Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Received: *January 1, 2024*
Revised: *June 12, 2024*
Accepted: *June 18, 2024*
Published: *July 8, 2024*

Reconstructing the long-wavelength matter density fluctuation modes from the scalar-type clustering fossils

Zhenyuan Wang [a,b] **and Donghui Jeong** [a,b,c]

[a] *Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 16802, U.S.A.*
[b] *Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, U.S.A.*
[c] *School of Physics, Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, Seoul, 02455, Republic of Korea*

E-mail: zzw173@psu.edu, djeong@psu.edu

Abstract: Revealing the large-scale structure from the 21cm intensity mapping surveys is only possible after the foreground cleaning. However, most current cleaning techniques relying on the smoothness of the foreground spectrum lead to a severe side effect of removing the large-scale structure signal along the line of sight. On the other hand, the clustering fossil, a coherent variation of the small-scale clustering over large scales, allows us to recover the long-wavelength density modes from the off-diagonal correlation between short-wavelength modes. In this paper, we revisit the reconstruction based on the short-wavelength matter density modes in real space and scrutinize the requirements for an unbiased and optimal clustering-fossil estimator. We show that (A) the estimator is unbiased only when using an accurate bispectrum model for the long-short-short mode coupling and (B) including the connected four-point correlation functions is essential for characterizing the noise power spectrum of the estimated long mode. For matter in real space, the clustering fossil estimator based upon the leading-order bispectrum yields an unbiased estimation of the long-wavelength ($k \lesssim 0.01$ [h/Mpc]) modes with the cross-correlation coefficient of 0.7 at redshifts $z = 0$ to 3.

Keywords: galaxy clustering, power spectrum

ArXiv ePrint: 2312.17321

https://doi.org/10.1088/1475-7516/2024/07/020

JCAP07(2024)020

Contents

1 Introduction

The intensity mapping of the 21cm hyper-fine transition line (21cmIM) can be a powerful probe of the large-scale structure of the Universe [1, 2]. With the 21cm line's low opacity and spectroscopic nature, the 21cmIM will provide a three-dimensional map of neutral hydrogen distribution over the unprecedently large volume. The ongoing and future 21cmIM projects, such as Tianlai [3], CHIME [4], HIRAX [5], BINGO [6], and SKA [7], are designed to observe the large-scale structure at redshifts $z = 0.5-6$, to measure, for example, the Hubble expansion rate and the angular diameter distance from the baryon acoustic oscillation (BAO) [8].

One of the main challenges of these 21cmIM experiments is the contamination from the extremely loud foreground from synchrotron, free-free, and thermal dust emissions. For example, the dominating galactic synchrotron emission is more than five orders of magnitude louder compared to the targeted large-scale structure signal in 21cmIM [9, 10]. The standard foreground cleaning method [11–21] takes advantage of the fact that synchrotron radiation is smooth in the frequency domain, while the 21cmIM signals show genuine cosmological density fluctuations along the radial direction. After the foreground cleaning, however, the long-wavelength Fourier modes parallel to the line of sight become inaccessible [10, 22]. Furthermore, the frequency-dependence response of the instrument can lead to a *foreground wedge* contamination in the Fourier space for $k_\parallel < k_\perp \tan\psi$, where the angle ψ increases with the field of view [23–29].

JCAP07(2024)020

These long-wavelength modes plagued by foreground cleaning are usually in the linear regime where the observed galaxy clustering statistical can be modeled by the Kaiser formula [30, 31], or near horizon scales where general relativistic corrections are important (see ref. [32] for a review). For both cases, the theoretical model is well understood in the framework of linear perturbation theory, and the accurate measurement of these long-wavelength modes will be translated to the measurement of the growth rate of the cosmic density field and metric perturbations as well as the local-type primordial non-Gaussianities [33]. Measuring these parameters, in turn, will inform us of the nature of dark energy and the physics of the primordial universe, respectively. In particular, given the large volume that 21cmIM covers, the large number of long-wavelength modes could provide an unprecedented tight constraint on cosmological parameters [34–37]. Therefore, to fully exploit the information from 21cmIM, it is crucial to recover the contaminated long-wavelength Fourier modes (hereafter, "long mode," for short).

To recover the contaminated long modes, ref. [38] have proposed an innovative technique called *cosmic-tide* reconstruction. The basic idea of the reconstruction is to exploit the non-Gaussianity of the density field coming from the nonlinear gravitational evolution that couples Fourier modes of different wavelengths. Namely, knowing the details of this coupling should allow us to infer the long-wavelength Fourier modes from the short-wavelength Fourier modes (hereafter, "short mode," for short). Based on this idea, ref. [39] developes a quadratic estimator for the long modes based on the anisotropic modulation of short modes' power spectrum by the large-scale tidal field. The coupling coefficient in the estimator is determined by the leading-order tidal interaction [40]. Applying this quadratic estimator to the N-body simulation results shows that the cross-correlation coefficient between the reconstructed long mode and the original long mode can reach 0.9 at $k < 0.1\ h/\mathrm{Mpc}$. This implies that the phase of the reconstructed long mode is highly correlated to the true long mode. Ref. [41] has further confirmed this result. Additional insights into tide reconstruction from tracers in real and redshift space are provided by [42] and [43]. Also see [44] for its application to recovering missing radial long modes in 21cm intensity mapping surveys, and [45] in the context of lensing reconstruction of line intensity mapping.

While following the spirit of cosmic-tide reconstruction, this paper investigates a different mathematical framework, called *clustering fossil*, for reconstructing the long modes. The clustering fossil utilizes the three-point non-Gaussian correlation between long mode and two short modes, which generates, in Fourier space, the off-diagonal two-point correlators between two short modes [46]. In a statistically homogeneous universe, the two-point correlation function $\langle \delta^*(\boldsymbol{k})\delta(\boldsymbol{k}')\rangle$ in Fourier space only has diagonal components proportional to $\delta^D(\boldsymbol{k}-\boldsymbol{k}')$: the power spectrum. In the presence of long modes, however, the non-Gaussian coupling between short modes and long modes makes locally measured two-point correlators differ from place to place. The statistical homogeneity is broken up locally, and the non-zero off-diagonal two-point correlators among short modes contain information about the long mode coupling with them. As we show in section 2.2, this off-diagonal correlation is proportional to the amplitude of the long mode and the coupling coefficient, which can be read from the squeezed-limit bispectrum.

Following ref. [46], we construct a quadratic clustering-fossil estimator for the long modes based on the off-diagonal correlation between short modes. Unlike ref. [46], which pursues

JCAP07(2024)020

inflationary spectator fields using the non-Gaussian coupling in the early universe [47, 48]; however, we focus on the non-Gaussianities caused by nonlinear matter clustering. In particular, applying the method to the case of 21cmIM, we treat the contaminated long mode as a scalar-type fossil field. Along this line, refs. [49, 50] have applied the clustering-fossil estimator to reconstruct the long mode based on the leading-order (tree-level) bispectrum in standard cosmological perturbation theory (SPT) [51]. Analyzing a suite of N-body simulations, they have recovered the long mode, with morphological features similar to the ground truth. They found, however, that the power spectrum of the recovered long mode is biased, which is consistent with the findings of ref. [39]. A further study [52] has applied this technique to the biased tracer fields to recover the matter density long mode and forecast the improved constraint on primordial non-Gaussianity through multi-tracer method, where the galaxy bias are also included into the estimator.

To proceed with the clustering-fossil-based reconstruction method, we need a more thorough theoretical understanding of the method's applicability and regime of validity. The goal of this paper is to scrutinize the method. We aim to provide a theoretical framework to explain the systematic bias observed in the N-body simulations and to characterize the statistical uncertainties of the recovered long mode. To solve the problem for the fully nonlinear density field, we need to include the coupling effects between long and short modes beyond the leading order and the non-Gaussian statistics of the short modes. To assess the requirement for a robust reconstruction method, we take a simpler approach here. Namely, instead of analyzing the result from full N-body simulations, we test the reconstruction method against a *controlled sample* density field only containing the first- and second-order density perturbations from the GridSPT (grid-based calculation of standard perturbation theory) [53–55]. Because the GridSPT density field strictly follows SPT, the fossil estimator constructed from the leading-order bispectrum can fully capture the coupling between the long and short modes. Therefore, testing the clustering fossil estimator in this way allows us to disentangle the effect of higher-order coupling from other effects, such as the non-Gaussianity of short modes. To inject the effect from the higher-order nonlinear couplings, we use the 2LPT (second-order Lagrangian Perturbation Theory) because, while agreeing with the SPT to second order, the 2LPT density field also contains higher-order contributions. We take advantage of the efficiency of both GridSPT and 2LPT, which enable us to generate a large number of realizations to suppress the sampling variance and find any underlying systematic errors in the estimator.

By testing the estimator on well-controlled datasets, we find that the clustering-fossil estimator is unbiased only when an accurate bispectrum model is used for the coupling between long and short modes. Furthermore, to reconstruct the long-mode power spectrum, the connected four-point correlation function must be included to compute the noise power spectrum. It is important to note that the goal of this paper is to assess the potential systematics of the quadratic estimator, so we restrict our study to matter clustering in real space for clarity and transparency. To apply this method to galaxy surveys or 21cm intensity mapping experiments, we need to include the galaxy bias and redshift space distortions into account(see e.g. [43, 52]), which would make the accurate reconstruction even harder.

The rest of this paper is organized as follows. In section 2, we review the clustering fossil technique and construct the optimal estimator of the large-scale matter density modes.

JCAP07(2024)020

In section 3, we introduce GridSPT and 2LPT we use to generate the nonlinear short-wavelength modes and implement the fossil estimator to reconstruct the large-scale mode. We compare the results with the theoretical prediction in section 4. We conclude and discuss the future applications in section 5. We provide the fossil estimator in the continuous limit in appendix A.

Throughout this paper, we use the following convention of Fourier transformation,

$$f(\boldsymbol{k}) = \int \mathrm{d}^3x f(\boldsymbol{x}) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{1.1}$$

$$f(\boldsymbol{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3} f(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{1.2}$$

Note that we use the same character for a function f and distinguish the Fourier-space representation and configuration-space representation by the argument. For the compactness of the equations, we use the following convention for the sum of multiple vectors,

$$\boldsymbol{k}_{1\cdots n} = \sum_{i=1}^{n} \boldsymbol{k}_i. \tag{1.3}$$

2 Estimating long modes from the clustering-fossil estimator

We begin the section by motivating the clustering-fossil method [46] for a generic non-Gaussian density field in section 2.1. In section 2.2, we construct the optimal quadratic estimator for measuring the long mode from the imprint of the squeeze-limit bispectrum, or long-short-short coupling. We also present the power spectrum of the reconstructed long modes and the cross-correlation coefficient between the original and the reconstructed long modes. Finally, we present the reconstructed long modes' covariance matrix, or noise power spectrum, in section 2.3.

2.1 Position-dependent correlations induced by squeezed-limit non-Gaussianity

The standard cosmological model based on the Friedman-Lemaître-Robertson-Walker world model is spatially homogeneous and isotropic. The spatial homogeneity extends to the clustering properties of galaxies, which is often referred to as statistical homogeneity. We begin this section by defining the statistical homogeneity through n-point correlation functions.

The fundamental quantity describing the galaxy clustering is the density contrast $\delta(\boldsymbol{x}) \equiv n(\boldsymbol{x})/\bar{n} - 1$, where $n(\boldsymbol{x})$ is the density at location $\boldsymbol{x}$ and $\bar{n}$ is the cosmic mean density. The two-point correlation function is defined in terms of the density contrast as

$$\xi(\boldsymbol{r}) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r}) \rangle \ . \tag{2.1}$$

Here, the bracket $\langle \ldots \rangle$ represents the average over the statistical ensemble of the cosmic density field. Note that $\xi(\boldsymbol{r})$ depends only on the separation vector $\boldsymbol{r}$ between two positions $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{r}$ but is independent of the position $\boldsymbol{x}$ where we make the measurement. This property manifests the statistical homogeneity. Furthermore, the statistical isotropy would restrict $\xi(\boldsymbol{r}) = \xi(r)$. Equivalently, the power spectrum, the Fourier transformation of the two-point correlation function, is defined as

$$\langle \delta(\boldsymbol{k}_1)\delta(\boldsymbol{k}_2) \rangle = (2\pi)^3 \delta^D(\boldsymbol{k}_{12}) P(\boldsymbol{k}_1), \tag{2.2}$$

JCAP07(2024)020

where the Dirac-delta operator, $\delta^D(\boldsymbol{k}_{12})$ on the right-hand side, encodes the statistical homogeneity. That is, the statistical homogeneity dictates that the correlation between the Fourier modes $\delta(\boldsymbol{k}_1)$ and $\delta(\boldsymbol{k}_2)$ vanishes unless $\boldsymbol{k}_1 + \boldsymbol{k}_2 = 0$. We call such two-point correlators in Fourier space *diagonal*.

We can extend this concept and write the n-point correlation function and the n-poly spectra for statistically homogeneous density contrast δ as follows:

$$\zeta^{(n)}(\boldsymbol{r}, \boldsymbol{s}, \cdots) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r})\delta(\boldsymbol{x}+\boldsymbol{s})\cdots \rangle \tag{2.3}$$

$$\langle \delta(\boldsymbol{k}_1)\cdots\delta(\boldsymbol{k}_n)\rangle = (2\pi)^3\delta^D(\boldsymbol{k}_{12\cdots n})P^{(n)}(\boldsymbol{k}_1, \cdots, \boldsymbol{k}_n)\,. \tag{2.4}$$

We call a density field Gaussian when the two-point correlation function is the only non-vanishing connected n-point correlation function [56]; otherwise, the density field is non-Gaussian.

The clustering-fossil estimator exploits the fact that non-Gaussianities in density fields can yield a spatial variation of correlation functions. That is, even if the density field satisfies statistical homogeneity when taking the ensemble average as in equation (2.3), the locally measured n-point correlation function can have an explicit position dependence through the non-Gaussian correlation. This is because the non-Gaussian correlation function involving one or more long-wavelength modes (δ_ℓ) can be broken down as the multiplication of the conditional correlation functions and the conditioning long modes:

$$\langle \delta_\ell(\boldsymbol{x}_1)\cdots\delta_\ell(\boldsymbol{x}_n)\delta(\boldsymbol{r})\delta(\boldsymbol{s})\cdots\rangle = \left\langle \langle \delta(\boldsymbol{r})\delta(\boldsymbol{s})\cdots\rangle|_{\delta_\ell(\boldsymbol{x}'_1)\cdots\delta_\ell(\boldsymbol{x}'_n)}\,\delta_\ell(\boldsymbol{x}_1)\cdots\delta_\ell(\boldsymbol{x}_n)\right\rangle\,, \tag{2.5}$$

where $\langle\rangle|_X$ stands for the correlator evaluated with the condition X. The expression follows from the definition of conditional probability $P(A\cap B) = P(A|B)P(B)$.

The most well-studied example of clustering fossil comes from the squeezed (or soft) limit of non-Gaussian correlation functions. Namely, the squeezed-limit $(n+1)$-point correlation function can modulate the n-point correlation function, inducing the position-dependent power spectrum [57, 58] or the position-dependent bispectrum [59]. In this case, we may write the conditional correlation function as

$$\langle \delta(\boldsymbol{r})\delta(\boldsymbol{s})\cdots\rangle|_{\delta_\ell(\boldsymbol{x}')} = C + \frac{\mathrm{d}}{\mathrm{d}\delta_\ell}\langle \delta(\boldsymbol{r})\delta(\boldsymbol{s})\cdots\rangle\Big|_{\delta_\ell=0}\delta_\ell(\boldsymbol{x}') + \cdots\,, \tag{2.6}$$

where C is a function independent from δ_ℓ, and we truncate the expansion at the linear-order response. On large scales where the long mode δ_ℓ is in the linear regime, we compute the original correlation function as

$$\langle \delta_\ell(\boldsymbol{x})\delta(\boldsymbol{r})\delta(\boldsymbol{s})\cdots\rangle \simeq \frac{\mathrm{d}}{\mathrm{d}\delta_\ell}\langle \delta(\boldsymbol{r})\delta(\boldsymbol{s})\cdots\rangle\Big|_{\delta_\ell=0}\xi_\ell(\boldsymbol{x}-\boldsymbol{x}')\,. \tag{2.7}$$

The general idea of the clustering fossil estimator [46] is to use equation (2.6) as a starting point for estimating the long-wavelength mode δ_ℓ for the non-Gaussian density field with underlying $(n+1)$-point correlation function taking the form of equation (2.7).

As shown in [46], the same argument applies when using a long mode h^s_ℓ with general spin s. The density field is a spin-0 case.

JCAP07(2024)020

2.2 The optimal clustering fossil estimator

Let us focus on the quadratic clustering-fossil estimator using the squeezed-limit three-point correlation function, or its Fourier transform, bispectrum. The Fourier-space counterpart of equation (2.7), when considering the three-point correlator, becomes

$$\langle \delta_\ell(\boldsymbol{k})\delta(\boldsymbol{k}_1)\delta(\boldsymbol{k}_2)\rangle = (2\pi)^3\delta^D(\boldsymbol{k}_{12}+\boldsymbol{k})f(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k})P_\ell(k), \tag{2.8}$$

with kernel $f(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k})$ encoding the details of the coupling between the long mode $\delta_\ell(\boldsymbol{k})$ and two short modes $\delta(\boldsymbol{k}_1)$ and $\delta(\boldsymbol{k}_2)$. Hereafter, we drop the explicit subscript ℓ and implicitly use $\boldsymbol{k} \ll \boldsymbol{k}_i$ to indicate the long mode. We emphasize that the three wave vectors in $f(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k})$ must satisfy $\boldsymbol{k}_{12}+\boldsymbol{k}=0$.

As we have shown in section 2.1, equation (2.8) is equivalent to the following conditional two-point correlator:

$$\langle \delta(\boldsymbol{k}_1)\delta(\boldsymbol{k}_2)\rangle\,|_{\delta(\boldsymbol{k})} = f(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k})\delta^*(\boldsymbol{k})\delta^{\rm K}_{\boldsymbol{k}_{12},-\boldsymbol{k}}. \tag{2.9}$$

Here, $\delta^{\rm K}$ refers to the Kronecker delta that we use instead of Dirac delta, after absorbing appropriate dimensionful constants inside $f(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k})$, and the superscript $*$ is the complex conjugate. The locally broken statistical homogeneity is now manifested by the off-diagonal correlator in equation (2.9).

Note that equation (2.8), more generally equation (2.7), assumes that the long mode is in the linear regime where we can neglect the higher order terms in equation (2.7); so the fossil estimator is unbiased only in that limit. A generic unbiased fossil estimator would require additional corrections to include higher-order effects. In section 4.3.2 of this paper, however, we show that the nonlinear part of the long-mode power spectrum is less than a few percent of the cosmic variance, which ensures that the nonlinear correction is unimportant at the scale within with we implement the estimator.

Let us translate the fossil equation, equation (2.9), to the quadratic estimator for the long mode $\delta(\boldsymbol{k})$. A simple estimator turning equation (2.9) around would be

$$\hat{\delta}_r^{\rm naive}(\boldsymbol{k}) = \left[\sum_{-\boldsymbol{q}_i}\frac{\delta(-\boldsymbol{q}_i)\delta(-\boldsymbol{k}+\boldsymbol{q}_i)}{f(-\boldsymbol{q}_i,-\boldsymbol{k}+\boldsymbol{q}_i;\boldsymbol{k})}\right]^* = \sum_{\boldsymbol{q}_i}\frac{\delta(\boldsymbol{q}_i)\delta(\boldsymbol{k}-\boldsymbol{q}_i)}{f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})}\,. \tag{2.10}$$

In the second equality, we use the reality of the density field $[\delta(\boldsymbol{k})]^* = \delta(-\boldsymbol{k})$ and that bispectra and $f(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k})$ must be real for even-parity density field: $f^*(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k}) = f(-\boldsymbol{k}_1,-\boldsymbol{k}_2;-\boldsymbol{k})$ [60]. For a given $\boldsymbol{k}$, this naive estimator sums over all possible pairs $\delta(\boldsymbol{q}_i)$ and $\delta(\boldsymbol{k}-\boldsymbol{q}_i)$ weighted by a known kernel function $f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})$ determined by the bispectrum as in equation (2.8). In order to exclude the duplication of counting the same quadratic contribution twice, we demand the condition $k < |\boldsymbol{k}-\boldsymbol{q}_i| \le q_i$. Since the estimator is based on the multiplication of two short modes, it is called a quadratic estimator. A similar quadratic estimator has already been used widely in the CMB lensing reconstruction [61, 62].

Noticing that the individual contribution in equation (2.10) is subject to different variances, we can introduce a normalized weight W_i

$$\hat{\delta}_r(\boldsymbol{k}) = \sum_i W_i\frac{\delta(\boldsymbol{q}_i)\delta(\boldsymbol{k}-\boldsymbol{q}_i)}{f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})}, \tag{2.11}$$

JCAP07(2024)020

to find an optimal estimator which minimizes the variance of $\hat{\delta}_r(\boldsymbol{k})$:

$$\sigma^2\left[\hat{\delta}_r(\boldsymbol{k})\right] = \sum_{ij} W_i W_j C_{ij}. \tag{2.12}$$

First, we compute the covariance matrix C_{ij} for individual contribution in the naive estimator [equation (2.10)]:

$$C_{ij}(\boldsymbol{q}_i, \boldsymbol{q}_j; \boldsymbol{k}) = V^2 \frac{P(q_i)P(|\boldsymbol{k}-\boldsymbol{q}_i|)}{f^2(\boldsymbol{q}_i, \boldsymbol{k}-\boldsymbol{q}_i; -\boldsymbol{k})} \delta^{\mathrm{K}}_{\boldsymbol{q}_j, -\boldsymbol{q}_i}, \tag{2.13}$$

by ignoring all the connected parts of four-point correlators. Here $P(q)$ and $P(|\boldsymbol{k}-\boldsymbol{q}|)$ are the measured nonlinear power spectra of the short modes. In this case, the inverse-variance weight

$$W_i = \sum_j C_{ij}^{-1} / \sum_{ij} C_{ij}^{-1},\ \sigma^2\left[\hat{\delta}_r(\boldsymbol{k})\right] = 1/\sum_j C_{ij}^{-1}. \tag{2.14}$$

gives the minimum variance estimator. Using the inverse-variance weight in equation (2.14), the optimal fossil estimator for the long mode becomes

$$\hat{\delta}_r(\boldsymbol{k}) = P_G^N(k) \sum_{\boldsymbol{q}} \frac{f(\boldsymbol{q}, \boldsymbol{k}-\boldsymbol{q}; -\boldsymbol{k})}{VP(\boldsymbol{q})P(\boldsymbol{k}-\boldsymbol{q})} \delta(\boldsymbol{q})\delta(\boldsymbol{k}-\boldsymbol{q}), \tag{2.15}$$

the variance of the optimal estimator $\hat{\delta}_r$ becomes

$$P_G^N(k) = \left[\sum_{\boldsymbol{q}} \frac{|f(\boldsymbol{q}, \boldsymbol{k}-\boldsymbol{q}; -\boldsymbol{k})|^2}{VP(\boldsymbol{q})P(\boldsymbol{k}-\boldsymbol{q})}\right]^{-1}. \tag{2.16}$$

The minimum variance $P_G^N(k)$ turns out to be the noise term of the recovered long mode's auto power spectrum, when taking only the Gaussian (disconnected) four-point correlator of the density contrast δ:

$$P_{rr}(k) \equiv \frac{1}{V}\frac{1}{N_k} \sum_{\hat{\boldsymbol{k}}} \left\langle |\hat{\delta}_r(\boldsymbol{k})|^2 \right\rangle = P_L(k) + P_G^N(k)\,. \tag{2.17}$$

Here, N_k is the number of Fourier modes for a fixed k. Equation (2.17) suggests that we have to subtract the noise power spectrum $P_G^N(k)$ from the auto power spectrum, in order to recover an unbiased long-mode power spectrum.

Let us turn into the cross-correlation coefficient $r(k)$ between the recovered long mode $\hat{\delta}_r(\boldsymbol{k})$ and the original one $\delta(\boldsymbol{k})$ that we designate with subscript m:

$$r(k) \equiv \frac{P_{rm}(k)}{\sqrt{P_{rr}(k)P_{mm}(k)}} \tag{2.18}$$

Here, $\left\langle \hat{\delta}_r(\boldsymbol{k})\delta_m(\boldsymbol{k}')\right\rangle = (2\pi)^3 P_{rm}(k)\delta^D(\boldsymbol{k}+\boldsymbol{k}')$ is the cross power spectrum

$$P_{rm}(k) \equiv \frac{1}{V}\frac{1}{N_k} \sum_{\hat{\boldsymbol{k}}} \mathrm{Re}\left\langle \hat{\delta}_r(\boldsymbol{k})\delta^*(\boldsymbol{k})\right\rangle = \sum_{\boldsymbol{q}_i} W_i \frac{B(\boldsymbol{q}_i, \boldsymbol{k}-\boldsymbol{q}_i, -\boldsymbol{k})}{f(\boldsymbol{q}_i, \boldsymbol{k}-\boldsymbol{q}_i; -\boldsymbol{k})} \simeq P_L(k)\,, \tag{2.19}$$

which approaches the linear power spectrum for the squeezed ($k \to 0$) limit. In section 3.1, we calculate the correction term to this approximation when using the full bispectrum

JCAP07(2024)020

expression. The leading contribution of the correction term comes from the correlation between the second-order of the long mode and two first-order short modes, which vanishes in the squeezed limit.

Combining all results, we calculate the leading-order expression for the cross-correlation coefficients in the squeezed limit [$k \to 0$ and $P_{mm} \to P_L(k)$] as,

$$r(k) \simeq \frac{1}{\sqrt{1 + P_G^N(k)/P_L(k)}}. \tag{2.20}$$

From equation (2.20), we see that $r(k) \to 1$ if and only if when the noise power spectrum is much suppressed compared to the linear power spectrum. The noise power spectrum [equation (2.16)] becomes smaller as we include more and more short modes in the quadratic estimator. Meanwhile, the noise power spectrum is insensitive to the lower limit of q and $|\boldsymbol{k} - \boldsymbol{q}|$ because most reconstruction power comes from the high-frequency ($|\boldsymbol{q}| \gg k$) modes. In this paper, we use the short modes $\delta(\boldsymbol{q})$, $\delta(\boldsymbol{k} - \boldsymbol{q})$ with wavenumbers satisfying

$$k < |\boldsymbol{k} - \boldsymbol{q}| \leq q \leq q_{\rm max}, \tag{2.21}$$

where $q_{\rm max}$ is the wavenumber of the smallest scale used for reconstruction. Ideally, we want to make $q_{\rm max}$ as large as possible to suppress the noise P_G^N. In practice, however, we are limited by strong nonlinearities on small scales, so we can only keep $q_{\rm max}$ in the quasi-linear scale where perturbation theory accurately models the nonlinearities [63, 64]. As we shall show below, using the short modes beyond the quasi-linear regime leads to the systematic bias in the reconstructed long modes.

As stated earlier, the summations in the equations above excludes the duplications (that is, $\boldsymbol{q}'=\boldsymbol{k}-\boldsymbol{q}$ is identical to $\boldsymbol{q}$), which leads to a factor of two difference between equation (2.16) here and the expressions shown in previous studies [46, 49, 50, 52]. The restriction in the summation, however, cancels the factor-of-two difference, so the expressions are identical.

However, just like our derivation in this section, most previous studies to date have ignored the connected four-point correlation function in deriving equation (2.13), so the noise power spectrum is underestimated (see figure 3 below). We now present the full noise power spectrum, including the connected four-point function. Readers with interest can find related discussions of connected four-point correlation function in the noise power spectrum of quadratic estimators in [45, 52] for long-mode reconstruction and [65] for CIB lensing.

2.3 The full noise power spectrum

In section 2.2, we use the Gaussian approximation for when calculating the covariance matrix of naive quadratic terms estimators in equation (2.10). In this section, we supplement that calculation by deriving the full covariance matrix, including the connected four-point function. We also present the improved noise power spectrum by using the full covariance matrix but still adopting the inverse-variant weight using the Gaussian approximation.

JCAP07(2024)020

We first compute the full covariance matrix of each term contributing to the naive estimator equation (2.10) as

$$\begin{aligned} C_{ij}(\boldsymbol{q}_i,\boldsymbol{q}_j;\boldsymbol{k}) &= \left\langle \frac{\delta(\boldsymbol{q}_i)\delta(\boldsymbol{k}-\boldsymbol{q}_i)}{f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})}\frac{\delta(\boldsymbol{q}_j)\delta(-\boldsymbol{k}-\boldsymbol{q}_j)}{f(\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j;\boldsymbol{k})}\right\rangle\Big|_{\delta(\boldsymbol{k}),\delta(-\boldsymbol{k})} \\ &\quad - \left\langle \frac{\delta(\boldsymbol{q}_i)\delta(\boldsymbol{k}-\boldsymbol{q}_i)}{f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})}\right\rangle\Big|_{\delta(\boldsymbol{k})}\left\langle \frac{\delta(\boldsymbol{q}_j)\delta(-\boldsymbol{k}-\boldsymbol{q}_j)}{f(\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j;\boldsymbol{k})}\right\rangle\Big|_{\delta(-\boldsymbol{k})} \\ &= V^2\frac{P(q_i)P(|\boldsymbol{k}-\boldsymbol{q}_i|)}{|f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})|^2}\left(\delta^{\rm K}_{\boldsymbol{q}_j,-\boldsymbol{q}_i}+\delta^{\rm K}_{\boldsymbol{q}_j,-(\boldsymbol{k}-\boldsymbol{q}_i)}\right) \\ &\quad + V\frac{T\left(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i,\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j\right)}{f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})f(\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j;\boldsymbol{k})}\,, \end{aligned} \tag{2.22}$$

where $T\left(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i,\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j\right)$ is the connected four-point function, or trispectrum, defined as follows:

$$\begin{aligned} T\left(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i,\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j\right) &= \langle\delta(\boldsymbol{q}_i)\delta(\boldsymbol{k}-\boldsymbol{q}_i)\delta(\boldsymbol{q}_j)\delta(-\boldsymbol{k}-\boldsymbol{q}_j)\rangle \\ &\quad - \langle\delta(\boldsymbol{q}_i)\delta(\boldsymbol{k}-\boldsymbol{q}_i)\rangle\langle\delta(\boldsymbol{q}_j)\delta(-\boldsymbol{k}-\boldsymbol{q}_j)\rangle \\ &\quad - \langle\delta(\boldsymbol{q}_i)\delta(\boldsymbol{q}_j)\rangle\langle\delta(\boldsymbol{k}-\boldsymbol{q}_i)\delta(-\boldsymbol{k}-\boldsymbol{q}_j)\rangle \\ &\quad - \langle\delta(\boldsymbol{q}_i)\delta(-\boldsymbol{k}-\boldsymbol{q}_j)\rangle\langle\delta(\boldsymbol{k}-\boldsymbol{q}_i)\delta(\boldsymbol{q}_j)\rangle\,. \end{aligned} \tag{2.23}$$

After removing the duplicated contributions by imposing the inequality in equation (2.21), the covariance matrix becomes

$$C_{ij}(\boldsymbol{q}_i,\boldsymbol{q}_j;\boldsymbol{k}) = V^2\frac{P(q_i)P(|\boldsymbol{k}-\boldsymbol{q}_i|)}{|f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})|^2}\delta^{\rm K}_{\boldsymbol{q}_j,-\boldsymbol{q}_i} + V\frac{T\left(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i,\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j\right)}{f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i;-\boldsymbol{k})f(\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j;\boldsymbol{k})}\,. \tag{2.24}$$

From the covariance matrix, we compute the full noise power spectrum for the optimal quadratic estimator, equation (2.15):

$$P^N_{\rm full}(k) = P^N_G(k) + \left\{\left[P^N_G(k)\right]^2\sum_{ij}\frac{T\left(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i,\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j\right)f(\boldsymbol{q}_i,\boldsymbol{k}-\boldsymbol{q}_i)f(\boldsymbol{q}_j,-\boldsymbol{k}-\boldsymbol{q}_j)}{V^2P(q_i)P(q_j)P(|\boldsymbol{k}-\boldsymbol{q}_i|)P(|-\boldsymbol{k}-\boldsymbol{q}_j|)}\right\}. \tag{2.25}$$

Note that the estimator equation (2.15) is optimal under the Gaussian assumption but may not remain optimal when the trispectrum contributes significantly to the covariance matrix. Since the relative contribution from the trispectrum rises toward small scales [66, 67], we expect that the noise power spectrum $P^N_{\rm full}(k)$ would saturate beyond some $q_{\rm max}$. Later, we verify the statements by comparing numerically the two noise power spectra $P^N_{\rm full}(k)$ and $P^N_G(k)$ for the two nonlinear density realizations: 2LPT [68] and GridSPT [53].

3 Leading-order clustering-fossil estimator and nonlinear density fields

The main goal of this paper is to assess the requirement for an accurate and unbiased clustering-fossil estimator in equation (2.15). As for the simplest but realistic nonlinear density field, as shown in section 3.1, we study the clustering-fossil estimator using the second-order density perturbations as defined in the SPT (Standard Perturbation Theory) [51]. Specifically,

JCAP07(2024)020

we apply the quadratic clustering-fossil estimator to measure the long modes from the squeezed-limit bispectrum that encodes an off-diagonal power spectrum of the short modes.

Our approach is different from the previous studies [49, 50], which tested the quadratic estimator based upon the second-order SPT (equation (3.4) below) against a few N-body simulation results, or tested the second-order EFT against dark matter halo fields [52]. In these studies, they have drawn affirmative conclusions about the possibility of reconstructing large-scale modes. At the same time, the quadratic estimator must fail beyond some $q_{\rm max}$, because the squeezed-limit bispectrum deviates from the tree-level SPT or EFT prediction. Also, the noise power spectrum of the reconstructed lone mode must be underestimated because the trispectrum contribution becomes important on small scales. Neglecting the trispectrum contribution would bias the power spectrum of reconstructed long modes. We cannot distinguish these two effects in N-body simulations. Also, we need multiple realizations to draw a robust statistical conclusion, but N-body simulations are too expensive for this purpose.

In contrast, our method based on the SPT allows a more systematic and in-depth study of the method because we can generate nonlinear density fields with a well-controlled nonlinear order. In this paper, we adopt a novel grid-based standard perturbation theory (GridSPT) method to generate the density fields up to second-order (section 3.2.1). We then study the higher-order contribution's effect on the fossil estimator by comparing the estimated long mode from GridSPT realizations with the result from the second-order Lagrangian Perturbation Theory (2LPT) realizations (section 3.2.2) and fourth-order GridSPT (section 3.2.3). Both GridSPT and 2LPT are fast enough to generate each realization in less than one minute.

We have performed the comparison study at two redshifts ($z = 1$ and $z = 0$). For all studies, we use the simulation box of $V = (1000\ {\rm Mpc}/h)^3$, and we adopt the following cosmological parameters [24]: $\Omega_{b0}h^2 = 0.022307$, $\Omega_{c0}h^2 = 0.11865$, $h = 0.6778$, $n_s = 0.9672$, $\Omega_{\nu 0}h^2 = 0.000638$, $\Omega_{\Lambda 0} = 0.69179$, $\mathcal{A}_s = 2.147 \times 10^{-9}$ and $\sigma_8 = 0.8166$. We use CAMB [69] to generate the input linear power spectrum of GridSPT and 2LPT.

3.1 The leading-order matter bispectrum from standard perturbation theory

Constructing the quadratic clustering-fossil estimator requires the squeeze-limit bispectrum to set the fossil kernel $f(\boldsymbol{k}_1, \boldsymbol{k}_2; \boldsymbol{k})$ in equation (2.15). In SPT, the leading-order matter bispectrum from nonlinear gravity is given as

$$B(k_1, k_2, k_3) = 2F_2(\boldsymbol{k}_1, \boldsymbol{k}_2)P_L(k_1)P_L(k_2) + (2\ \text{cyclic})\,, \tag{3.1}$$

with the second-order SPT kernel

$$F_2(\boldsymbol{k}_1, \boldsymbol{k}_2) = \frac{5}{7} + \frac{2}{7}\left(\frac{\boldsymbol{k}_1 \cdot \boldsymbol{k}_2}{k_1 k_2}\right)^2 + \frac{1}{2}\frac{\boldsymbol{k}_1 \cdot \boldsymbol{k}_2}{k_1 k_2}\left(\frac{k_1}{k_2} + \frac{k_2}{k_1}\right) \tag{3.2}$$

and the linear matter power spectrum $P_L(k)$. Note that the kernel $F_2(\boldsymbol{k}_1, \boldsymbol{k}_2)$ only depends on the three wavenumbers (k_1, k_2, k_3) because of the statistical homogeneity and isotropy.

The clustering-fossil estimator facilitates the squeezed limit of the bispectrum:

$$B(k, k_1, k_2) \xrightarrow{k\to 0} \left[2F_2(\boldsymbol{k}_1, \boldsymbol{k})P_L(k_1) + 2F_2(\boldsymbol{k}_2, \boldsymbol{k})P_L(k_2)\right] P_L(k)\,, \tag{3.3}$$

and we can read off the following fossil-kernel expression from equation (2.8) as:

$$f(\boldsymbol{k}_1, \boldsymbol{k}_2; \boldsymbol{k}) = 2F_2(\boldsymbol{k}_1, \boldsymbol{k})P_L(k_1) + 2F_2(\boldsymbol{k}_2, \boldsymbol{k})P_L(k_2). \tag{3.4}$$

JCAP07(2024)020

Note that the third term $2F_2(\boldsymbol{k}_1, \boldsymbol{k}_2)P_L(k_1)P_L(k_2)$ vanishes in the squeezed limit because $F(\boldsymbol{q}, -\boldsymbol{q}) = 0$. For a finite $\boldsymbol{k}$, this term indeed contributes to the statistics of the recovered long mode. For example, when computing the cross power spectrum between the recovered long modes (using equation (2.11)) and the original one, the term yields the correction to the squeezed-limit expression in equation (2.19) as

$$\begin{aligned} P_{rm}(k) &= \sum_i W_i \frac{B(\boldsymbol{q}_i, \boldsymbol{k} - \boldsymbol{q}_i, -\boldsymbol{k})}{f(\boldsymbol{q}_i, \boldsymbol{k} - \boldsymbol{q}_i; -\boldsymbol{k})} \\ &= P_L(k) + 2P^N(k) \sum_i \frac{F_2(\boldsymbol{q}_i, \boldsymbol{k} - \boldsymbol{q}_i)P_L(q_i)P_L(|\boldsymbol{k} - \boldsymbol{q}_i|)}{f(\boldsymbol{q}_i, \boldsymbol{k} - \boldsymbol{q}_i; -\boldsymbol{k})} \,. \end{aligned} \tag{3.5}$$

3.2 Nonlinear density fields

3.2.1 Second-order GridSPT

The Grid-based standard perturbation theory (GridSPT) is a method of computing the n-th order SPT density and velocity fields from the recursion relation. Standard Perturbation Theory [63, 70–76] approximates the evolution as a pressure-less fluid with the following evolution equations for the density contrast $\delta(\boldsymbol{x}, \tau) \equiv \rho_m(\boldsymbol{x}, \tau)/\bar{\rho}_m(\tau) - 1$ and the peculiar velocity $\boldsymbol{v}(\boldsymbol{x}, \tau)$:

$$\dot{\delta} + \nabla \cdot [(1 + \delta)\boldsymbol{v}] = 0 \,, \tag{3.6}$$

$$\dot{\boldsymbol{v}} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} + \frac{\dot{a}}{a}\boldsymbol{v} = -\nabla\Phi \,, \tag{3.7}$$

along with the Poisson equation:

$$\nabla^2\Phi = 4\pi G\bar{\rho}_{\rm m}\, a^2\delta \,. \tag{3.8}$$

Here, dot represents the conformal-time derivative, $d\tau = dt/a$ with $a(t)$ being the scale factor and t being the cosmic time, ∇ is comoving-coordinate derivative, $\bar{\rho}_{\rm m}$ is the mean matter density, and Φ is the peculiar gravitational potential. Note that we omit the spacetime coordinate in equations to avoid clutter.

The set of equations describes the non-relativistic-matter (cold-dark matter and baryon) fluid on scales larger than the baryonic Jeans scale. Note that the nonlinearities in equations (3.6)–(3.7) comes from the second-order terms such as $\nabla \cdot (\delta\boldsymbol{v})$ and $(\boldsymbol{v} \cdot \nabla)\boldsymbol{v}$. In SPT, we solve equations (3.6)–(3.8) by expanding the nonlinear density contrast δ and velocity-gradient field $\theta \equiv -\nabla \cdot \boldsymbol{v}/(aHf)$ as

$$\delta(\boldsymbol{x}, \tau) = \sum_n [D(\tau)]^n \delta^{(n)}(\boldsymbol{x}) \,, \quad \theta(\boldsymbol{x}, \tau) = \sum_n [D(\tau)]^n \theta^{(n)}(\boldsymbol{x}) \,, \tag{3.9}$$

where $D(\tau)$ is the linear growth factor.

For a given realization of the linear density field on regular grid points, the GridSPT [53] provides a way to compute the matter density field δ and the velocity field $\boldsymbol{v}$ of LSS perturbatively by solving the fluid equations [equations (3.6)–(3.8)], which becomes the recursion relation as:

$$\begin{pmatrix} \delta^{(n)}(\boldsymbol{x}) \\ \theta^{(n)}(\boldsymbol{x}) \end{pmatrix} = \frac{1}{(2n+3)(n-1)} \begin{pmatrix} 2n+1 & 1 \\ 3 & n \end{pmatrix} \sum_{m=1}^{n-1} \begin{pmatrix} \nabla \cdot \left(\delta^{(m)}\boldsymbol{v}^{(n-m)}\right) \\ \nabla^2 \left(\boldsymbol{v}^{(n-m)} \cdot \boldsymbol{v}^{(m)}\right) \end{pmatrix} . \tag{3.10}$$

JCAP07(2024)020

From a given set of linear density field, $\delta_1 = \theta_1 = \delta_L$, we can use equation (3.10) to compute the nonlinear density $\delta^{(n)}$ and velocity-gradient $\theta^{(n)}$ fields order by order manner. Using the FFT (Fast Fourier Transform) to evaluate the ∇ operators on the right-hand side, the GridSPT enables us to quickly generate the nth order quantities.

We use the second-order GridSPT, which contains first- and second-order density contrast, to test the fossil estimator. The squeezed bispectrum in the second-order GridSPT strictly follows the tree-level bispectrum in SPT so the constructed estimator can fully capture the coupling between long and short modes. Therefore, the second-order GridSPT is an ideal tool to conduct a proof-of-concept study of the fossil estimator without any uncontrolled systematic error.

We compute the GridSPT density fields on the $N_{\rm grid} = 512^3$ grids. We adopt the empirical cutoff $k_1 = 1.0\,h/\mathrm{Mpc}$ for the first-order density contrast and $k_2 = 1.33\,h/\mathrm{Mpc}$ for higher-order density contrasts used in [53].

3.2.2 Second-order Lagrangian perturbation theory (2LPT)

Lagrangian perturbation theory (LPT) is an alternative to the SPT in modeling the nonlinear density fields. The fundamental object in LPT is the displacement field $\boldsymbol{\Psi}(\boldsymbol{q}, \tau)$ from the regular Lagrangian position $\boldsymbol{q}$, which makes the Eulerian position $\boldsymbol{x}$ as

$$\boldsymbol{x}(\boldsymbol{q}, \tau) = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}, \tau)\,. \tag{3.11}$$

We can obtain the LPT solutions by solving the equation of motion in an expanding universe

$$\ddot{\boldsymbol{x}} + \frac{\dot{a}}{a}\dot{\boldsymbol{x}} = -\nabla_{\boldsymbol{x}}\Phi \tag{3.12}$$

for an irrotational displacement field $\nabla \times \boldsymbol{\Psi} = 0$ perturbatively. The second-order solution for the displacement field is given as (see, for example, appendix E of [56] for a full derivation)

$$\boldsymbol{\Psi}(\boldsymbol{q}, \tau) = -\nabla_{\boldsymbol{q}}\phi^{(1)}(\boldsymbol{q}, \tau) + \nabla_{\boldsymbol{q}}\phi^{(2)}(\boldsymbol{q}, \tau)\,, \tag{3.13}$$

where the linear Lagrangian potential is related to the linear density contrast as

$$\nabla^2_{\boldsymbol{q}}\phi^{(1)}(\boldsymbol{q}, \tau) = \delta^{(1)}(\boldsymbol{x}, \tau)\,, \tag{3.14}$$

and the second-order Lagrangian potential is related to $\phi^{(1)}$ as

$$\nabla^2_{\boldsymbol{q}}\phi^{(2)}(\boldsymbol{q}, \tau) = -\frac{D_2(\tau)}{D(\tau)^2}\sum_{i>j}\left\{\phi^{(1)}_{,ii}(\boldsymbol{q}, \tau)\phi^{(1)}_{,jj}(\boldsymbol{q}, \tau) - \left[\phi^{(1)}_{,ij}(\boldsymbol{q}, \tau)\right]^2\right\}. \tag{3.15}$$

Here, $D_2(\tau)$ is the solution of the following differential equation:

$$\ddot{D}_2(\tau) + \frac{\dot{a}}{a}\dot{D}_2(\tau) - \frac{3}{2}\left(\frac{\dot{a}}{a}\right)^2\left[\Omega_m(\tau)D_2(\tau) - D^2(\tau)\right] = 0\,, \tag{3.16}$$

and $D_2 = -3/7D^2(\tau)$ for the Einstein de-Sitter (spatially flat and matter-dominated) universe.

The 2LPT (seond-order LPT) prescription is to displace regularly spaced particles using equation (3.11) and equation (3.13). Since the nonlinearities are modeled by particle

JCAP07(2024)020

displacement, the resulting density contrast, while agreeing with the SPT prediction to second order, contains a myriad of higher-order nonlinear contributions (see, e.g., refs. [68] and [77]). The bispectrum of the 2LPT density field, therefore, deviates from the tree-level predictions in equation (3.3), particularly on small scales. By applying the same fossil estimator using the kernel given in equation (3.4) to the 2LPT density fields, we can test the behavior of the fossil estimator when ignoring the higher-order nonlinear couplings in data.

We implement the 2LPT by using the displacement of 512^3 Lagrangian particles in the cubic box of volume $(1000\ \mathrm{Mpc}/h)^3$. We then estimate the density with $N_{\rm grid} = 256^3$ grids and preserve the density modes of wavenumber $k < k_{\rm Nyq}/2$ to avoid the aliasing effect [56].

3.2.3 Fourth-order GridSPT

While we test the effect of neglecting higher-order nonlinear couplings by comparing the second-order GridSPT and 2LPT, the density fields in these two toy cases are far from reality on small scales. For example, the nonlinearity in the density field is too strong in second-order GridSPT [78] while too weak in 2LPT [53] compared to N-body simulations. To overcome this, we also test the performance of the tree-level fossil estimator in a more realistic density field using the fourth-order GridSPT. The power spectrum and bispectrum in the fourth-order GridSPT can be accurately modeled by the complete one-loop power spectrum and one-loop bispectrum in SPT, which are closer to the N-body result than either second-order GridSPT or 2LPT [78]. As for the $q_{\rm max}(z)$, the smallest scale to be included in the reconstruction at each redshift, we use the result of ref. [64], which has measured the maximum wavenumber below which the tree-level bispectrum accurately model the nonlinear bispectrum from a suite of N-body simulations.

4 Results

We present the results of the analysis in reconstructing the long modes by applying the tree-level fossil estimator to the following three nonlinear density fields: second-order and fourth-order GridSPT and 2LPT.

4.1 The tree-level fossil estimator on the 2LPT density field

First, we applied the tree-level fossil estimator to the 2LPT density fields to reconstruct the long modes.

4.1.1 The reconstructed long modes vs. the original modes

In figure 1, we compare the reconstructed (left panel) and true (right panel) large-scale density field. We show the projected density distribution at $z = 1$ in the x-y plane of thickness 2 Mpc/h. We use short modes up to $q_{\rm max}$ =0.4 h/Mpc for reconstruction, and we smooth both fields with a Gaussian filter of radius $R = 15$ Mpc/h. In general, the reconstructed density field resembles the morphology of the original density field, but we can clearly see the visible differences between the two. This is most apparent for small-scale features in that some features in the original 2LPT field are missing in the reconstructed field, which implies that the recovered mode is more noise-dominated on smaller scales.

JCAP07(2024)020

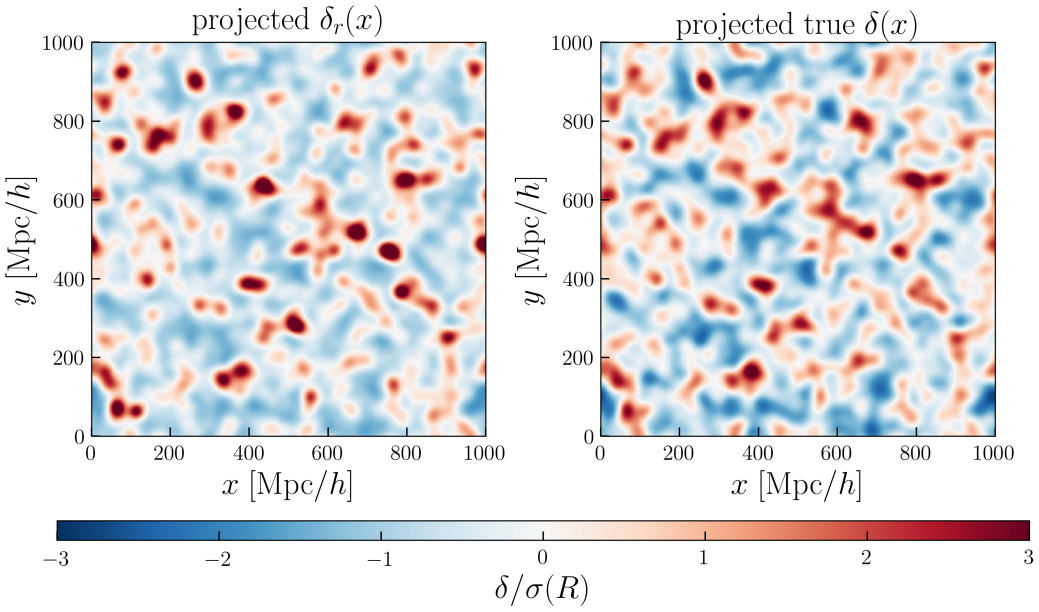

Figure 1. The two-dimensional (the 2 h/Mpc slice on the x-y ($z = 0$) plane) morphology of the reconstructed density field (*Left*) and the 2LPT density fields (*Right*) at $z = 1$. We smooth all fields using the spherical Gaussian filter with the radius $R = 15$ h/Mpc. The maximum wavenumber of modes used for reconstruction is $q_{\rm max} = 0.4\,h$/Mpc. As indicated by the color bar, both are normalized to their own variance.

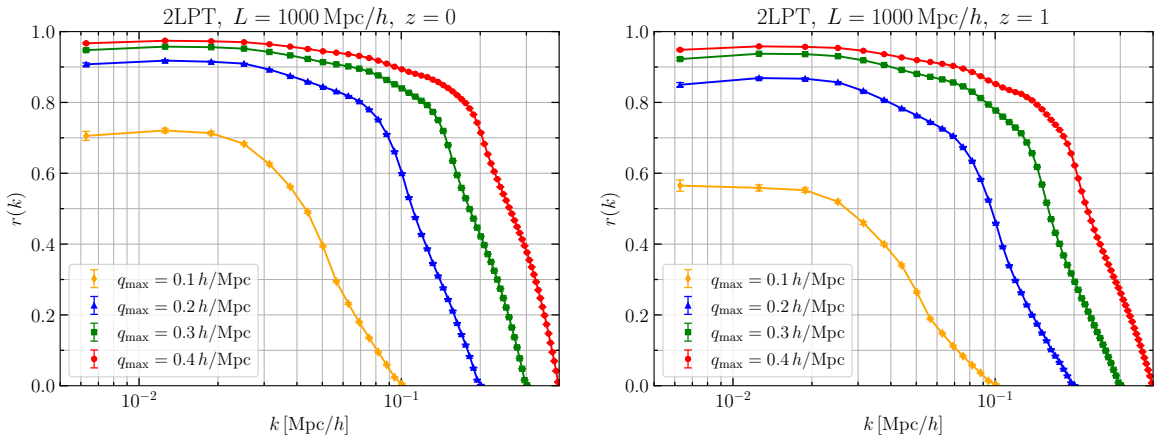

Figure 2. The cross-correlation coefficient, $r(k)$ defined in equation (2.18), between the reconstructed and true long modes in 2LPT simulations at $z = 0$ (left) and $z = 1$ (right). Each line shows the result from different maximum wavenumber of short modes for reconstruction, $q_{\rm max} = 0.1$, 0.2, 0.3, and 0.4 h/Mpc. The error bars indicate the standard deviation of the mean measured from 100 2LPT realizations.

JCAP07(2024)020

For a more quantitative comparison, we compute the cross-correlation coefficients between the recovered and original long modes, and the result is shown in figure 2 for the results at redshifts $z = 0$ (left) and $z = 1$ (right). At each redshift, we show four different results corresponding to four different values of $q_{\rm max} = 0.1$, 0.2, 0.3, and 0.4 $h/$ Mpc.

First of all, the cross-correlation coefficient quickly drops as the long-mode wavenumber k approaches $q_{\rm max}$. That is because there are fewer and fewer short modes for reconstruction as k approaches $q_{\rm max}$. Then the recovered long modes are dominated by noise and correlate weakly with the original long modes. This is consistent with the lack of small-scale features we observe in the morphological comparison in figure 1.

We also notice that the cross-correlation coefficients increase with $q_{\rm max}$, with an especially large improvement from $q_{\rm max} = 0.1$ h/Mpc to 0.2 h/Mpc. With $q_{\rm max} = 0.4$ h/Mpc, the cross-correlation coefficient reaches 0.95 on the large-scale end. This implies that the phase of the recovered long mode from the fossil estimator is highly correlated with the true long mode. Clearly, including more short modes leads to a more accurate reconstruction of the phase of the long modes.

Comparing the two panels, for a fixed $q_{\rm max}$, the cross-correlation coefficient at $z = 0$ is larger than $z = 1$ for all four $q_{\rm max}$ cases. As derived in equation (2.20), the cross-correlation coefficient is close to unity when $P_G^N/P_L \ll 1$. Since the noise power spectrum P_G^N on large scales only weakly depends on the redshift [equation (2.16) and equation (3.4)] while the amplitude of linear power spectrum grows in time, the cross-correlation coefficient is closer to unity at lower redshifts. In other words, a higher signal-to-noise ratio of the power spectrum leads to a larger cross-correlation coefficient at lower redshifts. This phenomenon can also be interpreted as a consequence of the stronger nonlinear coupling between the long and short modes at lower redshift. That is, at higher redshifts, we can reach the same level of the cross-correlation coefficient only with an increasing dynamic range of the short modes used for the reconstruction.

4.1.2 Signal-to-noise ratio

To test the statistical significance of the reconstruction, we measure the noise power spectrum of the reconstructed long modes and show them in figure 3 for two redshifts $z = 0$ (left) and $z = 1$ (right), and for four different maximum short mode wavenumber, $q_{\rm max}$. For each case, we compare the measured noise power spectrum (data points with errorbar) with the theoretical estimation from Gaussian assumption equation (2.16) (dashed line) and the full computation [equation (2.25)] including the trispectrum contribution (solid line). We find that while the full noise power spectrum estimate captures the measurement quite well, the Gaussian approximation always underestimates the noise. As coming from the non-Gaussian trispectrum, the discrepancy is most apparent from cases including more small-scale contributions or increasing $q_{\rm max}$.

The signal-to-noise ratio of the long-mode reconstruction significantly improves as increasing $q_{\rm max}$, because more short-modes are used for the reconstruction. At the same time, including short modes in the nonlinear scales deviates the noise power spectrum from the Gaussian approximated one. Such a deviation does not merely cause an underestimation of the errorbar, because, as shown in equation (2.17), estimating the long-mode power spectrum

JCAP07(2024)020

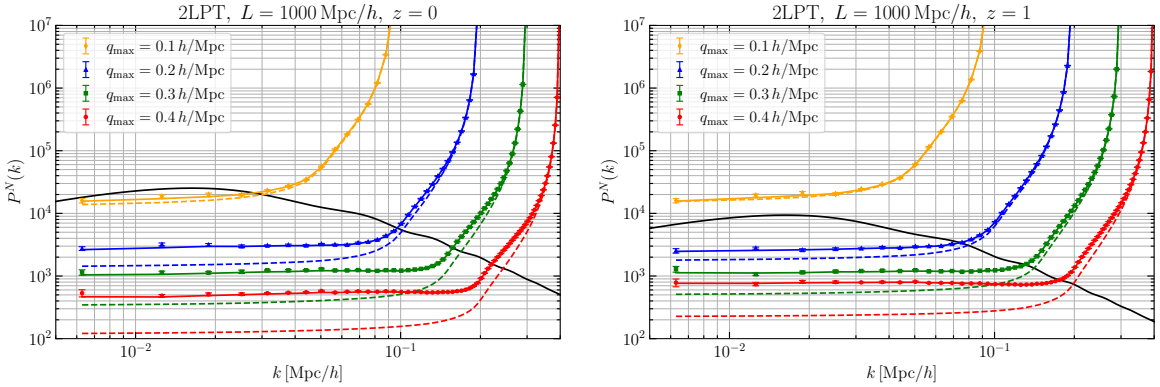

Figure 3. The significance of the long-mode reconstruction at $z = 0$ (left panel) and $z = 1$(right panel). For both panels, we show the linear matter power spectrum (signal) as a solid black line and the noise power spectrum with four different $q_{\rm max} = 0.1$, 0.2, 0.3, and 0.4 h/Mpc as different colors: measured from ten 2LPT realizations (dots with error bars), Gaussian estimate [equation (2.16)] (dashed lines), and non-Gaussian estimate, full expressions in equation (2.25) (solid lines).

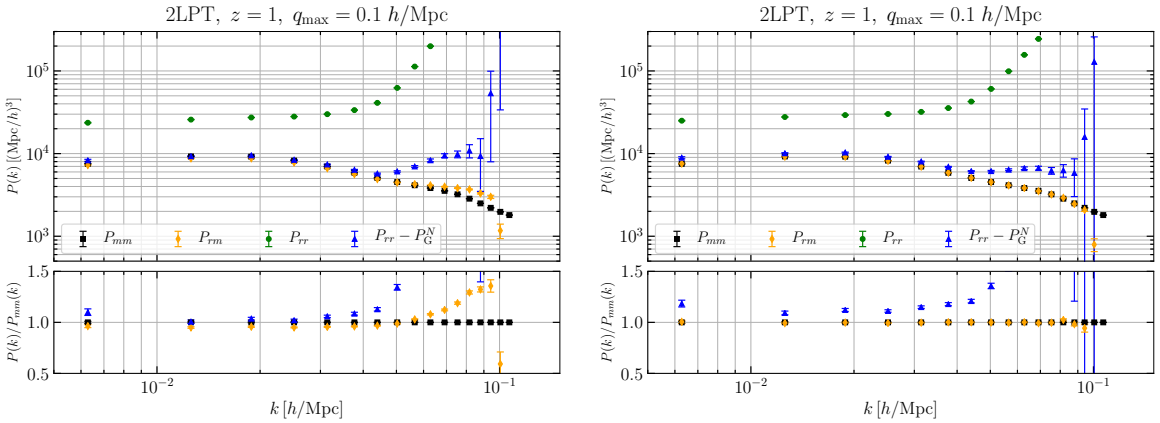

Figure 4. The ensemble mean of the power spectrum from 1,000 2LPT simulations at $z = 1$, using $q_{\rm max} = 0.1\,h/\mathrm{Mpc}$. We highlight the importance of the fossil kernel by comparing the fossil kernel $f(\boldsymbol{k}_1, \boldsymbol{k}_2; \boldsymbol{k})$ from the tree-level bispectrum (*Left*) and from the bispectrum and power spectrum measured from an average of 10,000 2LPT realizations (*right*). For both panels, black (sqaures), orange (diamonds), blue (triangles), and green (dots) are, respectively, the true matter power spectrum, the cross power spectrum between the recovered and the original field, and the recovered power spectrum subtracting the noise calculated from Gaussian approximation. We also show the ratio of each power spectrum to the $P_{mm}(k)$ in the bottom panels.

requires subtracting the noise power spectrum. That is, underestimation of the noise power spectrum leads to the systematic overestimation bias of the long-mode power spectrum. Section 4.1.3 demonstrates this point explicitly.

4.1.3 The power spectrum of the recovered long modes

In figure 4, we show the ensemble mean of the power spectra from the long modes, both auto and cross-correlation with the original mode, reconstructed from the 1,000 2LPT realizations

JCAP07(2024)020

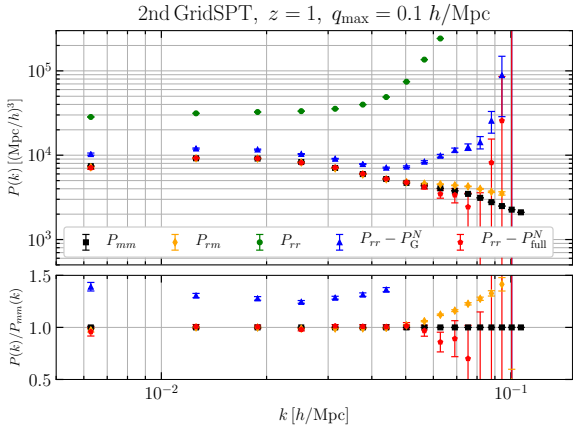

Figure 5. The reconstructed long mode power spectrum when applying the fossil estimator to the second-order GridSPT density field. The symbols are the same as figure 4 except for the red pentagon, which shows the recovered power spectrum with the full noise subtracted. The recovered long-mode power spectrum matches the ground truth (black square) only after subtracting the noise power spectrum taking into account the non-Gaussian covariance.

at $z = 1$ (with $q_{\rm max} = 0.1\,h/{\rm Mpc}$) and compare them with the power spectrum $P_{mm}(k)$ of the original long mode. A perfect reconstruction would yield a perfect match.

The left panel of figure 4 shows the reconstruction with the tree-level fossil kernel $f(\boldsymbol{k}_1, \boldsymbol{k}_2; \boldsymbol{k})$ in equation (3.4) coming from the tree-level bispectrum, and the right panel shows the results with the measured fossil kernel. We measured the fossil kernel by taking the ratio between the average of the bispectrum and the average long-mode power spectrum using the 10,000 2LPT simulations.

The cross-power spectrum $P_{rm}(k)$ from the tree-level fossil kernel (left panel) deviates from $P_{mm}(k)$, while that from the measured fossil kernel (right panel) is on top of the $P_{mm}(k)$. It is because the 2LPT bispectrum, even at $q_{\rm max} = 0.1\ h/{\rm Mpc}$ at $z = 1$, deviates from the tree-level prediction. This result highlights the importance of having an accurate fossil kernel for the reconstruction. Equation (3.5) shows that the cross power spectrum P_{rm} is determined by the ratio between the true bispectrum and the fossil kernel, which is equal to the linear matter power spectrum P_L on large scales. The high-k deviation of P_{rm} from P_{mm} is due to the correlation between the second-order long mode and two first-order short modes (the second term in equation (3.5)).

The auto power spectrum of the recovered modes P_{rr} exceeds P_{mm} due to the presence of the noise power spectrum (see equation (2.17)). Therefore, P_{rr} increase rapidly when k approaches $q_{\rm max}$. This is also consistent with the quick drop of the cross-correlation coefficient at high k as we showed in figure 2. Subtracting the noise power spectrum P^N_G estimated using the Gaussian approximation reduces them closer to $P_{mm}(k)$, but still $P_{rr} - P^N_G$ disagrees with P_{mm}. As we shall show in section 4.2, this is due to the trispectrum contribution to the noise power spectrum.

JCAP07(2024)020

4.2 The ideal toy: reconstruction from the second-order GridSPT

In section 3.2.2, we show that using an accurate squeezed-limit bispectrum for the fossil kernel is essential to get the correct cross-correlation power spectrum. At the same time, the estimated auto-correlation of the recovered long mode deviates from that of the original field when subtracting the noise power spectrum estimated using the Gaussian approximation. To show the proof-of-concept case where we can recover both cross-correlation and auto power spectrum of the long mode, we apply the fossil estimator to the second-order GridSPT realizations. The second-order GridSPT realizations contain the nonlinear density field up to the second order so that we can estimate all relevant correlations.

Figure 5 show the result of the reconstruction analysis using 1,000 second-order GridSPT realizations with $q_{\rm max} = 0.1\,h/{\rm Mpc}$. Just like in section 3.2.2, when estimating the noise power spectrum with Gaussian approximation, the recovered power spectrum (the blue triangles) is about 30 percent higher than P_{mm}. We reach the agreement (red pentagon symbols) only after including the full covariance matrix with the trispectrum contribution. To do so, we measure the off-diagonal covariance matrix numerically from 10,000 realizations of second-order GridSPT and calculate the full noise power spectrum $P^N_{\rm full}$ according to equation (2.25).

Therefore, it is essential to model the correct noise power spectrum, including the off-diagonal terms in the covariance matrix or trispectrum. Neglecting these terms not only underestimates the uncertainties in the estimated long-mode power spectrum, but also introduces systematic bias in the power spectrum of recovered long modes.

4.3 A more realistic toy: reconstruction with the fourth-order GridSPT

Thus far, we have reconstructed the long modes from the controlled toy nonlinear density fields generated from the 2LPT (section 3.2.2) and the second-order GridSPT (section 4.2), from which we find that we have to use an accurate fossil kernel and to incorporate the non-Gaussian covariance (including the trispectrum) in order to achieve an unbiased reconstruction of the long modes. Although useful for the analysis, of course, none of the test nonlinear density fields is close to the real cosmic density field. In this section, we study the limitations of the tree-level fossil estimator by using the fourth-order GridSPT density field.

4.3.1 The reach of the tree-level fossil estimator

We apply the tree-level fossil estimator in fourth-order GridSPT, using $q_{\rm max} = 0.1$, 0.133, 0.167, 0.2 h/Mpc, respectively, at $z = 0, 1, 2, 3$, which are the highest wavenumber where the tree-level bispectrum models the measured squeezed bispectrum in N-body simulations within the 2% accuracy [64]. That is, the reconstruction must be unbiased within these dynamic ranges.

In figure 6, we present the best cross-correlation coefficient that the estimator can achieve at $z = 0$, 1, 2, 3. For all four redshifts on large scales, the cross-correlation coefficients reach $r = 0.7$, which indicates the best signal-to-noise ratios of the recovered long mode based on tree-level bispectrum theory.

Figure 7 presents the measured auto- and cross- power spectra at the four redshifts. One notable feature is that the cross-power spectra match the true matter power spectra on large scales. This fact ensures that the tree-level fossil estimator is unbiased in the dynamical

JCAP07(2024)020

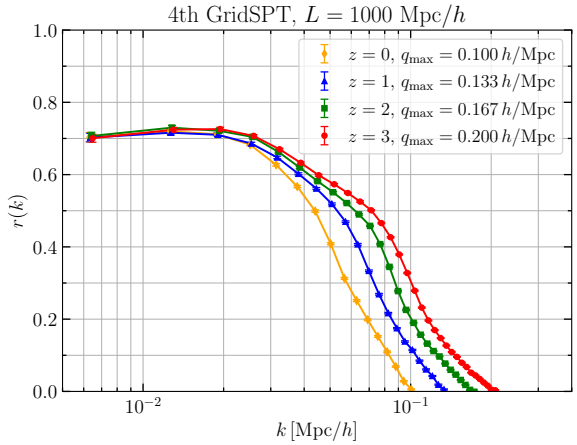

Figure 6. The cross-correlation coefficient between the recovered field and the original field constructed using four-order GridSPT realizations. For the reconstruction, we use four different $q_{\rm max} = 0.1$, 0.133, 0.167, $0.2\,h$/Mpc, respectively, at $z = 0$, 1, 2, 3 such that the squeezed bispectrum works within 2% accuracy, as determined by [64]. The error bars indicate the standard deviation of the mean measured from 100 GridSPT realizations.

ranges. To enhance the signal-to-noise ratio of the reconstructed mode, we need to find a more accurate bispectrum model, e.g. one-loop SPT bispectrum, such that the estimator can safely include more short modes on smaller scales without biasing the recovered long mode.

4.3.2 The limit of the tree-level fossil estimator

In section 2.2, we have proved that the fossil estimator is equivalent to the squeezed-limit bispectrum. Therefore, the fossil estimator reconstructs the linear-order density mode while the nonlinear part of the density mode is neglected. We test the scheme in this section by comparing the reconstructed power spectrum with the original density power spectrum which contains the nonlinear terms.

Figure 8 shows the difference between the reconstructed power spectrum and the original long-mode power spectrum in a unit of the cosmic variance uncertainty $\sigma(P_{rr})$. We use 100 fourth-order GridSPT realizations for this analysis. For all $(z = 0, 1, 2, 3)$ redshifts. This plot shows that the nonlinear corrections are only a few percent of the cosmic variance at all four redshifts, which indicates the nonlinear correction of the recovered mode can be safely neglected with the range of q and the simulation box size we use in this study. Note that the nonlinear correction would become important if we work with either a bigger simulation box or a larger dynamical range of q, because the former raises the number of modes while the latter suppresses the noise power spectrum in P_{rr}. In either case, the nonlinear correction could be comparable to the cosmic variance of P_{rr}. And we have to carefully pick out the range of wavenumbers that the fossil estimator works.

5 Conclusion & discussion

Here, we investigate the reconstruction of long-wavelength density modes based on the off-diagonal correlation between short-wavelength modes in the Fourier space, a phenomenon

JCAP07(2024)020

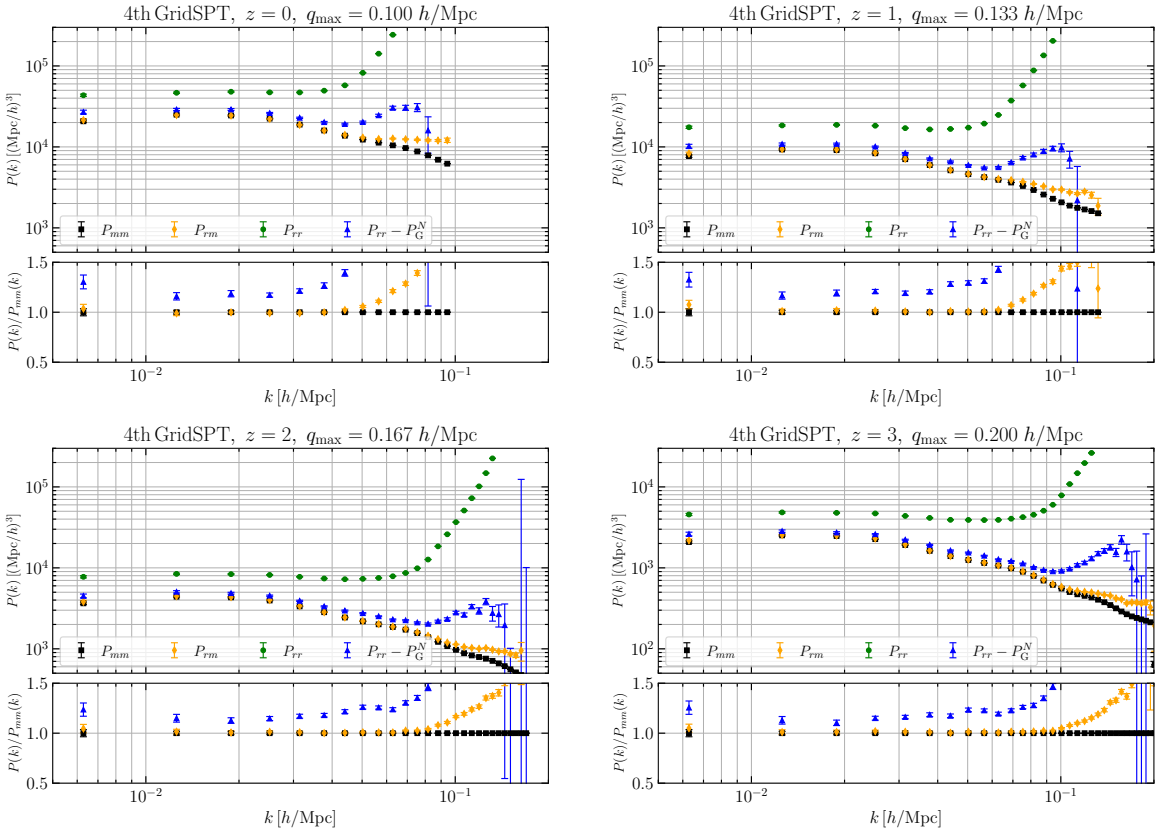

Figure 7. The ensemble mean of the power spectrum measured from 1,000 GridSPT realizations at $z = 0$, 1, 2, and 3, using the $q_{\rm max}$ determined as one-loop bispectrum models the squeezed-limit with 2 % accuracy [64]. The blue triangles are the recovered power spectrum subtracting the noise calculated from the Gaussian approximation. The orange diamonds are the cross-power spectrum between the recovered and the original field. Symbols are identical to figure 4. For all cases, the bottom panel shows the ratio between the power spectra and the true matter power spectrum P_{mm}.

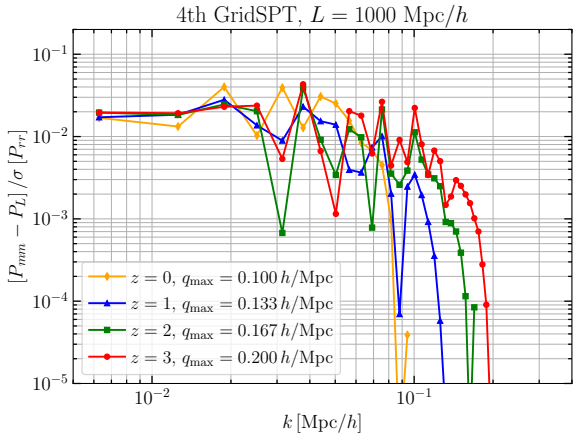

Figure 8. The ratio between the neglected nonlinear correction to the reconstructed matter power spectrum and the cosmic variance of the recovered auto power spectrum at $z = 0, 1, 2, 3$. The results are measured in 100 fourth-order GridSPT realizations. We use the same $q_{\rm max}$ values as in figure 6.

JCAP07(2024)020

called "clustering fossil". The off-diagonal correlation can be understood as the local inhomogeneities introduced by the long mode. The coupling coefficient between short modes can be modeled from the squeezed-limit bispectrum and be used to pick up the specific nonlinear correlation that we use to estimate the long modes.

Throughout this paper, we use the tree-level bispectrum in standard perturbation theory to write down the quadratic estimator for the long mode in equation (2.15), i.e., we only consider the coupling between the second-order short modes and the linear-order long mode. We have tested the same estimator in both second-order GridSPT and 2LPT simulations. By calculating the ensemble mean of the cross power spectrum $P_{rm}(k)$, we show that the estimator is unbiased in second-order GridSPT but biased in 2LPT. We manage to remove the bias of the recovered long mode in 2LPT simulations with the fossil kernel term measured from the 2LPT bispectrum. The results imply that the coupling coefficient from an accurate squeezed-limit bispectrum model guarantees an unbiased estimator.

We also aim to reconstruct the long-mode power spectrum. The auto power spectrum of the recovered mode P_{rr} involves the four-point correlation function in Fourier space. P_{rr} contains both the long mode power spectrum and the noise power spectrum. The noise power term is minimized if we use the inverse variance weight with the assumption that the short density modes are Gaussian, which is not true in the real universe. In fact, the noise power spectrum also receives the contribution from the connected four-point correlation function, which becomes more significant as we increase the largest wavenumber of short modes for reconstruction. Therefore, it is crucial to include the connected four-point correlation function into the noise power spectrum to accurately reconstruct the long-mode power spectrum.

We also demonstrate that the cross-correlation coefficient between the recovered and true long mode is determined by the ratio between the long-mode power spectrum and noise. To get a higher cross-correlation coefficient, we can suppress the noise power spectrum by including more short modes for reconstruction. With the estimation from fourth-order GridSPT, we estimate the best cross-correlation coefficient the tree-level fossil estimator could achieve in N-body simulation is 0.7 at redshifts $z = 0, 1, 2, 3$ while keeping the recovered long mode unbiased.

In order to apply the fossil estimator in N-body simulations and enhance the reconstruction power, we need to include more short modes of larger wavenumbers for the reconstruction. Therefore, we need to use higher-order coupling terms beyond tree-level in the perturbation theory into the estimator. However, the estimator can still be biased because the perturbation theory fails on the fully nonlinear scale. Fortunately, the response approach [79, 80] enables us to measure the squeezed-limit bispectrum in mild-sized N-body simulations accurately, which can potentially extend $q_{\rm max}$ from quasi-linear scale to fully nonlinear scale without biasing the estimator. We leave this part to the future work.

To make this method more feasible in future galaxy surveys and 21cmIM experiments, we need to investigate the reconstruction from the biased tracer field. Using perturbative galaxy bias expansion, we can calculate the squeezed-limit galaxy-galaxy-matter bispectrum and reconstruct the matter density long mode from the galaxy short modes. We also need to include the redshift-space distortion effect and the foreground wedge in the 21cmIM [52]. Our next step, then, is to understand how galaxy bias and the observational effects impact the reconstruction power.

JCAP07(2024)020

There are many interesting applications of fossil estimators. By implementing the estimator in galaxy survey, the recovered matter long mode and observed galaxy long mode naturally form multi tracers and can constrain the local-type primordial non-Gaussianity during the inflation without cosmic variance [81, 82] (See [52] for more details.). Furthermore, we can construct the tensor-type fossil estimator to infer the primordial gravitational waves from the galaxy short modes, if the galaxy-galaxy-gravitational wave bispectrum can be modeled accurately. Based on the clustering fossil theory, this reconstruction method could become a powerful new statistical approach to probing the large-scale structure and increasing the scientific output of the future galaxy and 21cm surveys.

Acknowledgments

We want to thank Caryl Gronwall and Michael Sigel for useful comments on the early draft of this paper. ZW wants to acknowledge Hong-ming Zhu and Oliver Philcox for useful discussions and Yuanheng Wang for helpful support during the completion of this paper. This research made use of the Roar Collab Supercomputer at Penn State University. ZW and DJ acknowledge support from the National Science Foundation Grant No. AST-2307026 at Penn State University. DJ is supported by KIAS Individual Grant PG088301 at Korea Institute for Advanced Study and was supported by NASA ATP program (80NSSC18K1103) at Penn State University.

A Fossil estimator in the continuous limit

To get a fast estimation of the noise power spectrum of the reconstructed modes, we can write down the fossil estimator in the continuous limit.

$$\begin{aligned}\delta_r(k) &= P_G^N(\boldsymbol{k})\int_q \frac{\delta(\boldsymbol{q})\delta(\boldsymbol{k}-\boldsymbol{q})f(\boldsymbol{q},\boldsymbol{k}-\boldsymbol{q})}{P(q)P(|\boldsymbol{k}-\boldsymbol{q}|)} \\ &= P_G^N(k)\int_k^{q_{\max}}\frac{\mathrm{d}q}{(2\pi)^2}q^2\int \mathrm{d}\mu\frac{\delta(\boldsymbol{q})\delta(\boldsymbol{k}-\boldsymbol{q})f(\boldsymbol{q},\boldsymbol{k}-\boldsymbol{q})}{P(q)P(|\boldsymbol{k}-\boldsymbol{q}|)}.\end{aligned} \tag{A.1}$$

In the continuous limit, the noise power spectrum under the Gaussian assumption is

$$P_G^N(k) = \left[\int_k^{q_{\max}}\frac{\mathrm{d}q}{(2\pi)^2}q^2\int \mathrm{d}\mu\frac{|f(\boldsymbol{q},\boldsymbol{k}-\boldsymbol{q})|^2}{P(q)P(|\boldsymbol{k}-\boldsymbol{q}|)}\right]^{-1} \tag{A.2}$$

Here we have used the relation between the integration and the grid-based discrete summation over the Fourier space

$$\frac{1}{V}\sum = \int\frac{\mathrm{d}^3q}{(2\pi)^3}. \tag{A.3}$$

The range of $\mu \equiv \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{q}}$ is constrained by $q \geq |\boldsymbol{k}-\boldsymbol{q}| > k$, which is

$$\frac{k}{2q} \leq \mu \leq \min\left\{1, \frac{q}{2k}\right\}. \tag{A.4}$$

JCAP07(2024)020

References

[1] T.-C. Chang, U.-L. Pen, J.B. Peterson and P. McDonald, *Baryon acoustic oscillation intensity mapping as a test of dark energy*, *Phys. Rev. Lett.* **100** (2008) 091303 [arXiv:0709.3672] [INSPIRE].

[2] A. Loeb and S. Wyithe, *Precise measurement of the cosmological power spectrum with a dedicated 21 cm survey after reionization*, *Phys. Rev. Lett.* **100** (2008) 161301 [arXiv:0801.1677] [INSPIRE].

[3] X. Chen, *The Tianlai project: a 21 cm cosmology experiment*, *Int. J. Mod. Phys. Conf. Ser.* **12** (2012) 256 [arXiv:1212.6278] [INSPIRE].

[4] K. Bandura et al., *Canadian Hydrogen Intensity Mapping Experiment (CHIME) pathfinder*, *Proc. SPIE Int. Soc. Opt. Eng.* **9145** (2014) 22 [arXiv:1406.2288] [INSPIRE].

[5] L.B. Newburgh et al., *HIRAX: a probe of dark energy and radio transients*, *Proc. SPIE Int. Soc. Opt. Eng.* **9906** (2016) 99065X [arXiv:1607.02059] [INSPIRE].

[6] R.A. Battye et al., *HI intensity mapping: a single dish approach*, *Mon. Not. Roy. Astron. Soc.* **434** (2013) 1239 [arXiv:1209.0343] [INSPIRE].

[7] COSMOLOGY-SWG and EOR/CD-SWG collaborations, *Cosmology from EoR/Cosmic Dawn with the SKA*, *PoS* **AASKA14** (2015) 012 [arXiv:1501.04291] [INSPIRE].

[8] H.-J. Seo et al., *A ground-based 21 cm baryon acoustic oscillation survey*, *Astrophys. J.* **721** (2010) 164 [arXiv:0910.5007] [INSPIRE].

[9] E. Chapman and V. Jelić, *Foregrounds and their mitigation*, arXiv:1909.12369 [INSPIRE].

[10] A. Liu and J.R. Shaw, *Data analysis for precision 21 cm cosmology*, *Publ. Astron. Soc. Pac.* **132** (2020) 062001 [arXiv:1907.08211] [INSPIRE].

[11] M.F. Morales, J.D. Bowman and J.N. Hewitt, *Improving foreground subtraction in statistical observations of 21 cm emission from the epoch of reionization*, *Astrophys. J.* **648** (2006) 767 [astro-ph/0510027] [INSPIRE].

[12] X.-M. Wang, M. Tegmark, M. Santos and L. Knox, *Twenty-one centimeter tomography with foregrounds*, *Astrophys. J.* **650** (2006) 529 [astro-ph/0501081] [INSPIRE].

[13] J.D. Bowman, M.F. Morales and J.N. Hewitt, *Foreground contamination in interferometric measurements of the redshifted 21 cm power spectrum*, *Astrophys. J.* **695** (2009) 183 [arXiv:0807.3956] [INSPIRE].

[14] A. Liu, M. Tegmark and M. Zaldarriaga, *Will point sources spoil 21 cm tomography?*, *Mon. Not. Roy. Astron. Soc.* **394** (2009) 1575 [arXiv:0807.3952] [INSPIRE].

[15] A. Liu and M. Tegmark, *A method for 21 cm power spectrum estimation in the presence of foregrounds*, *Phys. Rev. D* **83** (2011) 103006 [arXiv:1103.0281] [INSPIRE].

[16] A. Parsons et al., *A sensitivity and array-configuration study for measuring the power spectrum of 21 cm emission from reionization*, *Astrophys. J.* **753** (2012) 81 [arXiv:1103.2135] [INSPIRE].

[17] E. Chapman et al., *Foreground removal using FastICA: a showcase of LOFAR-EoR*, *Mon. Not. Roy. Astron. Soc.* **423** (2012) 2518 [arXiv:1201.2190] [INSPIRE].

[18] E. Chapman et al., *The scale of the problem: recovering images of reionization with GMCA*, *Mon. Not. Roy. Astron. Soc.* **429** (2013) 165 [arXiv:1209.4769] [INSPIRE].

[19] J.S. Dillon, A. Liu and M. Tegmark, *A fast method for power spectrum and foreground analysis for 21 cm cosmology*, *Phys. Rev. D* **87** (2013) 043005 [arXiv:1211.2232] [INSPIRE].

JCAP07(2024)020

[20] L. Wolz et al., *Erasing the Milky Way: new cleaning technique applied to GBT intensity mapping data*, *Mon. Not. Roy. Astron. Soc.* **464** (2017) 4938 [arXiv:1510.05453] [INSPIRE].

[21] I.P. Carucci, M.O. Irfan and J. Bobin, *Recovery of 21 cm intensity maps with sparse component separation*, *Mon. Not. Roy. Astron. Soc.* **499** (2020) 304 [arXiv:2006.05996] [INSPIRE].

[22] S. Furlanetto, S.P. Oh and F. Briggs, *Cosmology at low frequencies: the 21 cm transition and the high-redshift universe*, *Phys. Rept.* **433** (2006) 181 [astro-ph/0608032] [INSPIRE].

[23] M.F. Morales, B. Hazelton, I. Sullivan and A. Beardsley, *Four fundamental foreground power spectrum shapes for 21 cm cosmology observations*, *Astrophys. J.* **752** (2012) 137 [arXiv:1202.3830] [INSPIRE].

[24] PLANCK collaboration, *Planck 2015 results. XIII. Cosmological parameters*, *Astron. Astrophys.* **594** (2016) A13 [arXiv:1502.01589] [INSPIRE].

[25] C.M. Trott, R.B. Wayth and S.J. Tingay, *The impact of point source subtraction residuals on 21 cm epoch of reionization estimation*, *Astrophys. J.* **757** (2012) 101 [arXiv:1208.0646] [INSPIRE].

[26] J.S. Dillon, *Power spectrum estimation for 21 cm cosmology*, *Amer. Astron. Soc. Meet. Abs.* **224** (2014) 318.04.

[27] A. Liu, A.R. Parsons and C.M. Trott, *Epoch of reionization window. I. Mathematical formalism*, *Phys. Rev. D* **90** (2014) 023018 [arXiv:1404.2596] [INSPIRE].

[28] J.C. Pober et al., *What next-generation 21 cm power spectrum measurements can teach us about the epoch of reionization*, *Astrophys. J.* **782** (2014) 66 [arXiv:1310.7031] [INSPIRE].

[29] J.C. Pober, *The impact of foregrounds on redshift space distortion measurements with the highly-redshifted 21 cm line*, *Mon. Not. Roy. Astron. Soc.* **447** (2015) 1705 [arXiv:1411.2050] [INSPIRE].

[30] N. Kaiser, *On the spatial correlations of Abell clusters*, *Astrophys. J. Lett.* **284** (1984) L9 [INSPIRE].

[31] N. Kaiser, *Clustering in real space and in redshift space*, *Mon. Not. Roy. Astron. Soc.* **227** (1987) 1 [INSPIRE].

[32] D. Jeong and F. Schmidt, *Large-scale structure observables in general relativity*, *Class. Quant. Grav.* **32** (2015) 044001 [arXiv:1407.7979] [INSPIRE].

[33] V. Desjacques, D. Jeong and F. Schmidt, *Large-scale galaxy bias*, *Phys. Rept.* **733** (2018) 1 [arXiv:1611.09787] [INSPIRE].

[34] M. McQuinn et al., *Cosmological parameter estimation using 21 cm radiation from the epoch of reionization*, *Astrophys. J.* **653** (2006) 815 [astro-ph/0512263] [INSPIRE].

[35] Y. Mao et al., *How accurately can 21 cm tomography constrain cosmology?*, *Phys. Rev. D* **78** (2008) 023529 [arXiv:0802.1710] [INSPIRE].

[36] P. Bull, P.G. Ferreira, P. Patel and M.G. Santos, *Late-time cosmology with 21 cm intensity mapping experiments*, *Astrophys. J.* **803** (2015) 21 [arXiv:1405.1452] [INSPIRE].

[37] D. Karagiannis, A. Slosar and M. Liguori, *Forecasts on primordial non-Gaussianity from 21 cm intensity mapping experiments*, *JCAP* **11** (2020) 052 [arXiv:1911.03964] [INSPIRE].

[38] U.-L. Pen et al., *Cosmic tides*, arXiv:1202.5804 [INSPIRE].

[39] H.-M. Zhu et al., *Cosmic tidal reconstruction*, *Phys. Rev. D* **93** (2016) 103504 [arXiv:1511.04680] [INSPIRE].

JCAP07(2024)020

[40] F. Schmidt, E. Pajer and M. Zaldarriaga, *Large-scale structure and gravitational waves III: tidal effects*, *Phys. Rev. D* **89** (2014) 083507 [arXiv:1312.5616] [INSPIRE].

[41] N.G. Karaçaylı and N. Padmanabhan, *Anatomy of cosmic tidal reconstruction*, *Mon. Not. Roy. Astron. Soc.* **486** (2019) 3864 [arXiv:1904.01387] [INSPIRE].

[42] H.-M. Zhu, T.-X. Mao and U.-L. Pen, *Cosmic tidal reconstruction with halo fields*, *Astrophys. J.* **929** (2022) 5 [arXiv:2108.01575] [INSPIRE].

[43] S.-H. Zang, H.-M. Zhu, M. Schmittfull and U.-L. Pen, *Cosmic tidal reconstruction in redshift space*, *Astrophys. J.* **962** (2024) 21 [arXiv:2212.04294] [INSPIRE].

[44] H.-M. Zhu, U.-L. Pen, Y. Yu and X. Chen, *Recovering lost 21 cm radial modes via cosmic tidal reconstruction*, *Phys. Rev. D* **98** (2018) 043511 [arXiv:1610.07062] [INSPIRE].

[45] S. Foreman, P.D. Meerburg, A. van Engelen and J. Meyers, *Lensing reconstruction from line intensity maps: the impact of gravitational nonlinearity*, *JCAP* **07** (2018) 046 [arXiv:1803.04975] [INSPIRE].

[46] D. Jeong and M. Kamionkowski, *Clustering fossils from the early universe*, *Phys. Rev. Lett.* **108** (2012) 251301 [arXiv:1203.0302] [INSPIRE].

[47] L. Dai, D. Jeong and M. Kamionkowski, *Seeking inflation fossils in the cosmic microwave background*, *Phys. Rev. D* **87** (2013) 103006 [arXiv:1302.1868] [INSPIRE].

[48] E. Dimastrogiovanni, M. Fasiello, D. Jeong and M. Kamionkowski, *Inflationary tensor fossils in large-scale structure*, *JCAP* **12** (2014) 050 [arXiv:1407.8204] [INSPIRE].

[49] P. Li, S. Dodelson and R.A.C. Croft, *Large scale structure reconstruction with short-wavelength modes*, *Phys. Rev. D* **101** (2020) 083510 [arXiv:2001.02780] [INSPIRE].

[50] P. Li, R.A.C. Croft and S. Dodelson, *New probes of large scale structure*, arXiv:2007.00226 [INSPIRE].

[51] F. Bernardeau, S. Colombi, E. Gaztanaga and R. Scoccimarro, *Large scale structure of the universe and cosmological perturbation theory*, *Phys. Rept.* **367** (2002) 1 [astro-ph/0112551] [INSPIRE].

[52] O. Darwish et al., *Density reconstruction from biased tracers and its application to primordial non-Gaussianity*, *Phys. Rev. D* **104** (2021) 123520 [arXiv:2007.08472] [INSPIRE].

[53] A. Taruya, T. Nishimichi and D. Jeong, *Grid-based calculation for perturbation theory of large-scale structure*, *Phys. Rev. D* **98** (2018) 103532 [arXiv:1807.04215] [INSPIRE].

[54] A. Taruya, T. Nishimichi and D. Jeong, *Covariance of the matter power spectrum including the survey window function effect: N-body simulations versus fifth-order perturbation theory on grids*, *Phys. Rev. D* **103** (2021) 023501 [arXiv:2007.05504] [INSPIRE].

[55] A. Taruya, T. Nishimichi and D. Jeong, *Grid-based calculations of redshift-space matter fluctuations from perturbation theory: UV sensitivity and convergence at the field level*, *Phys. Rev. D* **105** (2022) 103507 [arXiv:2109.06734] [INSPIRE].

[56] D. Jeong, *Cosmology with high ($z > 1$) redshift galaxy surveys*, Ph.D. thesis, University of Texas, Austin, TX, U.S.A. (2010).

[57] C.-T. Chiang, C. Wagner, F. Schmidt and E. Komatsu, *Position-dependent power spectrum of the large-scale structure: a novel method to measure the squeezed-limit bispectrum*, *JCAP* **05** (2014) 048 [arXiv:1403.3411] [INSPIRE].

[58] C.-T. Chiang, *Position-dependent power spectrum: a new observable in the large-scale structure*, Ph.D. thesis, Munich U., Munich, Germany (2015) [arXiv:1508.03256] [INSPIRE].

JCAP07(2024)020

[59] S. Adhikari, D. Jeong and S. Shandera, *Constraining primordial and gravitational mode coupling with the position-dependent bispectrum of the large-scale structure*, *Phys. Rev. D* **94** (2016) 083528 [arXiv:1608.05139] [INSPIRE].

[60] D. Jeong and F. Schmidt, *Parity-odd galaxy bispectrum*, *Phys. Rev. D* **102** (2020) 023530 [arXiv:1906.05198] [INSPIRE].

[61] W. Hu and T. Okamoto, *Mass reconstruction with CMB polarization*, *Astrophys. J.* **574** (2002) 566 [astro-ph/0111606] [INSPIRE].

[62] T. Okamoto and W. Hu, *CMB lensing reconstruction on the full sky*, *Phys. Rev. D* **67** (2003) 083002 [astro-ph/0301031] [INSPIRE].

[63] D. Jeong and E. Komatsu, *Perturbation theory reloaded: analytical calculation of non-linearity in baryonic oscillations in the real space matter power spectrum*, *Astrophys. J.* **651** (2006) 619 [astro-ph/0604075] [INSPIRE].

[64] J. Tomlinson and D. Jeong, *Spherical bispectrum: a novel visualization scheme for facilitating comparisons*, *JCAP* **08** (2023) 040 [arXiv:2204.00668] [INSPIRE].

[65] E. Schaan, S. Ferraro and D.N. Spergel, *Weak lensing of intensity mapping: the cosmic infrared background*, *Phys. Rev. D* **97** (2018) 123539 [arXiv:1802.05706] [INSPIRE].

[66] E. Sefusatti and R. Scoccimarro, *Galaxy bias and halo-occupation numbers from large-scale clustering*, *Phys. Rev. D* **71** (2005) 063001 [astro-ph/0412626] [INSPIRE].

[67] D. Gualdi and L. Verde, *Integrated trispectrum detection from BOSS DR*12 *NGC CMASS*, *JCAP* **09** (2022) 050 [arXiv:2201.06932] [INSPIRE].

[68] M. Crocce, S. Pueblas and R. Scoccimarro, *Transients from initial conditions in cosmological simulations*, *Mon. Not. Roy. Astron. Soc.* **373** (2006) 369 [astro-ph/0606505] [INSPIRE].

[69] A. Lewis, A. Challinor and A. Lasenby, *Efficient computation of CMB anisotropies in closed FRW models*, *Astrophys. J.* **538** (2000) 473 [astro-ph/9911177] [INSPIRE].

[70] E.T. Vishniac, *Why weakly non-linear effects are small in a zero-pressure cosmology*, *Mon. Not. Roy. Astron. Soc.* **203** (1983) 345.

[71] J.N. Fry, *The galaxy correlation hierarchy in perturbation theory*, *Astrophys. J.* **279** (1984) 499 [INSPIRE].

[72] M.H. Goroff, B. Grinstein, S.J. Rey and M.B. Wise, *Coupling of modes of cosmological mass density fluctuations*, *Astrophys. J.* **311** (1986) 6 [INSPIRE].

[73] Y. Suto and M. Sasaki, *Quasi nonlinear theory of cosmological selfgravitating systems*, *Phys. Rev. Lett.* **66** (1991) 264 [INSPIRE].

[74] N. Makino, M. Sasaki and Y. Suto, *Analytic approach to the perturbative expansion of nonlinear gravitational fluctuations in cosmological density and velocity fields*, *Phys. Rev. D* **46** (1992) 585 [INSPIRE].

[75] B. Jain and E. Bertschinger, *Second order power spectrum and nonlinear evolution at high redshift*, *Astrophys. J.* **431** (1994) 495 [astro-ph/9311070] [INSPIRE].

[76] R. Scoccimarro and J. Frieman, *Loop corrections in nonlinear cosmological perturbation theory 2. Two point statistics and selfsimilarity*, *Astrophys. J.* **473** (1996) 620 [astro-ph/9602070] [INSPIRE].

[77] N. McCullagh, D. Jeong and A.S. Szalay, *Toward accurate modelling of the non-linear matter bispectrum: standard perturbation theory and transients from initial conditions*, *Mon. Not. Roy. Astron. Soc.* **455** (2016) 2945 [arXiv:1507.07824] [INSPIRE].

JCAP07(2024)020

[78] Z. Wang et al., *Perturbation theory remixed: improved nonlinearity modeling beyond standard perturbation theory*, *Phys. Rev. D* **107** (2023) 103534 [arXiv:2209.00033] [INSPIRE].

[79] C. Wagner, F. Schmidt, C.-T. Chiang and E. Komatsu, *Separate universe simulations*, *Mon. Not. Roy. Astron. Soc.* **448** (2015) L11 [arXiv:1409.6294] [INSPIRE].

[80] A. Barreira and F. Schmidt, *Responses in large-scale structure*, *JCAP* **06** (2017) 053 [arXiv:1703.09212] [INSPIRE].

[81] U. Seljak, *Extracting primordial non-gaussianity without cosmic variance*, *Phys. Rev. Lett.* **102** (2009) 021302 [arXiv:0807.1770] [INSPIRE].

[82] P. McDonald and U. Seljak, *How to measure redshift-space distortions without sample variance*, *JCAP* **10** (2009) 007 [arXiv:0810.0323] [INSPIRE].

JCAP07(2024)020