate the survey via galaxy number counts N_x^g in each cell x of the box. In order to account for instrumental noise, our model is a non-uniform Gaussian process, meaning that in every cell N_x^g is drawn from a Gaussian distribution with mean μ_x^g and standard deviation σ_x,

$$N_x^\mathrm{g} \curvearrowleft \mathcal{G}(\mu_x^\mathrm{g}|\sigma_x) \tag{35}$$

The mean $\mu_x^\mathrm{g} \equiv W_x \rho_x^\mathrm{g} = W_x \bar{N}(1 + b\,\delta_x^\mathrm{f})$ characterises the expected number of galaxies. It accounts for all physical effects via δ_x^f and for the survey response operator W_x. The standard deviation is defined as $\sigma_x \equiv \sigma\sqrt{W_x \bar{N}}$,

[3] In principle, our approach would require treating cosmological parameters as nuisance parameters in the inference of the primordial power spectrum, and to marginalise over them. In this work, for simplicity, we keep them fixed to the values used to predict the expansion point.

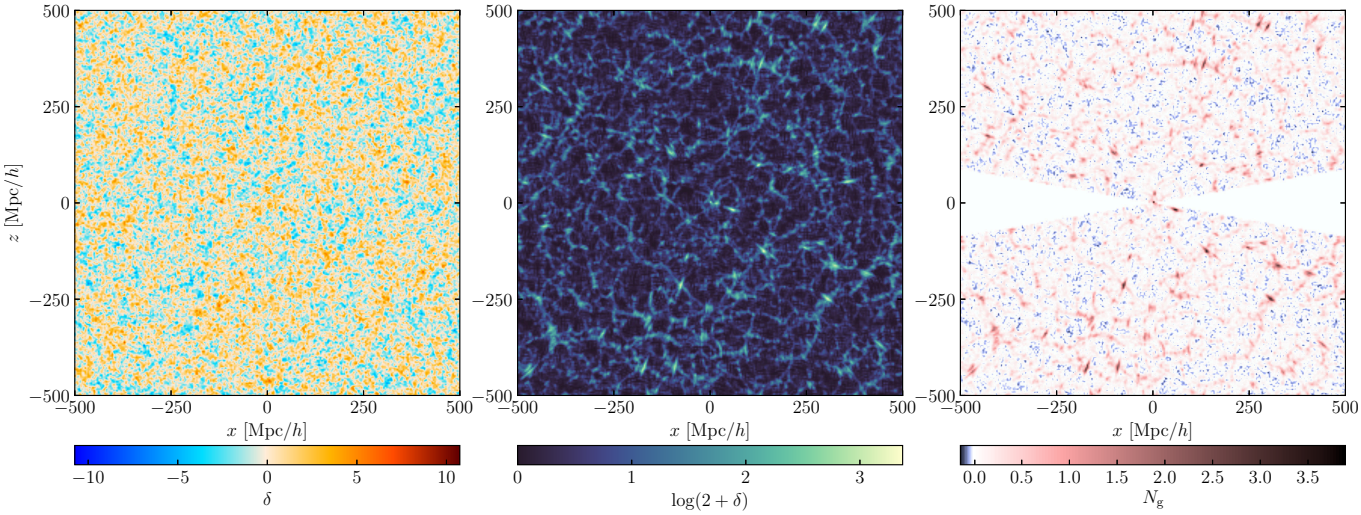

FIG. 4. Slices through one realisation of the initial density field (left) scaled to redshift zero with the linear growth factor, the corresponding evolved density field in redshift space (middle), and the field of observed galaxy number counts (right), according to the data model used in this work. The fields are defined on a grid of 256^3 cells covering a total volume of $(1\ \mathrm{Gpc}/h)^3$.

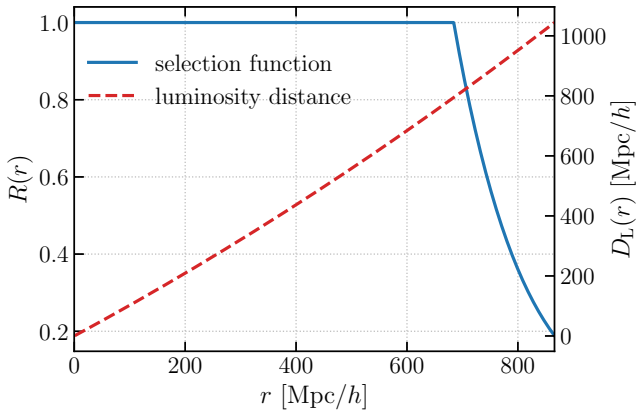

FIG. 5. Radial selection function $R(r)$ (solid blue line) used in our synthetic survey as a function of the comoving distance to the observer r. The model used is a Schechter luminosity function with parameters $\alpha = -1.05$ and $M_* = -20.44$, a limiting apparent magnitude of $m = 18.5$ and absolute magnitude cuts $-25 \leq M \leq -21$. The luminosity distance D_L (dashed red line, right vertical axis) is also shown. The survey is complete up to $D_\mathrm{L} = 794$ Mpc$/h$, corresponding to $r = 684$ Mpc$/h$.

where we set the overall noise level $\sigma \equiv 10^{-1}$. A slice through one realisation of the galaxy number count field $\mathbf{N}^\mathrm{g}$ is shown in the right panel of figure 4.

C. Summary statistic

As is standard practice in cosmological data analysis, we do not work at the level of the entire galaxy number count map $\mathbf{N}^\mathrm{g}$, but use a compression $\mathcal{C}$ of the map to well-chosen statistical summaries. In this work, we limit ourselves to an estimator of the final power spectrum of the survey. It should be remarked however, that the method can use any other statistical summary and even combinations of those.

We obtain the binned data power spectrum by taking the squared modulus of the Fourier transform of $\mathbf{N}^\mathrm{g}$, summing the contributions over all the Fourier cells within wavenumber shell k_r, and normalizing:

$$P^\mathrm{f}(k_r) \equiv C \times \sum_{|\mathbf{k}| \in k_r} \frac{|N^\mathrm{g}_k|^2}{N_{k_r} - 2}. \tag{36}$$

The overall constant factor $C = 1000^3/256^2\ (\mathrm{Mpc}/h)^3$ arises from our Fourier transform convention; N_{k_r} represents the number of modes within the shell k_r; and the factor -2 arises from the assumption that the data power spectrum is inverse-Γ distributed with shape parameter $N_{k_r}/2$ and scale parameter $C \times |N^\mathrm{g}_k|^2/2$ (see Jasche *et al.*, 2010).

As the galaxy data model relies on a full cosmological simulation, the predictions for the final power spectrum remain reasonable even at scales that have experienced substantial non-linearity. In spite of the approximations made, we trust our data model at the percent level up to $k = 0.5\ h/\mathrm{Mpc}$. For this reason, we use $P = 43\ k_r$-bins in the range $[0.02, 0.5]\ h/\mathrm{Mpc}$, ensuring that each bin contains at least 100 modes. These bins are logarithmically spaced for $k_r \geq 0.04\ h/\mathrm{Mpc}$. For convenience, we normalise the output of the black-box using the expansion point, so that $\mathbf{\Phi}$ is the P-dimensional vector of components $A \times P^\mathrm{f}(k_r)/P_0(k_r)$ with $A = 50$. The estimation of $P^\mathrm{f}(k)$ is completely deterministic once the galaxy number counts are given. Therefore, we now have a complete

model to generate artificial realisations of $\mathbf{\Phi}$ for a given primordial power spectrum $P(k)$ and specific realisations of initial phases and noise.

Due to the breakdown of models based on perturbation theory in the non-linear regime and to the difficulties in incorporating the impact of small-scale observational processes, state-of-the-art large-scale structure analyses are typically limited to $k_{\max} \lesssim 0.3\ h/\mathrm{Mpc}$ (e.g. Ross *et al.*, 2015). Pushing the analysis to $k_{\max} = 0.5\ h/\mathrm{Mpc}$ represents an increase by a factor of ~ 5 in the number of modes used (scaling as $k_{\max}^3$), which is expected to yield substantial improvements in the inference results.

D. Idealised data model

For testing purposes, we also define an idealised data model corresponding to a Gaussian random field. More specifically, using exactly the same setup as before, the black-box here simply consists of producing the initial density field $\boldsymbol{\delta}_{\mathrm{i}}$ (see section III B), scaling it to redshift zero using the linear growth factor, and measuring its normalised power spectrum $\mathbf{\Phi}$ (see section III C). The use of the two different black-boxes (Gaussian random field and realistic mock survey) within our method will quantify the effect of non-linear gravity, redshift-space distortions, and survey complications on primordial power spectrum inference, in particular the detectability of BAOs.

IV. RESULTS

This section describes the results obtained by applying the statistical method proposed in section II in conjunction with the data-generating process described in section III to an artificial galaxy survey, itself generated using the same process.

We use for $P_0(k)$ the "wiggle-less" BBKS power spectrum (Bardeen *et al.*, 1986) under Planck 2015 cosmology (see table III). Unknown ground truth cosmological parameters $\boldsymbol{\omega}_{\mathrm{gt}}$ are drawn from the (marginalised, Gaussian) Planck priors:

$$\begin{pmatrix} h \\ \Omega_{\mathrm{b}} \\ \Omega_{\mathrm{m}} \\ n_{\mathrm{S}} \\ \sigma_8 \end{pmatrix} \curvearrowleft \mathcal{G}\left[\begin{pmatrix} 0.6774 \\ 0.04860 \\ 0.3089 \\ 0.9667 \\ 0.8159 \end{pmatrix}, \mathrm{diag} \begin{pmatrix} 0.0046^2 \\ 0.00030^2 \\ 0.0062^2 \\ 0.0040^2 \\ 0.0086^2 \end{pmatrix} \right]. \quad (37)$$

The "wiggly" ground truth power spectrum $P_{\mathrm{gt}}(k)$ is generated with the Eisenstein & Hu (1998) (EH) fitting function, using these cosmological parameters. It is used to simulate observed data $\mathbf{\Phi}_{\mathrm{O}}$, with unknown nuisance parameters (phase realisation and instrumental noise). For later use, the fiducial "wiggly" power spectrum $P_{\mathrm{fid}}(k)$ is also generated with the EH prescription, using Planck cosmology. The target parameters $(\boldsymbol{\theta})_s \equiv P(k_s)/P_0(k_s)$ are the values of the wiggle function at the $S = 100$ support wavenumbers defined in section III A. We note $\boldsymbol{\theta}_{\mathrm{gt}}$ and $\boldsymbol{\theta}_{\mathrm{fid}}$ the vectors of component $P_{\mathrm{gt}}(k_s)/P_0(k_s)$ and $P_{\mathrm{fid}}(k_s)/P_0(k_s)$, respectively.

A. Diagnostics of the black-box

We created an ensemble of $N_0 = 150$ mock realisations at the expansion point $\boldsymbol{\theta}_0 = \mathbf{1}_{\mathbb{R}^S}$ using different nuisance parameters. These are used to compute the average black-box at the expansion point, $\mathbf{f}_0 \equiv \hat{\mathbf{\Phi}}_{\boldsymbol{\theta}_0}$, and the covariance matrix $\mathbf{C}_0 \equiv \hat{\mathbf{\Sigma}}'_{\boldsymbol{\theta}_0}$, using their definitions (equations (12), (13), and (14)). The results are shown in figure 6. There, the left panel shows individual realisations $\mathbf{\Phi}_{\boldsymbol{\theta}_0}$ and the observed data vector $\mathbf{\Phi}_{\mathrm{O}}$. The average black-box $\mathbf{f}_0$ is also plotted, with a credible region corresponding to two standard deviations (i.e. $2\sqrt{\mathrm{diag}(\mathbf{C}_0)}$). The full estimated covariance matrix $\mathbf{C}_0$ is shown in the right panel. As expected, the measured variance is larger on large scales due to cosmic variance, with some anti-correlations between pairs of bins. The effect of the mask, which increases power at the largest scales found in the simulation box, is also clearly visible.

Using a step size of $h = 10^{-2}$ and an ensemble of $N_s = 100$ mock realisations at each of the expansion points $\boldsymbol{\theta}_s$, we measured the gradients of the black-box $(\nabla \mathbf{f}_0)_s^{\intercal}$ along all directions of parameter space (see equation (18)). The nuisance parameters (phase realisation and noise) are kept at fixed values (the ones corresponding to the first N_s realisations generated at the expansion point) for this calculation. The results are shown in figure 7, where the left panel shows $(\nabla \mathbf{f}_0)_s^{\intercal}$ for individual values of s and the right panel shows the full rectangular matrix $\nabla \mathbf{f}_0$. Some interesting phenomena can be observed. At large scales, (see e.g. for $k_{20} = 0.0364\ h/\mathrm{Mpc}$) exciting one initial mode ony triggers an answer in the bins closest to this scale; the gradient therefore resembles a multiple of the identity function. This is the result expected from linear perturbation theory. However, at small scales, the non-linear simulator couples modes. This implies that the response is smaller in amplitude but distributed over a much larger ranges of scales (see e.g. for $k_{70} = 0.3780\ h/\mathrm{Mpc}$). In the non-linear regime, the gradient is typically negative at large scales, crosses zero slightly before the excited scale k_s, then becomes positive at smaller scales.

As discussed in section II B, the linearised black-box $\mathbf{f}(\boldsymbol{\theta})$ is fully characterised by $\mathbf{f}_0$, $\mathbf{C}_0$ and $\nabla \mathbf{f}_0$. In this work, we used a total of $N_0 + N_s \times S = 10,150$ simulations to get very precise estimates of $\mathbf{C}_0$ and $\nabla \mathbf{f}_0$, although using fewer would have been possible.

B. The prior and its optimisation

As discussed in section II C, we choose a Gaussian prior centered on the expansion point $\boldsymbol{\theta}_0$ with a covariance

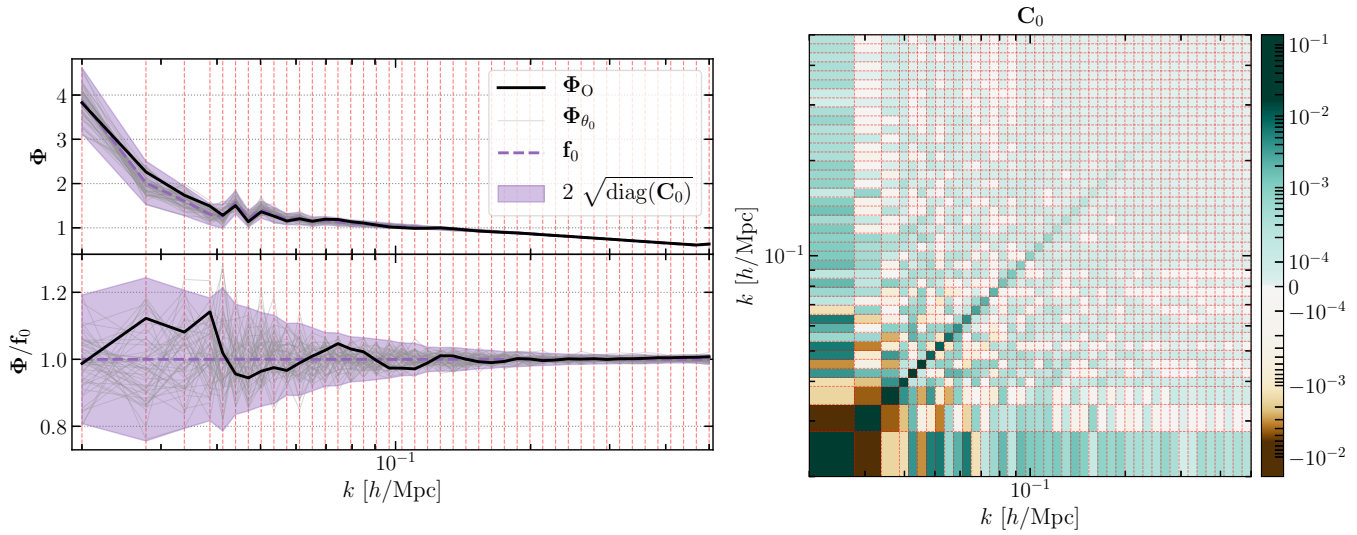

FIG. 6. Diagnostics of the black-box at the expansion point $\boldsymbol{\Phi}_0$. The left panel shows individual mock observations $\boldsymbol{\Phi}_{\theta_0}$ (grey lines), the observed data $\boldsymbol{\Phi}_\mathrm{O}$ (solid black line), and the average black-box $\mathbf{f}_0$ (dashed purple line). The shaded region corresponds to two standard deviations. The right panel shows the covariance matrix of the summaries at the expansion point, $\mathbf{C}_0$. The dashed red lines correspond to the positions of the bins at which summaries are measured in data space.

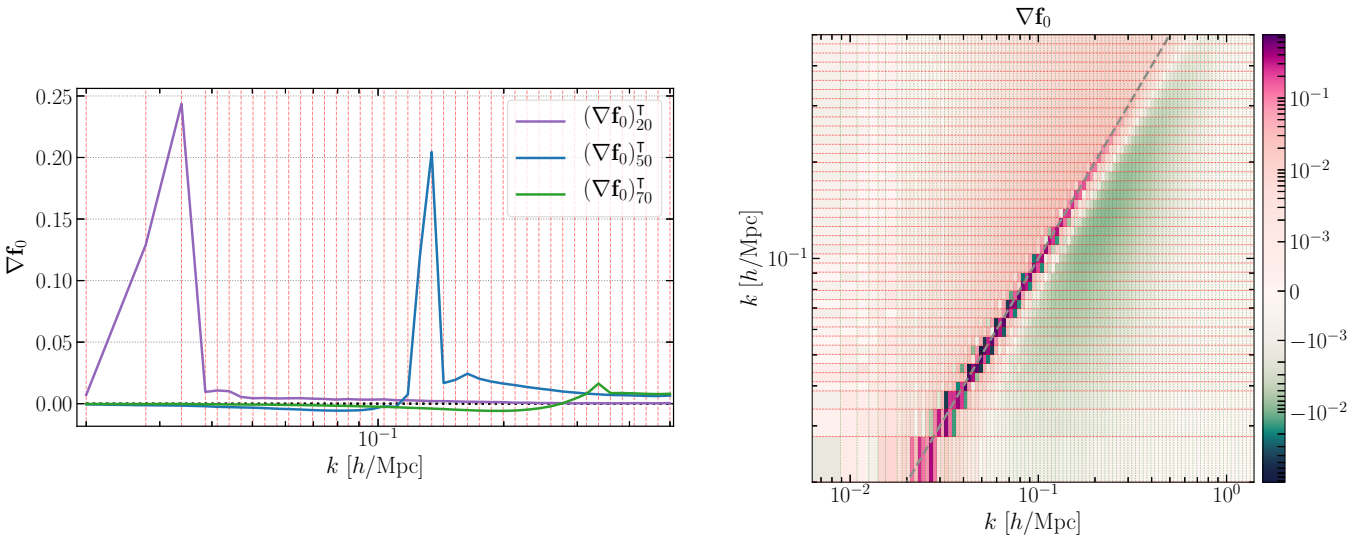

FIG. 7. Gradient of the black-box, measured via finite differencing. The right panel shows individual columns $(\nabla\mathbf{f}_0)^\intercal_s$ for $s = 20, 50$, and 70, corresponding to support wavenumbers $k_s = 0.0364, 0.1484$, and 0.3780 h/Mpc, respectively. The right panel shows the gradient matrix $\nabla\mathbf{f}_0$. The dashed grey line corresponds to the identity function; the dashed red lines correspond to the positions of the bins at which summaries are measured in data space; and the dotted green lines correspond to the support wavenumbers in parameter space.

matrix $\mathbf{S}$ given by equation (22) and characterised by three hyperparameters $\{k_\mathrm{corr}, \alpha_\mathrm{cv}, \theta_\mathrm{norm}\}$.

Following the method presented in section II E, we found optimal values for the prior hyperparameters. The strength of cosmic variance within our simulation volume shall satisfy $\alpha_\mathrm{cv} = \sqrt{k^3/N_k}$ at all scales k, where N_k is the number of modes. In our Fourier grid (described in section III A), we measured up to the Nyquist frequency $\alpha_\mathrm{cv} = 8.848 \times 10^{-4}$, value that we adopt.

Assuming that the target function $\boldsymbol{\theta}$ follows the functional shape of the fiducial wiggle function $\boldsymbol{\theta}_\mathrm{fid}$ calculated with Planck cosmology, the likelihood for k_corr and θ_norm is given by equation (27). We further assume broad, uncorrelated Gaussian hyperpriors on k_corr and θ_norm: $k_\mathrm{corr} \sim \mathcal{G}(0.020, 0.015^2)$ [h/Mpc] and $\theta_\mathrm{norm} \sim \mathcal{G}(0.2, 0.3^2)$. The posterior surface is plotted in figure

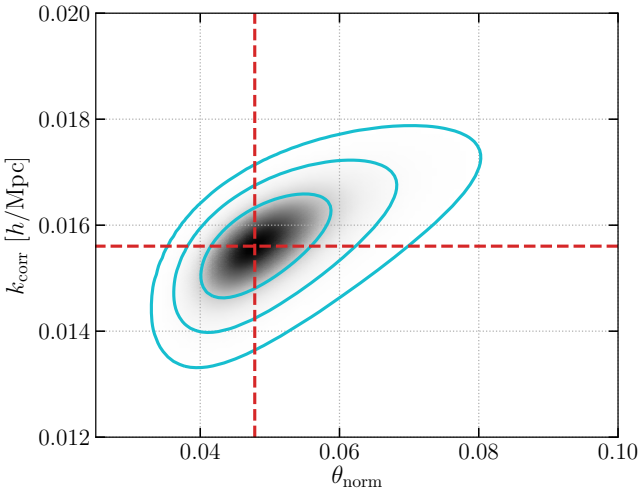

FIG. 8. Posterior surface for the prior hyperparameters $k_{\rm corr}$ and $\theta_{\rm norm}$. The 1σ, 2σ and 3σ credible contours are shown as solid blue lines. The dashed red lines mark the maximum *a posteriori* values.

8. We found the maximum *a posteriori* values using the popular optimiser L-BFGS (Byrd *et al.*, 1995). In our run, these are $k_{\rm corr} = 0.0156\ h/\mathrm{Mpc}$ and $\theta_{\rm norm} = 0.0478$. The resulting prior covariance matrix is shown in figure 9. For the idealised data model (see section III D), the same procedure is applied and the optimal parameters are found to be $k_{\rm corr} = 0.0158\ h/\mathrm{Mpc}$ and $\theta_{\rm norm} = 0.0535$.

C. The effective posterior

Using the effective likelihood, characterised by the observed data $\boldsymbol{\Phi}_{\rm O}$ and the linearised black-box described in section IV A, as well as the optimised prior discussed in section IV B, we obtained the effective posterior on $\boldsymbol{\theta}$. It is a Gaussian, with mean and covariance matrix given by the "filter equations" (25) and (26).

Figure 10 shows the inferred primordial wiggle function $\boldsymbol{\gamma}$ in comparison with the expansion point $\boldsymbol{\theta}_0$, the ground truth $\boldsymbol{\theta}_{\rm gt}$, and the fiducial function used to optimise prior hyperparameters, $\boldsymbol{\theta}_{\rm fid}$. 2σ credible regions are shown for the prior and the posterior (i.e. $2\sqrt{\mathrm{diag}(\mathbf{S})}$ and $2\sqrt{\mathrm{diag}(\boldsymbol{\Gamma})}$, respectively). The top panel corresponds to the result obtained using the idealised data model (a Gaussian random field, see section III D), and the bottom panel to the result obtained using the realistic mock survey data model. In both cases, the inference is unbiased since the ground truth always lies within the 2σ credible intervals of the reconstruction. As discussed in the introduction, this effective posterior can be seen as a largely model-independent parametrisation of the theory, containing all the available cosmological information under weak assumptions.

As can be read from the figure, the inferred vector contains the BAO wiggles, even far within the Silk damping tail (Silk, 1968). All visible oscillations are fully reconstructed in the idealised case, and in the realistic case, up to scales of $k \approx 0.3\ h/\mathrm{Mpc}$. In particular, 5 acoustic oscillations are unambiguously identified, which is competitive with the latest cosmic microwave background experiments and has been so far out of reach of galaxy surveys. Note that this result is obtained given a simulation volume of $(1\ \mathrm{Gpc}/h)^3$ and that further improvements could be obtained with a larger volume, as will be probed by upcoming surveys. The inferred wiggle function shows higher uncertainty in regions of small and large wavenumbers. This is due to cosmic variance and noise, respectively. Cosmic variance reflects the limited number of modes that we have at the largest scales in our simulation volume. This effect limits the significance of the determination of the cosmological power spectrum at these scales. For this reason, at $k \lesssim 0.05\ h/\mathrm{Mpc}$ the reconstruction is driven towards the prior mean, which is also the expansion point and the default answer in the absence of informative data. As expected, this effect is visible in the idealised as well as in the realistic case. On the other hand, noise (understood as the combined effect of the specific phase realisation of the data, non-linear gravity, redshift-space distortions and instrumental noise) acts on smaller scales ($0.2\ h/\mathrm{Mpc} \lesssim k \lesssim 0.5\ h/\mathrm{Mpc}$). Some of the primordial information is effectively destroyed at these scales – or at least is not captured by the statistical summaries $\boldsymbol{\Phi}$. The reconstruction is therefore also driven towards the prior mean, and the uncertainty is increased, because the data are less informative than in the idealised case. As expected, we recover the prior at $k \gtrsim 0.5\ h/\mathrm{Mpc}$ (in fact a little below, due to mode coupling), since the data $\boldsymbol{\Phi}$ do not contain measurements at these scales.

It is important to note that the prior $\mathcal{P}(\boldsymbol{\theta})$, used in this section to regularise the inference of the primordial power spectrum, does not appear in the inference of cosmological parameters in the next section (only the effective likelihood does). Thus, no bias is introduced in cosmological parameter inference when the amplitude of reconstructed BAO wiggles seems to undershoot the ground truth.

Since the proposed method is fully Bayesian, we do not simply obtain a point estimate, but a complete probability distribution, which provides a detailed quantification of uncertainties. Figure 11 shows the covariance matrix $\boldsymbol{\Gamma}$ of the Gaussian effective posterior; it can be compared with the prior covariance matrix $\mathbf{S}$ shown in figure 9.

D. Cosmological parameters

The last step in the analysis is to infer parameters $\boldsymbol{\omega}$ of specific cosmological models, given the observed data $\boldsymbol{\Phi}_{\rm O}$. In this section, we assume a flat ΛCDM model, characterised by 5 parameters $\{h, \Omega_{\rm b}, \Omega_{\rm m}, n_{\rm S}, \sigma_8\}$. For simplicity, our prescription to generate primordial power spectra given cosmological parameters is the EH fitting

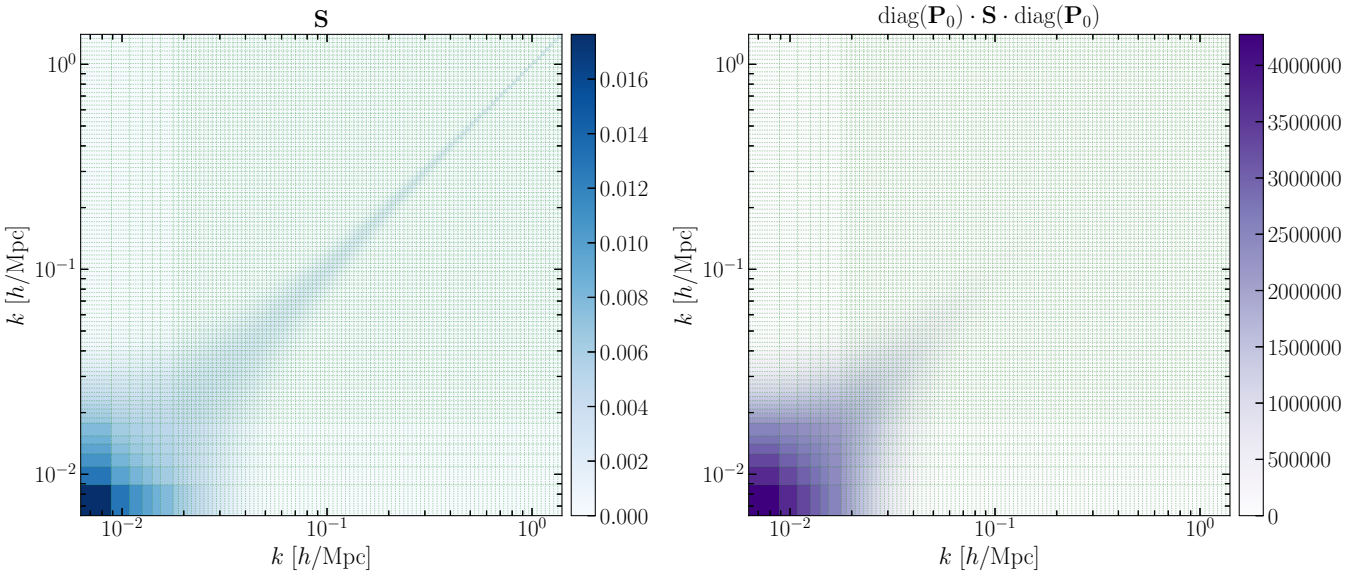

FIG. 9. Prior covariance matrix $\mathbf{S}$ for the wiggle function $\boldsymbol{\theta}$ (left panel). The corresponding prior covariance matrix for power spectrum amplitudes, $\mathrm{diag}(\mathbf{P}_0) \cdot \mathbf{S} \cdot \mathrm{diag}(\mathbf{P}_0)$, is shown in the right panel. The dotted green lines correspond to the support wavenumbers in parameter space.

function. The generative process $\mathcal{T}$ (from $\boldsymbol{\omega}$ to $\boldsymbol{\theta}$) is therefore the EH fitting function divided by P_0, sampled at the support wavenumbers k_s. It could easily be generalised to extensions of the flat ΛCDM model and to include a Boltzmann solver. The linearised data model $\mathbf{f}$ (from $\boldsymbol{\theta}$ to $\boldsymbol{\Phi}$) has already been characterised in section IV A.

The effective likelihood $\widehat{L}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega})$ for cosmological parameters is given by equation (31). Consistently with the expansion point used to linearise the black-box, we complement $\widehat{L}^N_{\boldsymbol{\omega}}(\boldsymbol{\omega})$ with a Gaussian prior $\mathcal{P}(\boldsymbol{\omega})$ centred on Planck cosmological parameters, but with broader variance: the diagonal covariance matrix given in equation (37) is multiplied by a factor of 3. We explored the effective posterior $\mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}}$ (equation (29)) via MCMC (performed using the EMCEE code, Foreman-Mackey *et al.*, 2013), ensuring sufficient convergence.

Results are shown in figure 12. Prior contours are shown in blue, and posterior contours are shown in red and purple for two different realisations of the data $\boldsymbol{\Phi}_\mathrm{O}$. The two data realisations have been generated using different ground truth cosmological parameters $\boldsymbol{\omega}_\mathrm{gt}$ (shown as dashed and dotted lines, respectively), as well as different nuisance parameters (phase realisation and noise). The plot demonstrates that cosmological parameter inference is unbiased and robust to nuisances imprinted in the data.

V. DISCUSSION AND CONCLUSION

The biggest challenge in galaxy survey analyses arises from the requirement of non-linear data modelling. In this work, we described the development of a novel simulation-based Bayesian approach, SELFI, which can be used to infer the primordial matter power spectrum and cosmological parameters from galaxy surveys. The main results are the "filter equations" (25) and (26). They can be applied to get an effective posterior for any model where the mean and covariance of the data are estimated from arbitrarily complex forward models. Essentially everything is obtained from a simulator, which can be treated as a black-box, without necessity to include any knowledge of its internal mechanisms into the statistical analysis.

We derived the "filter equations" under two assumptions: the availability of a black-box able to generate artificial data, and of strong prior constraints in parameter space, obtained from a previous experiment. We built an effective likelihood for this scenario and made its evaluation efficient by linearising the black-box around an expansion point. We devised a method to optimise the hyperparameters appearing in our power spectrum prior. Finally, we derived the cheap likelihood for parameters of specific cosmological models, to be used in our framework.

The approach presented in this paper relies on likelihood-free forward-modelling via ABC. It complements statistically exact, likelihood-based forward-modelling techniques. The principal differences are as follows.

- First, the numerical complexity of likelihood-based MCMC approaches typically requires to approximate complex data models to allow for fast execution speeds. In this work, we rather aimed at performing approximate inference, but with full-scale

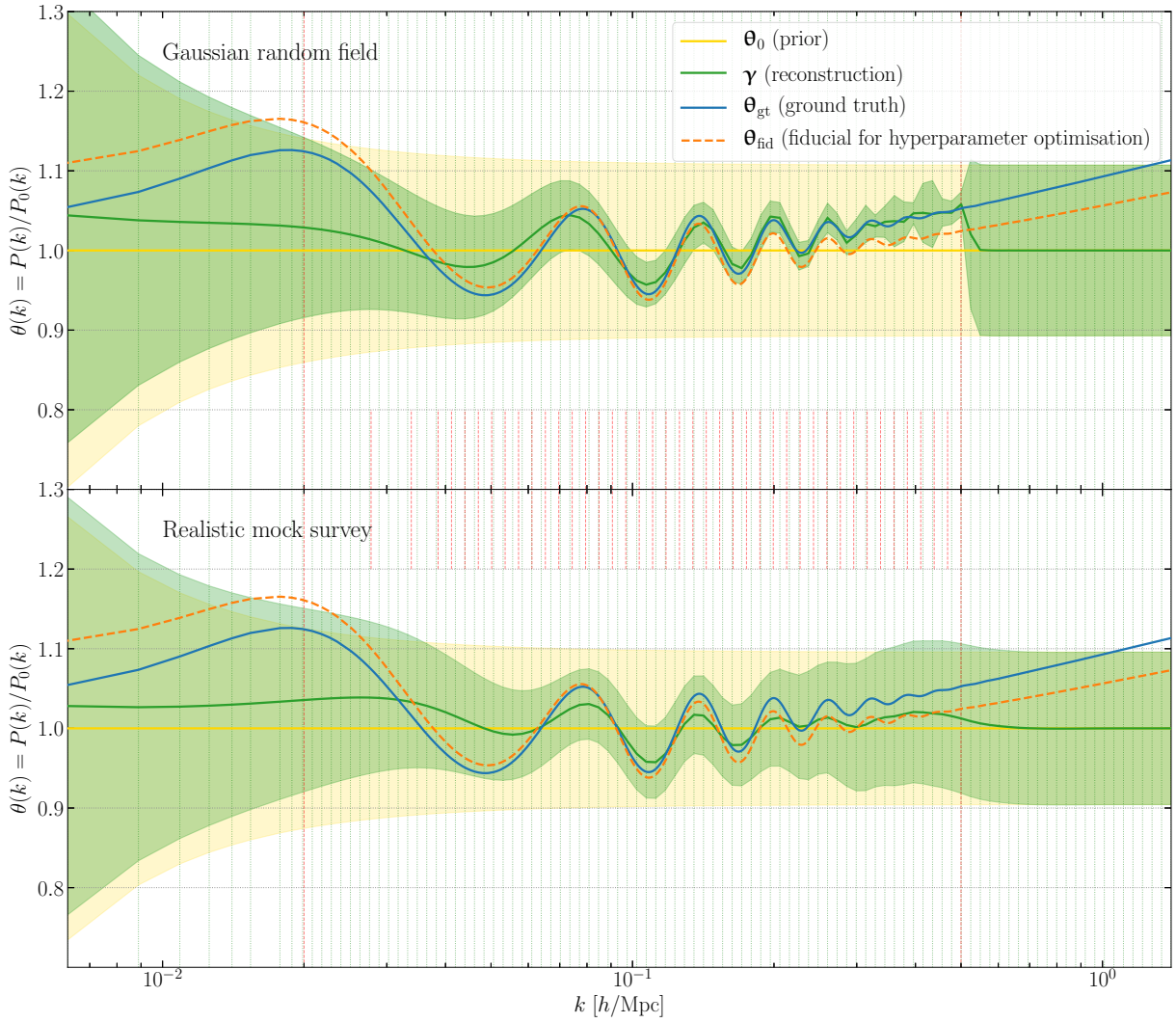

FIG. 10. Inference of the wiggle function $\theta(k) \equiv P(k)/P_0(k)$ as a function of wavenumber k, using as data model a Gaussian random field (top panel) or a realistic mock survey (bottom panel). The prior mean $\boldsymbol{\theta}_0$ and the effective posterior mean $\boldsymbol{\gamma}$ are represented as solid yellow and green lines, respectively, with their 2σ credible intervals (for the prior, $2\sigma = 2\,\theta_{\rm norm}(1+\alpha_{\rm cv}/k^{3/2})$). For comparison, the ground truth $\boldsymbol{\theta}_{\rm gt}$ and the fiducial "wiggly" function $\boldsymbol{\theta}_{\rm fid}$ used to optimise the prior hyperparameters are plotted as solid blue and dashed orange line, respectively. The dashed red lines correspond to the positions of the bins at which summaries are measured in data space; and the dotted green lines correspond to the support wavenumbers in parameter space. In the realistic case, in spite of survey complications which limit the information captured, the signature of BAOs is well reconstructed up to $k \approx 0.3$ h/Mpc, with 5 inferred acoustic peaks, result which could be improved using more volume. In the absence of informative data, the power spectrum reconstruction is driven towards the prior mean, but this effect does not affect cosmological parameter inference (see section IV C for details).

black-box models. This approach allows a much more accurate modelling of cosmological data, including in particular the complicated systematics they experience.

- Second, for MCMC methods, the number of data model evaluations is not fixed *a priori*, as some proposed samples are rejected during runtime. One has to assess the convergence of the chain. In contrast, our method only requires a fixed number of realisations to characterise the effective likelihood with the linearised black-box, all of which are used to obtain the inference result. For S target parameters and N_0 nuisance parameters realisations at the expansion point, N_s nuisance parameters realisations along each direction in parameter space, one has to perform $N_0 + N_s \times S$ data model evaluations. Increasing N_0 or N_s to get a better estimate of the required covariance matrix or gradient only increases the overall computational cost linearly.

- Third, while MCMC have to be computed sequen-

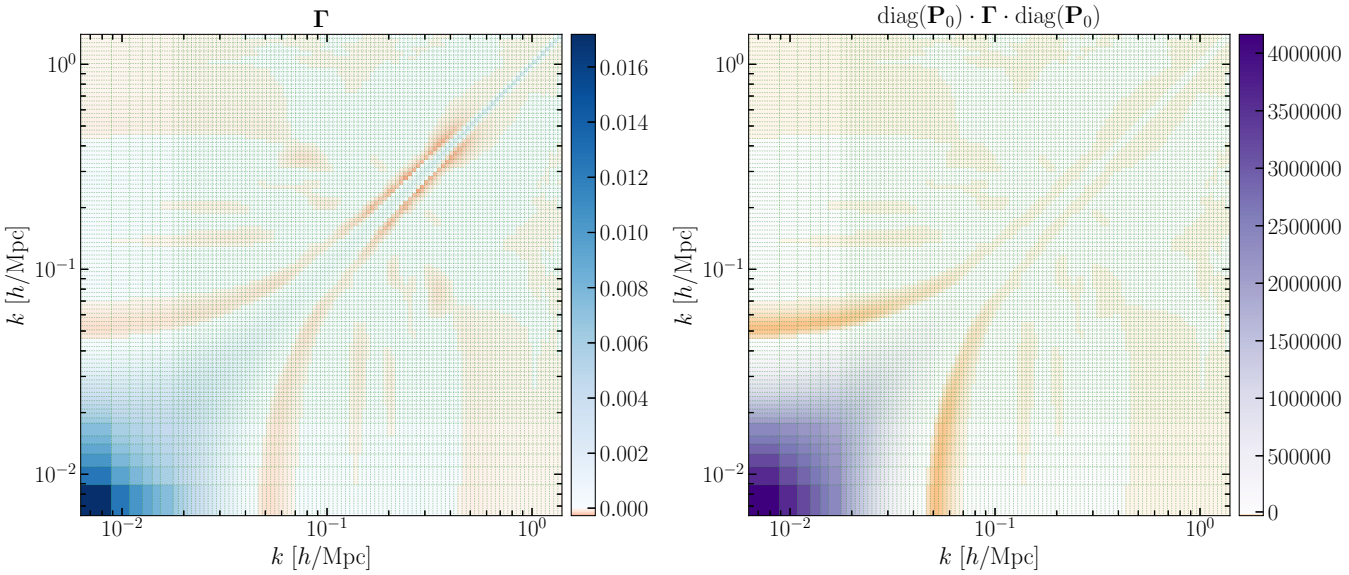

FIG. 11. Posterior covariance matrix $\mathbf{\Gamma}$ for the wiggle function $\boldsymbol{\theta}$ (left panel). The corresponding posterior covariance matrix for power spectrum amplitudes, $\mathrm{diag}(\mathbf{P}_0) \cdot \mathbf{\Gamma} \cdot \mathrm{diag}(\mathbf{P}_0)$, is shown in the right panel. The dotted green lines correspond to the support wavenumbers in parameter space. The correlation structure between different inferred parameters exhibits a complex behaviour (see also figure 9 for comparison).

tially, all the simulations required in the proposed method can be run simultaneously in parallel, or even on different machines. This allows a fast application of the method and makes it particularly suitable for grid computing.

- Finally, the linearised black-box is trained once and for all independently of the data. This means that if one acquires new data from the same survey, no additional black-box evaluations are required to perform inference, whereas likelihood-based techniques would require a new MCMC. Furthermore, if the cosmological simulations used are stored, they can even be used to perform inference from a different survey, by just replacing the part of the black-box corresponding to survey specifications.

SELFI also differs from other approaches to ABC (such as likelihood-free rejection sampling, Population Monte Carlo, DELFI, or BOLFI), which are limited by their inability to scale with the number of target parameters. By relying on an expansion of the simulator, SELFI allows the likelihood-free inference of $S \gtrsim 100$ parameters, as is necessary for a model-independent parametrisation of theory in cosmology.

In this work, we demonstrated that a "non-wiggly" expansion point $\boldsymbol{\theta}_0$ is sufficient to recover the target wiggle function $\boldsymbol{\theta}$ in the domain allowed by Planck priors. However, it shall be noted that if the solution is farther from the expansion point, then the method can be iterated. In this case, the posterior mean $\boldsymbol{\gamma}$ would be used as the new expansion point to train a new linearised black-box, used to obtain a new posterior. Using a sufficient number of iterations, we expect the effective posterior to converge to the true function, even if it strongly deviates from the first expansion point. We leave the detailed investigation of this idea to future studies.

In this paper, we showed a successful application of finite differencing to obtain the gradient of the averaged black-box $\nabla \mathbf{f}_0$, but our equations could be used with other techniques, such as automatic differentiation. The data covariance matrix at the expansion point $\mathbf{C}_0$ also needs to be evaluated; for this task, and for certain summary statistics, variance reduction techniques such as the use of fixed and paired simulations (Angulo & Pontzen, 2016; Villaescusa-Navarro *et al.*, 2018) or hybrid estimators (Hall & Taylor, 2019) could further be exploited.

As a proof of concept, we applied our technique in conjunction with the artificial galaxy survey simulator SIMBELMYNË, emulating relevant effects at play: non-linear gravitational structure formation, redshift-space distortions, a survey mask and selection function, and instrumental noise. As a result, the inferred primordial power spectrum is unbiased with a distinct identification of BAO wiggles, even far in the Silk damping tail. We also demonstrated that unbiased inference of cosmological parameters is possible. In spite of the non-linear evolution of structures on small scales, we are able to use the power spectrum of the galaxy field as summary statistic, up to $k_{\mathrm{max}} = 0.5\ h/\mathrm{Mpc}$. This represents an increase by a factor of ~ 5 in the number of modes used with respect to state-of-the-art perturbation theory and backward-modelling techniques, with perspectives for further improvements. Assuming that posteriors are Gaussian and modes are independent, this increase

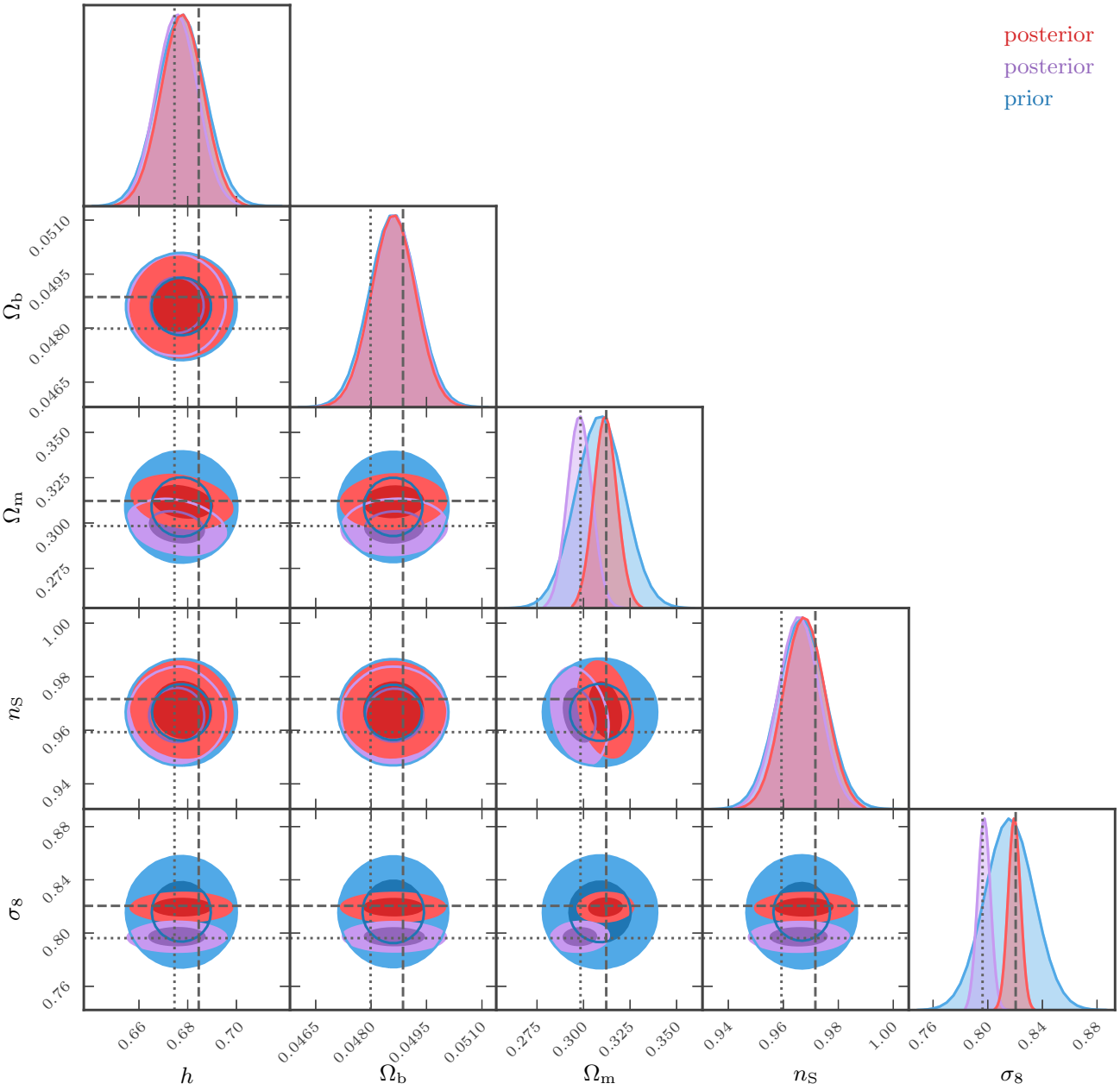

FIG. 12. Cosmological parameter inference using a linearised black-box model of galaxy surveys. The prior is shown in blue, and the effective posteriors for two different data realisations are shown in red and purple. The two different data realisations have different input cosmological parameters (shown as dashed and dotted lines), different phases and noise realisations. For all distributions, the 1σ and 2σ contours are shown.

translates into a reduction of the size of credible contours by a factor $\sim \sqrt{5}$.

The data model used in this work remains simplified with respect to some of the complications found in real galaxy surveys. However, due to the flexible nature of the method, it is straightforward to include additional aspects in the inference process: one only has to exchange the black-box for a more sophisticated one. We developed a python code reflecting this versatility, pySELFI, which we publicly released, together with documentation and the data necessary to reproduce the results of the present paper.[4] The application of this method to more complex models and to real survey data is left for future research.

[4] Currently, the code's homepage is hosted at

In conclusion, the method developed constitutes a computationally efficient and easily applicable framework to infer the primordial matter power spectrum and cosmological parameters from complex black-box mock observations. It allows the use of fully non-linear data models, as required for an optimal analysis of galaxy surveys. Other applications may include the cosmic microwave background, weak gravitational lensing, or the 21 cm signal of hydrogen. The prize for using full forward-modelling in these problems is a potentially vast gain of precision in cosmological constraints.

Appendix A: Derivation of the effective likelihood

In this appendix, we derive the approximate likelihood given in equation (11), starting from equation (9) and the assumptions detailed in section II A 3. We have

$$\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}|\boldsymbol{\theta}) \propto \int \mathcal{P}(\boldsymbol{\Phi}|\mathbf{s}) \left[\prod_{n=1}^{N} \mathcal{P}(\boldsymbol{\Phi}_\theta^{(i)}|\mathbf{s})\right] \mathrm{d}\mathbf{s}. \quad \text{(A1)}$$

Using the parametric form for $\mathcal{P}(\boldsymbol{\Phi}|\mathbf{s})$ (equation (10)) yields $\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}|\boldsymbol{\theta}) \propto \int \exp\left[\hat{\ell}_1(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}, \mathbf{s})\right] \mathrm{d}\mathbf{s}$, with

$$\begin{aligned} -2\hat{\ell}_1(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}, \mathbf{s}) \equiv\ & (\boldsymbol{\Phi} - \mathbf{s})^\intercal \boldsymbol{\Sigma}_\theta^{-1} (\boldsymbol{\Phi} - \mathbf{s}) \\ & + \sum_{i=1}^{N} (\boldsymbol{\Phi}_\theta^{(i)} - \mathbf{s})^\intercal \boldsymbol{\Sigma}_\theta^{-1} (\boldsymbol{\Phi}_\theta^{(i)} - \mathbf{s}) \\ & + (N+1) \log |2\pi \boldsymbol{\Sigma}_\theta|. \end{aligned} \quad \text{(A2)}$$

In order to evaluate the integral, we complete the square with respect to $\mathbf{s}$ in the argument of the exponential,

$$\begin{aligned} -2\hat{\ell}_1(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}, \mathbf{s}) &= \boldsymbol{\Phi}^\intercal \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Phi} + \sum_{i=1}^{N} \boldsymbol{\Phi}_\theta^{(i)\intercal} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Phi}_\theta^{(i)} \\ &\quad - 2(\boldsymbol{\Phi} + N\hat{\boldsymbol{\Phi}}_\theta)^\intercal \boldsymbol{\Sigma}_\theta^{-1} \mathbf{s} \\ &\quad + (N+1)\, \mathbf{s}^\intercal \boldsymbol{\Sigma}_\theta^{-1} \mathbf{s} \\ &\quad + (N+1) \log |2\pi \boldsymbol{\Sigma}_\theta| \\ &= \boldsymbol{\Phi}^\intercal \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Phi} + \sum_{i=1}^{N} \boldsymbol{\Phi}_\theta^{(i)\intercal} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Phi}_\theta^{(i)} \\ &\quad - \boldsymbol{\eta}^\intercal (N+1)\, \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\eta} \\ &\quad + (N+1) \big[(\mathbf{s} - \boldsymbol{\eta})^\intercal\, \boldsymbol{\Sigma}_\theta^{-1} (\mathbf{s} - \boldsymbol{\eta}) \\ &\quad\quad + \log |2\pi \boldsymbol{\Sigma}_\theta|\big], \end{aligned} \quad \text{(A3)}$$

where we have recognised $\hat{\boldsymbol{\Phi}}_\theta = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Phi}_\theta^{(i)}$ (equation (12)) and introduced $\boldsymbol{\eta} \equiv (\boldsymbol{\Phi} + N\hat{\boldsymbol{\Phi}}_\theta)/(N+1)$. After integration over $\mathbf{s}$, the last term gives a constant factor, so that $\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}|\boldsymbol{\theta}) \propto \exp\left[\hat{\ell}_2(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\})\right]$, with

$$\begin{aligned} -2\hat{\ell}_2(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}) \equiv\ & \boldsymbol{\Phi}^\intercal \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Phi} + \sum_{i=1}^{N} \boldsymbol{\Phi}_\theta^{(i)\intercal} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Phi}_\theta^{(i)} \\ & - (\boldsymbol{\Phi} + N\hat{\boldsymbol{\Phi}}_\theta)^\intercal \frac{1}{N+1}\, \boldsymbol{\Sigma}_\theta^{-1} (\boldsymbol{\Phi} + N\hat{\boldsymbol{\Phi}}_\theta). \end{aligned} \quad \text{(A4)}$$

We now complete the square with respect to $\boldsymbol{\Phi}$ to obtain

$$\begin{aligned} -2\hat{\ell}_2(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}) =\ & (\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}_\theta)^\intercal \left(\frac{N+1}{N} \boldsymbol{\Sigma}_\theta\right)^{-1} (\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}_\theta) \\ & + \text{constant terms}. \end{aligned} \quad \text{(A5)}$$

In order to obtain a computable approximation of the likelihood, the unknown covariance $\boldsymbol{\Sigma}_\theta$ in $\hat{\ell}_2(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\})$ has to be approximated by $\hat{\boldsymbol{\Sigma}}_\theta$, defined by equation (14). The covariance of the effective likelihood is therefore $\hat{\boldsymbol{\Sigma}}'_\theta \equiv \frac{N+1}{N} \hat{\boldsymbol{\Sigma}}_\theta$. The unknown inverse covariance $\boldsymbol{\Sigma}_\theta^{-1}$ is also replaced by its unbiased computable approximation $\hat{\boldsymbol{\Sigma}}_\theta^{-1}$, defined by equation (15). Finally, we use the normalisation condition $\int \mathcal{P}(\boldsymbol{\Phi}|\{\boldsymbol{\Phi}_\theta^{(i)}\}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\Phi} = \int \frac{\mathcal{P}(\boldsymbol{\Phi}, \{\boldsymbol{\Phi}_\theta^{(i)}\}|\boldsymbol{\theta})}{\mathcal{P}(\{\boldsymbol{\Phi}_\theta^{(i)}\}|\boldsymbol{\theta})}\, \mathrm{d}\boldsymbol{\Phi} = 1$ and evaluate at $\boldsymbol{\Phi} = \boldsymbol{\Phi}_\mathrm{O}$, as prescribed by equation (7), to obtain $\hat{\ell}^N(\boldsymbol{\theta})$ given by equation (11). When $N \to \infty$, $\hat{\boldsymbol{\Phi}}_\theta \longrightarrow \mathbf{s}$ and $\hat{\boldsymbol{\Sigma}}'_\theta \longrightarrow \boldsymbol{\Sigma}_\theta$, thus the limiting approximation is $\tilde{\ell}(\boldsymbol{\theta})$ given by equation (16).

Appendix B: Derivation of the effective posterior

We recall the canonical form of the Gaussian distribution with mean $\mathbf{x}_0$ and covariance matrix $\mathbf{X}$, given as

$$\begin{aligned} -2 \log \mathcal{P}(\mathbf{x}) &= \log |2\pi \mathbf{X}| + (\mathbf{x} - \mathbf{x}_0)^\intercal \mathbf{X}^{-1} (\mathbf{x} - \mathbf{x}_0) \\ &= \log |2\pi \mathbf{X}| + \boldsymbol{\xi}_0^\intercal \mathbf{X} \boldsymbol{\xi}_0 - 2\boldsymbol{\xi}_0^\intercal \mathbf{x} + \mathbf{x}^\intercal \mathbf{X}^{-1} \mathbf{x}, \end{aligned} \quad \text{(B1)}$$

where $\boldsymbol{\xi}_0 \equiv \mathbf{X}^{-1} \mathbf{x}_0$.

Using the linearised data model (equation (17)) in the expression of the effective likelihood (equation (19)), we get

$$\begin{aligned} -2\hat{\ell}^N(\boldsymbol{\theta}) &= \log |2\pi \mathbf{C}_0| + [\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0 - \nabla \mathbf{f}_0 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)]^\intercal \cdot \\ &\qquad \mathbf{C}_0^{-1} [\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0 - \nabla \mathbf{f}_0 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)] \\ &= \log |2\pi \mathbf{C}_0| + (\mathbf{y}_0 - \boldsymbol{\theta})^\intercal \mathbf{N}_0^{-1} (\mathbf{y}_0 - \boldsymbol{\theta}), \end{aligned} \quad \text{(B2)}$$

where we have defined

$$\mathbf{N}_0 \equiv \left[(\nabla \mathbf{f}_0)^\intercal \mathbf{C}_0^{-1} \nabla \mathbf{f}_0\right]^{-1} \quad \text{(B3)}$$

and

$$\mathbf{y}_0 \equiv \boldsymbol{\theta}_0 + (\nabla \mathbf{f}_0)^{-1} \cdot (\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0), \quad \text{(B4)}$$

$(\nabla \mathbf{f}_0)^{-1}$ denoting the adjoint of the Jacobian characterising the linearised black-box (its computation will not

---

http://pyselfi.florent-leclercq.eu; the sources are available on GitHub at https://github.com/florent-leclercq/pyselfi; and the documentation is on Read the Docs at https://pyselfi.readthedocs.io.

be necessary). In canonical form, the Gaussian effective likelihood is written

$$-2\hat{\ell}^N(\boldsymbol{\theta}) = \log|2\pi\mathbf{C}_0| + \boldsymbol{\mu}_0^\intercal\mathbf{N}_0\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0\boldsymbol{\theta} + \boldsymbol{\theta}^\intercal\mathbf{N}_0^{-1}\boldsymbol{\theta}, \tag{B5}$$

with $\boldsymbol{\mu}_0 \equiv \mathbf{N}_0^{-1}\mathbf{y}_0 = \mathbf{N}_0^{-1}\boldsymbol{\theta}_0 + (\nabla\mathbf{f}_0)^\intercal\mathbf{C}_0^{-1}(\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0)$. Similarly, the prior (equation (23)) is written

$$-2\log\mathcal{P}(\boldsymbol{\theta}) = \log|2\pi\mathbf{S}| + \boldsymbol{\eta}_0^\intercal\mathbf{S}\boldsymbol{\eta}_0 - 2\boldsymbol{\eta}_0^\intercal\boldsymbol{\theta} + \boldsymbol{\theta}^\intercal\mathbf{S}^{-1}\boldsymbol{\theta} \tag{B6}$$

where $\boldsymbol{\eta}_0 \equiv \mathbf{S}^{-1}\boldsymbol{\theta}_0$.

Adding the two expressions, we find that the effective posterior verifies

$$\begin{aligned} -2\log\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_\mathrm{O}} = &-2(\boldsymbol{\mu}_0 + \boldsymbol{\eta}_0)^\intercal\boldsymbol{\theta} \\ &+\boldsymbol{\theta}^\intercal(\mathbf{N}_0^{-1} + \mathbf{S}^{-1})\boldsymbol{\theta} \\ &+ \text{constant terms.} \end{aligned} \tag{B7}$$

This is the canonical form of a Gaussian distribution, where the covariance matrix is identified as $\boldsymbol{\Gamma} \equiv (\mathbf{N}_0^{-1} + \mathbf{S}^{-1})^{-1}$, giving equation (26), and the mean is identified as

$$\begin{aligned} \boldsymbol{\gamma} &= \boldsymbol{\Gamma}(\boldsymbol{\mu}_0 + \boldsymbol{\eta}_0) \\ &= \boldsymbol{\Gamma}\mathbf{N}_0^{-1}\boldsymbol{\theta}_0 + \boldsymbol{\Gamma}(\nabla\mathbf{f}_0)^\intercal\mathbf{C}_0^{-1}(\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0) + \boldsymbol{\Gamma}\mathbf{S}^{-1}\boldsymbol{\theta}_0, \end{aligned} \tag{B8}$$

giving equation (25).

Note that the above calculation is analogous to the derivation of the Wiener filter equations: assuming a linear data model ($\mathbf{d} = \mathbf{s}+\mathbf{n}$), a prior with mean $\bar{\mathbf{s}}$ and signal covariance $\mathbf{S}$, and a likelihood with mean $\bar{\mathbf{d}}$ and noise covariance $\mathbf{N}$, the filter covariance is $(\mathbf{N}^{-1} + \mathbf{S}^{-1})^{-1}$ and the filtered signal is $\bar{\mathbf{s}} + (\mathbf{N}^{-1} + \mathbf{S}^{-1})^{-1}\mathbf{N}^{-1}(\mathbf{d}_\mathrm{O} - \bar{\mathbf{d}})$.

STATEMENT OF CONTRIBUTION

Study concept and design (JJ, FL, WE); design of prior optimisation and cosmological parameter inference (FL); original code implementation of the filter equations (WE); code rewriting and enhancements (FL); design and implementation of the data model (FL); running of the simulations (FL); drafting of the manuscript (WE); critical revision of the manuscript (FL); proofreading (FL, WE, JJ, AH); supervision (JJ, FL); support and interpretation of results (JJ, AH). All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

FL is grateful to Guilhem Lavaux and Andrew Jaffe for useful discussions. This work has made use of a modified version of PYGTC (Bocquet & Carter, 2016). Numerical computations were done on the cx1 cluster hosted by the Research Computing Service facilities at Imperial College London (doi:10.14469/hpc/2232). This work is done within the Aquila Consortium (https://aquila-consortium.org).

FL acknowledges funding from the Imperial College London Research Fellowship Scheme. This research was supported by the DFG cluster of excellence "Origin and Structure of the Universe" (www.universe-cluster.de).

REFERENCES

(Akeret *et al.*, 2015) J. Akeret, A. Refregier, A. Amara, S. Seehars, C. Hasner, *Approximate Bayesian computation for forward modeling in cosmology*, Journal of Cosmology and Astroparticle Physics **8**, 043 (2015), arXiv:1504.07245.

(Albrecht *et al.*, 2006) A. Albrecht, G. Bernstein, R. Cahn, W. L. Freedman, J. Hewitt, W. Hu, J. Huth, M. Kamionkowski, E. W. Kolb, L. Knox, J. C. Mather, S. Staggs, N. B. Suntzeff, *Report of the Dark Energy Task Force*, ArXiv Astrophysics e-prints (2006), astro-ph/0609591.

(Alsing, Wandelt & Feeney, 2018) J. Alsing, B. Wandelt, S. Feeney, *Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology*, Mon. Not. R. Astron. Soc. **477**, 2874 (2018), arXiv:1801.01497.

(Angulo & Pontzen, 2016) R. E. Angulo, A. Pontzen, *Cosmological N-body simulations with suppressed variance*, Mon. Not. R. Astron. Soc. **462**, L1 (2016), arXiv:1603.05253.

(Bardeen *et al.*, 1986) J. M. Bardeen, J. R. Bond, N. Kaiser, A. S. Szalay, *The statistics of peaks of Gaussian random fields*, Astrophys. J. **304**, 15 (1986).

(Blanton *et al.*, 2003) M. R. Blanton, D. W. Hogg, N. A. Bahcall, J. Brinkmann, M. Britton, A. J. Connolly, I. Csabai, M. Fukugita, J. Loveday, A. Meiksin, J. A. Munn, R. C. Nichol, S. Okamura, T. Quinn, D. P. Schneider, K. Shimasaku, M. A. Strauss, M. Tegmark, M. S. Vogeley, D. H. Weinberg, *The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1*, Astrophys. J. **592**, 819 (2003), astro-ph/0210215.

(Bocquet & Carter, 2016) S. Bocquet, F. W. Carter, *pygtc: beautiful parameter covariance plots (aka. Giant Triangle Confusograms)*, The Journal of Open Source Software **1** (2016), 10.21105/joss.00046.

(Bos, Kitaura & van de Weygaert, 2019) E. G. P. Bos, F.-S. Kitaura, R. van de Weygaert, *Bayesian cosmic density field inference from redshift space dark matter maps*, Mon. Not. R. Astron. Soc. **488**, 2573 (2019), arXiv:1810.05189.

(Bouchet, 1996) F. R. Bouchet, *Introductory Overview of Eulerian and Lagrangian Perturbation Theories*, in *Dark Matter in the Universe*, edited by S. Bonometto, J. R. Primack, A. Provenzale (1996) p. 565, astro-ph/9603013.

(Bouchet *et al.*, 1995) F. R. Bouchet, S. Colombi, E. Hivon, R. Juszkiewicz, *Perturbative Lagrangian approach to gravitational instability*, Astron. & Astrophys. **296**, 575 (1995), astro-ph/9406013.

(Burden, Percival & Howlett, 2015) A. Burden, W. J. Percival, C. Howlett, *Reconstruction in Fourier space*, Mon. Not. R. Astron. Soc. **453**, 456 (2015), arXiv:1504.02591.

(Byrd *et al.*, 1995) R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, *A Limited Memory Algorithm for Bound Constrained Optimization*, SIAM Journal on Scientific Computing **16**, 1190 (1995).

(Doumler *et al.*, 2013) T. Doumler, Y. Hoffman, H. Courtois, S. Gottlöber, *Reconstructing cosmological initial conditions from galaxy peculiar velocities - I. Reverse Zeldovich Approximation*, Mon. Not. R. Astron. Soc. **430**, 888 (2013), arXiv:1212.2806.

(Eisenstein, 2005) D. J. Eisenstein, *Dark energy and cosmic sound [review article]*, New Astronomy Reviews **49**, 360 (2005).

(Eisenstein & Hu, 1998) D. J. Eisenstein, W. Hu, *Baryonic Features in the Matter Transfer Function*, Astrophys. J. **496**, 605 (1998), astro-ph/9709112.

(Eisenstein, Seo & White, 2007) D. J. Eisenstein, H.-J. Seo, M. White, *On the Robustness of the Acoustic Scale in the Low-*

Redshift Clustering of Matter, Astrophys. J. **664**, 660 (2007), astro-ph/0604361.

(Eisenstein *et al.*, 2007) D. J. Eisenstein, H.-J. Seo, E. Sirko, D. N. Spergel, *Improving Cosmological Distance Measurements by Reconstruction of the Baryon Acoustic Peak*, Astrophys. J. **664**, 675 (2007), astro-ph/0604362.

(Foreman-Mackey *et al.*, 2013) D. Foreman-Mackey, D. W. Hogg, D. Lang, J. Goodman, *emcee: The MCMC Hammer*, Publications of the Astronomical Society of the Pacific **125**, 306 (2013), arXiv:1202.3665 [astro-ph.IM].

(Granett *et al.*, 2015) B. R. Granett, E. Branchini, L. Guzzo, U. Abbas, C. Adami, S. Arnouts, J. Bel, M. Bolzonella, D. Bottini, A. Cappi, J. Coupon, O. Cucciati, I. Davidzon, G. De Lucia, S. de la Torre, A. Fritz, P. Franzetti, M. Fumana, B. Garilli, O. Ilbert, A. Iovino, J. Krywult, V. Le Brun, O. Le Fèvre, D. Maccagni, K. Małek, F. Marulli, H. J. McCracken, M. Polletta, A. Pollo, M. Scodeggio, L. A. M. Tasca, R. Tojeiro, D. Vergani, A. Zanichelli, A. Burden, C. Di Porto, A. Marchetti, C. Marinoni, Y. Mellier, T. Moutard, L. Moscardini, R. C. Nichol, J. A. Peacock, W. J. Percival, G. Zamorani, *The VIMOS Public Extragalactic Redshift Survey. Reconstruction of the redshift-space galaxy density field*, Astron. & Astrophys. **583**, A61 (2015), arXiv:1505.06337.

(Hall & Taylor, 2019) A. Hall, A. Taylor, *A Bayesian method for combining theoretical and simulated covariance matrices for large-scale structure surveys*, Mon. Not. R. Astron. Soc. **483**, 189 (2019), arXiv:1807.06875.

(Hartlap, Simon & Schneider, 2007) J. Hartlap, P. Simon, P. Schneider, *Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix*, Astron. & Astrophys. **464**, 399 (2007), astro-ph/0608064.

(Hockney & Eastwood, 1981) R. W. Hockney, J. W. Eastwood, *Computer Simulation Using Particles* (McGraw-Hill, 1981).

(Ishida *et al.*, 2015) E. E. O. Ishida, S. D. P. Vitenti, M. Penna-Lima, J. Cisewski, R. S. de Souza, A. M. M. Trindade, E. Cameron, V. C. Busti, C. Collaboration, *COSMOABC: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation*, Astronomy and Computing **13**, 1 (2015), arXiv:1504.06129.

(Jasche & Lavaux, 2015) J. Jasche, G. Lavaux, *Matrix-free large-scale Bayesian inference in cosmology*, Mon. Not. R. Astron. Soc. **447**, 1204 (2015), arXiv:1402.1763.

(Jasche & Lavaux, 2019) J. Jasche, G. Lavaux, *Physical Bayesian modelling of the non-linear matter distribution: New insights into the nearby universe*, Astron. & Astrophys. **625**, A64 (2019), arXiv:1806.11117.

(Jasche & Wandelt, 2013a) J. Jasche, B. D. Wandelt, *Bayesian physical reconstruction of initial conditions from large-scale structure surveys*, Mon. Not. R. Astron. Soc. **432**, 894 (2013a), arXiv:1203.3639 [astro-ph.CO].

(Jasche & Lavaux, 2017) J. Jasche, G. Lavaux, *Bayesian power spectrum inference with foreground and target contamination treatment*, Astron. & Astrophys. **606**, A37 (2017), arXiv:1706.08971.

(Jasche & Wandelt, 2013b) J. Jasche, B. D. Wandelt, *Methods for Bayesian Power Spectrum Inference with Galaxy Surveys*, Astrophys. J. **779**, 15 (2013b), arXiv:1306.1821 [astro-ph.CO].

(Jasche, Leclercq & Wandelt, 2015) J. Jasche, F. Leclercq, B. D. Wandelt, *Past and present cosmic structure in the SDSS DR7 main sample*, Journal of Cosmology and Astroparticle Physics **1**, 036 (2015), arXiv:1409.6308 [astro-ph.CO].

(Jasche *et al.*, 2010) J. Jasche, F. S. Kitaura, B. D. Wandelt, T. A. Enßlin, *Bayesian power-spectrum inference for large-scale structure data*, Mon. Not. R. Astron. Soc. **406**, 60 (2010), arXiv:0911.2493 [astro-ph.CO].

(Jeffrey & Abdalla, 2018) N. Jeffrey, F. B. Abdalla, *Parameter inference and model comparison using theoretical predictions from noisy simulations*, arXiv e-prints (2018), arXiv:1809.08246.

(Jennings & Madigan, 2017) E. Jennings, M. Madigan, *astroABC : An Approximate Bayesian Computation Sequential Monte Carlo sampler for cosmological parameter estimation*, Astronomy and Computing **19**, 16 (2017), arXiv:1608.07606 [astro-ph.IM].

(Laureijs *et al.*, 2011) R. Laureijs, J. Amiaux, S. Arduini, J. . Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, et al., *Euclid Definition Study Report*, ArXiv e-prints (2011), arXiv:1110.3193 [astro-ph.CO].

(Lavaux & Jasche, 2016) G. Lavaux, J. Jasche, *Unmasking the masked Universe: the 2M++ catalogue through Bayesian eyes*, Mon. Not. R. Astron. Soc. **455**, 3169 (2016), arXiv:1509.05040 [astro-ph.CO].

(Leclercq, 2018) F. Leclercq, *Bayesian optimization for likelihood-free cosmological inference*, Phys. Rev. D **98**, 063511 (2018), arXiv:1805.07152.

(Leclercq, Jasche & Wandelt, 2015) F. Leclercq, J. Jasche, B. Wandelt, *Bayesian analysis of the dynamic cosmic web in the SDSS galaxy survey*, Journal of Cosmology and Astroparticle Physics **6**, 015 (2015), arXiv:1502.02690 [astro-ph.CO].

(LSST Science Collaboration, 2012) LSST Science Collaboration, *Large Synoptic Survey Telescope: Dark Energy Science Collaboration*, ArXiv e-prints (2012), arXiv:1211.0310 [astro-ph.CO].

(Majerotto *et al.*, 2012) E. Majerotto, L. Guzzo, L. Samushia, W. J. Percival, Y. Wang, S. de la Torre, B. Garilli, P. Franzetti, E. Rossetti, A. Cimatti, C. Carbone, N. Roche, G. Zamorani, *Probing deviations from general relativity with the Euclid spectroscopic survey*, Mon. Not. R. Astron. Soc. **424**, 1392 (2012), arXiv:1205.6215.

(Meiksin, White & Peacock, 1999) A. Meiksin, M. White, J. A. Peacock, *Baryonic signatures in large-scale structure*, Mon. Not. R. Astron. Soc. **304**, 851 (1999), astro-ph/9812214.

(Moutarde *et al.*, 1991) F. Moutarde, J.-M. Alimi, F. R. Bouchet, R. Pellat, A. Ramani, *Precollapse scale invariance in gravitational instability*, Astrophys. J. **382**, 377 (1991).

(Padmanabhan *et al.*, 2012) N. Padmanabhan, X. Xu, D. J. Eisenstein, R. Scalzo, A. J. Cuesta, K. T. Mehta, E. Kazin, *A 2 per cent distance to z = 0.35 by reconstructing baryon acoustic oscillations - I. Methods and application to the Sloan Digital Sky Survey*, Mon. Not. R. Astron. Soc. **427**, 2132 (2012), arXiv:1202.0090.

(Peacock, 1999) J. A. Peacock, *Cosmological Physics* (Cambridge University Press, 1999).

(Peacock & Heavens, 1985) J. A. Peacock, A. F. Heavens, *The statistics of maxima in primordial density perturbations*, Mon. Not. R. Astron. Soc. **217**, 805 (1985).

(Peebles, 1980) P. J. E. Peebles, *The large-scale structure of the universe* (Princeton University Press, 1980).

(Percival *et al.*, 2007) W. J. Percival, S. Cole, D. J. Eisenstein, R. C. Nichol, J. A. Peacock, A. C. Pope, A. S. Szalay, *Measuring the Baryon Acoustic Oscillation scale using the Sloan Digital Sky Survey and 2dF Galaxy Redshift Survey*, Mon. Not. R. Astron. Soc. **381**, 1053 (2007), arXiv:0705.3323.

(Planck Collaboration, 2018a) Planck Collaboration, *Planck 2018 results. I. Overview and the cosmological legacy of Planck*, arXiv e-prints (2018a), arXiv:1807.06205.

(Planck Collaboration, 2018b) Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, arXiv e-prints (2018b), arXiv:1807.06209.

(Planck Collaboration, 2016) Planck Collaboration, *Planck 2015 results. XIII. Cosmological parameters*, Astron. & Astrophys. **594**, A13 (2016), arXiv:1502.01589.

(Price *et al.*, 2018) L. F. Price, C. C. Drovandi, A. Lee, D. J. Nott, *Bayesian Synthetic Likelihood*, Journal of Computational and Graphical Statistics **27:1**, 1 (2018).

(Rasmussen & Williams, 2006) C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptative computation and machine learning series (University Press Group Limited, 2006).

(Ross *et al.*, 2015) A. J. Ross, L. Samushia, C. Howlett, W. J. Percival, A. Burden, M. Manera, *The clustering of the SDSS*

DR7 main Galaxy sample - I. A 4 per cent distance measure at z = 0.15, Mon. Not. R. Astron. Soc. **449**, 835 (2015), arXiv:1409.3242.

(Ross *et al.*, 2012) A. J. Ross, W. J. Percival, A. G. Sánchez, L. Samushia, S. Ho, E. Kazin, M. Manera, B. Reid, M. White, R. Tojeiro, C. K. McBride, X. Xu, D. A. Wake, M. A. Strauss, F. Montesano, M. E. C. Swanson, S. Bailey, A. S. Bolton, A. M. Dorta, D. J. Eisenstein, H. Guo, J.-C. Hamilton, R. C. Nichol, N. Padmanabhan, F. Prada, D. J. Schlegel, M. V. Magaña, I. Zehavi, M. Blanton, D. Bizyaev, H. Brewington, A. J. Cuesta, E. Malanushenko, V. Malanushenko, D. Oravetz, J. Parejko, K. Pan, D. P. Schneider, A. Shelden, A. Simmons, S. Snedden, G.-b. Zhao, *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: analysis of potential systematics*, Mon. Not. R. Astron. Soc. **424**, 564 (2012), arXiv:1203.6499.

(Schechter, 1976) P. Schechter, *An analytic expression for the luminosity function for galaxies.*, Astrophys. J. **203**, 297 (1976).

(Seljak *et al.*, 2017) U. Seljak, G. Aslanyan, Y. Feng, C. Modi, *Towards optimal extraction of cosmological information from nonlinear data*, Journal of Cosmology and Astroparticle Physics **12**, 009 (2017), arXiv:1706.06645.

(Sellentin & Heavens, 2016) E. Sellentin, A. F. Heavens, *Parameter inference with estimated covariance matrices*, Mon. Not. R. Astron. Soc. **456**, L132 (2016), arXiv:1511.05969 [astro-ph.CO].

(Silk, 1968) J. Silk, *Cosmic Black-Body Radiation and Galaxy Formation*, Astrophys. J. **151**, 459 (1968).

(Smith & Marian, 2015) R. E. Smith, L. Marian, *Towards optimal estimation of the galaxy power spectrum*, Mon. Not. R. Astron. Soc. **454**, 1266 (2015), arXiv:1503.06830.

(Springel, 2005) V. Springel, *The cosmological simulation code GADGET-2*, Mon. Not. R. Astron. Soc. **364**, 1105 (2005), astro-ph/0505010.

(Tassev, Zaldarriaga & Eisenstein, 2013) S. Tassev, M. Zaldarriaga, D. J. Eisenstein, *Solving large scale structure in ten easy steps with COLA*, Journal of Cosmology and Astroparticle Physics **6**, 036 (2013), arXiv:1301.0322 [astro-ph.CO].

(Verde *et al.*, 2002) L. Verde, A. F. Heavens, W. J. Percival, S. Matarrese, C. M. Baugh, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, M. Colless, C. Collins, W. Couch, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, P. Norberg, J. A. Peacock, B. A. Peterson, W. Sutherland, K. Taylor, *The 2dF Galaxy Redshift Survey: the bias of galaxies and the density of the Universe*, Mon. Not. R. Astron. Soc. **335**, 432 (2002), arXiv:astro-ph/0112161.

(Villaescusa-Navarro *et al.*, 2018) F. Villaescusa-Navarro, S. Naess, S. Genel, A. Pontzen, B. Wandelt, L. Anderson, A. Font-Ribera, N. Battaglia, D. N. Spergel, *Statistical Properties of Paired Fixed Fields*, Astrophys. J. **867**, 137 (2018), arXiv:1806.01871.

(Wandelt, Larson & Lakshminarayanan, 2004) B. D. Wandelt, D. L. Larson, A. Lakshminarayanan, *Global, exact cosmic microwave background data analysis using Gibbs sampling*, Phys. Rev. D **70**, 083511 (2004), astro-ph/0310080.

(Wang *et al.*, 2013) H. Wang, H. J. Mo, X. Yang, F. C. van den Bosch, *Reconstructing the Initial Density Field of the Local Universe: Methods and Tests with Mock Catalogs*, Astrophys. J. **772**, 63 (2013), arXiv:1301.1348 [astro-ph.CO].

(Wang *et al.*, 2014) H. Wang, H. J. Mo, X. Yang, Y. P. Jing, W. P. Lin, *ELUCID—Exploring the Local Universe with the Reconstructed Initial Density Field. I. Hamiltonian Markov Chain Monte Carlo Method with Particle Mesh Dynamics*, Astrophys. J. **794**, 94 (2014), arXiv:1407.3451 [astro-ph.CO].

(White, 2015) M. White, *Reconstruction within the Zeldovich approximation*, Mon. Not. R. Astron. Soc. **450**, 3822 (2015), arXiv:1504.03677.

(Wiener, 1964) N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, Principles of Electrical Engineering Series (M.I.T. Press, 1964).

(Wood, 2010) S. N. Wood, *Statistical inference for noisy nonlinear ecological dynamic systems*, Nature **466**, 1102 (2010).