Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **499,** 5902–5914 (2020) doi:10.1093/mnras/staa3165
Advance Access publication 2020 October 15

A new approach to observational cosmology using the scattering transform

Sihao Cheng (程思浩) [1]★ Yuan-Sen Ting (丁源森) [2,3,4,5] Brice Ménard[1] and Joan Bruna[3,6,7]

[1]*Department of Physics and Astronomy, The Johns Hopkins University, 3400 N Charles Street, Baltimore, MD 21218, USA*
[2]*Institute for Advanced Study, Princeton, NJ 08540, USA*
[3]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[4]*Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA*
[5]*Research School of Astronomy & Astrophysics, Australian National University, Cotter Road, Weston, ACT 2611, Australia*
[6]*Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA*
[7]*Center for Data Science, New York University, New York, NY 10011, USA*

Accepted 2020 October 7. Received 2020 September 24; in original form 2020 July 15

ABSTRACT

Parameter estimation with non-Gaussian stochastic fields is a common challenge in astrophysics and cosmology. In this paper, we advocate performing this task using the *scattering transform*, a statistical tool sharing ideas with convolutional neural networks (CNNs) but requiring neither training nor tuning. It generates a compact set of coefficients, which can be used as robust summary statistics for non-Gaussian information. It is especially suited for fields presenting localized structures and hierarchical clustering, such as the cosmological density field. To demonstrate its power, we apply this estimator to a cosmological parameter inference problem in the context of weak lensing. On simulated convergence maps with realistic noise, the scattering transform outperforms classic estimators and is on a par with the state-of-the-art CNN. It retains advantages of traditional statistical descriptors, has provable stability properties, allows to check for systematics, and importantly, the scattering coefficients are interpretable. It is a powerful and attractive estimator for observational cosmology and the study of physical fields in general.

Key words: gravitational lensing: weak – methods: statistical – cosmological parameters – large-scale structure of Universe.

1 INTRODUCTION

Non-Gaussian fields are ubiquitous in astrophysics. Analysing them is challenging, as the dimensionality of their description can be arbitrarily high. In addition, there is usually little guidance on which statistical estimator will be most appropriate for parameter inference. In this paper, we advocate using a novel approach, called the scattering transform (Mallat 2012), for the analysis of such fields and, in particular, the matter distribution in the Universe, a highly studied non-Gaussian field.

In many areas of astrophysics and, in particular, in cosmology, extracting non-Gaussian information has been attempted through *N*-point correlation functions (see e.g. Bernardeau, Mellier & van Waerbeke 2002; Takada & Jain 2003; Semboloni et al. 2011; Fu et al. 2014, for weak lensing applications) and polyspectra, their Fourier equivalents (see e.g. Sefusatti et al. 2006). Correlation functions are convenient for theoretical predictions and for measuring weak departures from Gaussianity. However, being high powers of the input field, these statistics suffer from an increasing variance and are not robust to outliers in real data, making them gradually less informative (Welling 2005). If the distribution of field intensity has a long tail, the amount of information accessible to *N*-point functions will quickly decrease (Carron 2011). In addition, the number of configurations to consider for *N*-point functions explodes with the number of points used. As a result, information is highly diluted among coefficients, and it becomes a challenge to efficiently extract information with *N*-point functions. Other methods, including performing a non-linear transformation before calculating correlation functions (Neyrinck, Szapudi & Szalay 2011; Simpson et al. 2011; Carron & Szapudi 2013; Giblin et al. 2018), using topological properties such as Minkowski functionals (Mecke, Buchert & Wagner 1994; Hikage et al. 2003; Kratochvil et al. 2012; Shirasaki & Yoshida 2014), and using biasing properties such as counts of clusters, peaks, and voids (Jain & Van Waerbeke 2000; Marian, Smith & Bernstein 2009; Kratochvil, Haiman & May 2010; Liu et al. 2015a, b; Pisani et al. 2019), have also been considered. However, in the cosmological context, these excursions into non-Gaussian signal analyses have had limited impact in improving existing constraints on cosmological parameters so far.

Recently, convolutional neural networks (CNNs; e.g. Lecun et al. 1998) have claimed supremacy in a wide variety of applications aimed at extracting information from complex data. They have also shown promises to efficiently retrieve cosmological information well beyond second-order statistics (see e.g. Gupta et al. 2018; Ribli, Pataki & Csabai 2019a; Ribli et al. 2019b). While the potential of this method is enormous, it also comes with a number of issues. To precisely and robustly estimate cosmological parameters, CNNs require a large training set. In addition, when applied to real data, systematic errors not included in the training process of CNN can hardly get checked and controlled, whereas for traditional statistics, a simple χ^2 test can do so. As such, the use of CNNs in real data comes with limitations regarding interpretability and validity.

In this paper, we advocate using a different approach called the *scattering transform* to efficiently and robustly extract statis-

★ E-mail: s.cheng@jhu.edu

tical information from non-Gaussian fields.[1] The operations and structure of the scattering transform have close similarities with those built in CNNs, but the scattering transform does not require any training, and like traditional statistics, it generates coefficients with proved properties. It can therefore hopefully overcome the aforementioned limitations encountered with CNNs. In Section 2, we introduce the scattering transform, present intuitive understanding of its coefficients, and visualize its key properties. In Sections 3 and 4, we demonstrate the power of the scattering transform to infer cosmological parameters ($\Omega_{\rm m}$ and σ_8) in the context of weak lensing using simulated convergence maps. As we will show, it outperforms the power spectrum and peak counts, and is on a par with the state-of-the-art CNN. Finally, we comment on the attractive properties of the scattering transform in Section 5 and conclude in Section 6.

2 THE SCATTERING TRANSFORM

In this section, we present the scattering transform and intuitive interpretations of its coefficients. The scattering transform generates a compact set of coefficients that captures substantial non-Gaussian information beyond the power spectrum. In contrast to N-point functions, the scattering coefficients are all proportional to the input data, and do not suffer from the increasing variance issue. Thus, the scattering coefficients, which form a representation of the input field, can be used to extract non-Gaussian information efficiently and robustly. This is particularly attractive from a data analysis point of view.

2.1 Motivation

The scattering transform was originally proposed by Mallat (2012) as a tool for signal processing to extract information from high-dimensional data. In contrast to neural networks, it comes with attractive provable properties, including translational invariance, non-expanding variance, and Lipschitz continuous to spatial deformation (Mallat 2012). Interestingly, the scattering transform has also provided key insights into deciphering the remarkable behaviour and performance of CNNs (Bruna & Mallat 2013). A perhaps counter-intuitive feature of CNNs is that the convolution, though restricting the flexibility of the neural network, dramatically boosts its performance on many types of data. In addition, a successful CNN architecture can often be re-purposed for very different tasks. These facts suggest that a certain mathematical structure enables efficient information extraction from a wide range of complex data. Understanding this structure may dramatically simplify the costly training process required when using neural networks.

The scattering transform has been successfully used in many areas, including audio signal processing (Andén & Mallat 2014), image classification (Bruna & Mallat 2013), texture classification (Sifre & Mallat 2013), materials science (Hirn, Mallat & Poilvert 2017; Eickenberg et al. 2018; Sinz et al. 2020), multifractal analysis in turbulence and finance (Bruna et al. 2015), and graph-structured data (Gama, Ribeiro & Bruna 2018). Several of these examples reached state-of-the-art performance compared to the CNNs in use at the time. In astrophysics, a pioneer application has been performed by Allys et al. (2019) to analyse the interstellar medium.

The scattering transform can be used with two possible goals in mind: representing a specific realization of a field (with a classification goal) or characterizing the global statistical properties of a field. The narratives in these two regimes are slightly different (Mallat 2012; Bruna & Mallat 2013). We will focus on the latter one, which is relevant to cosmological applications.

[1]This work was done simultaneously and independently of that presented in Allys et al. (2020), where the authors apply a different but related technique, the wavelet phase harmonic, to slices of the matter density field.

2.2 Formulation

Here, we present the formulation of the scattering transform in the context of characterizing random fields (Mallat 2012). We focus on the 2D case in this study, but it can be directly generalized to any other dimensionality. For clarity, we will attach the notation (x, y) for the spatial dependence of a field only when it is first introduced.

To extract information from an input field, the scattering transform first generates a group of new fields by recursively applying two operations: a wavelet convolution and a modulus. Then, the expected values of these fields are defined as the scattering coefficients and used to characterize statistical properties of the original field (see Fig. 1 for an illustration). This hierarchical structure, the use of localized convolution kernels, and the use of non-expansive non-linear operator are all elements found in the architecture of CNNs.

Formally, given an input field $I_0(x, y)$, the scattering transform generates a set of first-order fields $I_1(x, y)$ by convolving it with a family of wavelets $\psi^{j,l}(x, y)$ and then taking the modulus:

$$I_1 \equiv \left| I_0 \star \psi^{j_1,l_1} \right| , \tag{1}$$

where I_1 represents a group of fields labelled by the wavelet index j_1, l_1. Wavelets are localized oscillations and band-pass filters. Fig. 2 shows the profiles of Morlet wavelets in real and Fourier spaces. Morlet wavelets are used in our study and described in Appendix A. In general, a family of wavelets covers the whole Fourier space. They all have the same shape but different sizes and orientations, labelled by j and l, respectively. They can all be generated through dilating and rotating a prototype wavelet. In the scattering transform, the convention is to use a dilation factor of 2, such that for a pixelized field, the size of a wavelet $\psi^{j,l}$ in the real space is roughly 2^j pixels.

Having created the first-order fields, one can then iterate the same process to create second-order fields $I_2(x, y)$

$$\begin{aligned} I_2 &\equiv \left| I_1 \star \psi^{j_2,l_2} \right| \\ &= \left| |I_0 \star \psi^{j_1,l_1}| \star \psi^{j_2,l_2} \right| , \end{aligned} \tag{2}$$

where I_2 represents a group of fields labelled by the two sets of wavelet index j_1, l_1 and j_2, l_2. An illustration of these first two orders of scattering transform is shown in Fig. 1. Higher order scattering fields can be created with further iterations. We note that the iterations are not commutative, so maintaining the order of wavelets is important. In this paper, we will only show the scattering transform up to the second order, because we find that in our particular data set, little cosmological information is stored in the third order.

If the input field I_0 is homogeneous, then all the generated fields I_n remain homogeneous. Therefore, the expected values of their intensity can be used as translation-invariant descriptors of the input field

$$S_0 \equiv \langle I_0 \rangle \tag{3}$$

$$S_1^{j_1,l_1} \equiv \langle I_1^{j_1,l_1} \rangle = \langle \left| I_0 \star \psi^{j_1,l_1} \right| \rangle \tag{4}$$

$$S_2^{j_1,l_1,j_2,l_2} \equiv \langle I_2^{j_1,l_1,j_2,l_2} \rangle = \langle \left| |I_0 \star \psi^{j_1,l_1}| \star \psi^{j_2,l_2} \right| \rangle . \tag{5}$$

These expected values S_n are called the nth-order scattering coefficients. Due to homogeneity, these expected scattering coefficients can be estimated by taking the spatial average of a single

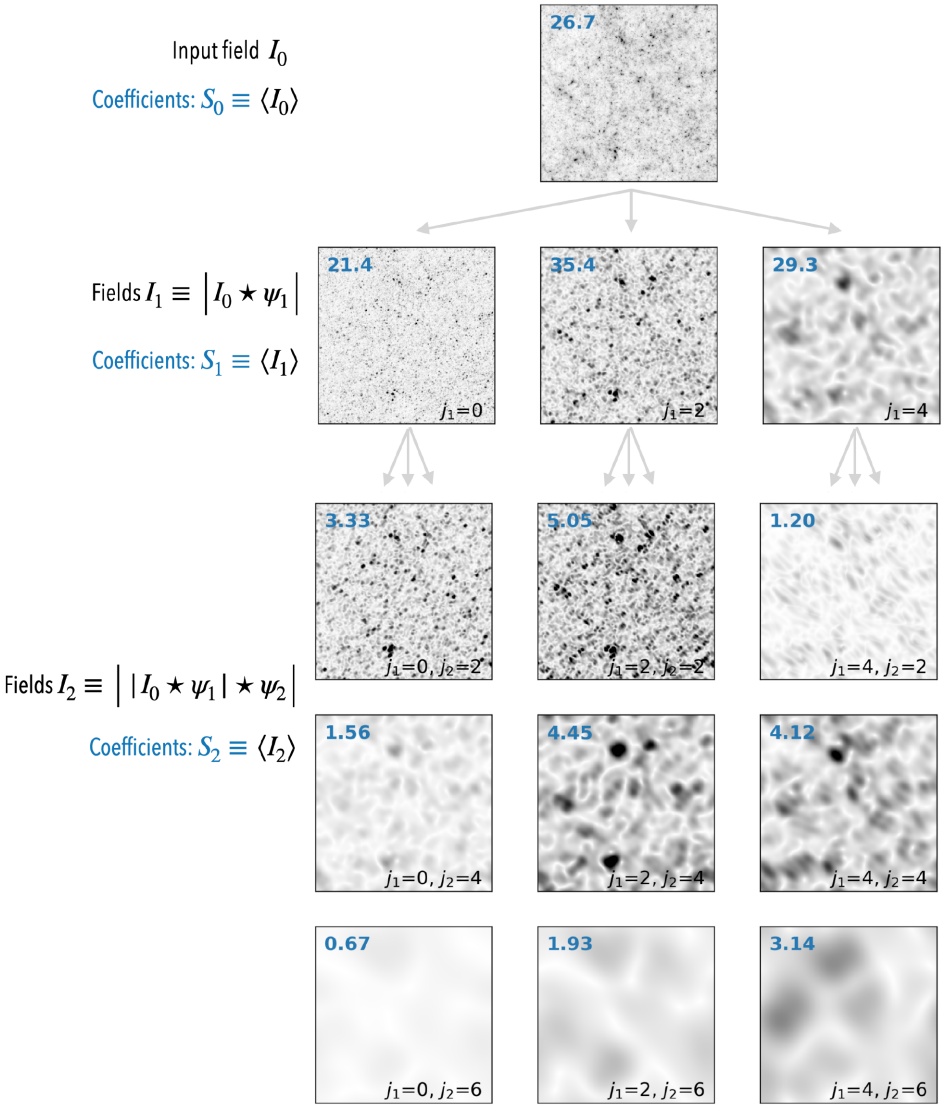

Figure 1. Illustration of the scattering transform on a weak lensing map. The azimuthal resolution is set to be $L = 4$. For clarity, we only show results using wavelets with orientation indices $l_1 = 1$ and $l_2 = 1$, and several selected scale indices j_1 and j_2. In the top left corner of each panel, we show the mean value of that field. They are the scattering coefficients (S_0, S_1, S_2) of the input field. For a convenient display, the blue numbers are 10^4 times the coefficients derived from the lensing map. For example, the S_0 coefficient of this lensing map is actually 0.002 67. The colour bar ranges for I_0, I_1, and I_2 fields are adjusted separately for better visualization.

realization

$$\hat{S}_n = \langle I_n \rangle_{x,y} , \tag{6}$$

where $\hat{S}_n$ is an unbiased estimator of S_n and $\langle \cdot \rangle_{x,y}$ represents the spatial average of a field.[2]

The number of scattering coefficients S_n is determined by the number of wavelets used. Setting J different scales (2^J cannot exceed the side length of the field) and L different orientations results in $J \times L$ different wavelets used in total. If all combinations of wavelets are used, then the number of coefficients at the nth-order will be $J^n L^n$.

[2]To follow the convention in cosmology, we use slightly different notations from Mallat (2012): we use S_n to represent the expected values, which characterize properties of a random field and which Mallat denotes as $\overline{S}_n$; we use $\hat{S}_n$ to represent S_n's estimators calculated from spatial average, which Mallat directly denotes as S_n.

When considering an isotropic field, which is the case of interest in cosmology, the scattering coefficients S_n can be further reduced. To construct isotropic statistics, we simply average over all orientation indices, which reduces the number of coefficients by an order of L^n and creates a more compact and robust set of statistical descriptors. We thus define our reduced scattering coefficients as

$$s_0 \equiv S_0 \tag{7}$$

$$s_1^{j_1} \equiv \langle S_1^{j_1,l_1} \rangle_{l_1} \tag{8}$$

$$s_2^{j_1,j_2} \equiv \langle S_2^{j_1,l_1,j_2,l_2} \rangle_{l_1,l_2} , \tag{9}$$

where S_n represent the standard scattering coefficients, s_n represent our reduced coefficients, and $\langle \rangle_l$ denotes an average over orientation indices. The reduced coefficients s_n can also be understood as the expected value of some 'reduced' fields $\langle I_n \rangle_{l_1,\ldots,l_n}$, which are 'stacks' of the I_n with same scale indices j but different orientation indices

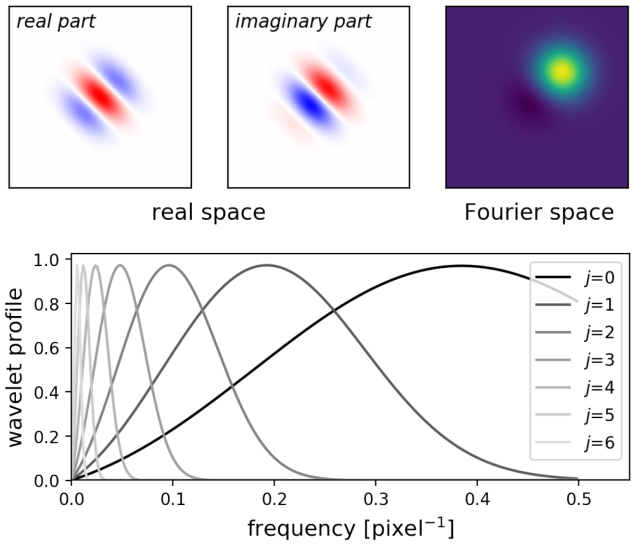

Figure 2. *Upper panel*: profile of a Morlet wavelet ($j = 6, l = 0$, image size 512×512 pixels) in the real space and another one ($j = 1, l = 1$) in Fourier space. The centre of the Fourier space represents zero frequency. *Lower panel*: radial frequency profiles of a family of wavelets. Dilating/contracting (by factor of 2) and rotating (by π/L) one wavelet give the whole family of wavelets used in the scattering transform.

$l_1, .., l_n$. We show several examples of second-order 'reduced' fields in Fig. 3, where information is condensed, so features look clearer than in Fig. 1. Our reduction is similar to the first group of isotropic coefficients used by Allys et al. (2019). Up to the second order, our reduced set includes $1 + J + J^2$ coefficients. As a result, probing the full range of scales for an image with 512×512 pixels ($J = 8$) yields in total 73 reduced scattering coefficients.

In general, performing azimuthal averages over both l_1 and l_2 leads to information loss. To preserve more isotropic information, one could keep $l_2 - l_1$ as an index of the reduced second-order scattering coefficients (Bruna & Mallat 2013; Allys et al. 2019) or apply the 'scattering strategy' again to rotation (Sifre & Mallat 2013). In the weak lensing study presented below, however, we checked that this additional information does not improve the performance of our analysis, probably due to the lack of anisotropic structures in the weak lensing maps we use. So, we do not take it into account.

Having introduced the mathematical formulation of the scattering transform, we will present in the next section some intuitive understanding of its key operations.

2.3 The role of wavelet convolution and modulus

The core operation $I \rightarrow |I \star \psi^{j,l}|$ employed by the scattering transform comprises two steps: a convolution by a complex-valued wavelet and a modulus operation. In short, the wavelet convolution selects scales, and the modulus converts fluctuations into their local strength.

Let us discuss the wavelet convolution first. As a wavelet is a band-pass filter, the wavelet convolution selects Fourier modes around a central frequency and coarsely separates information of different scales (see Fig. 2). Due to the locality of wavelets in real space, which is related to their logarithmic spacing and widths in Fourier space, the scattering coefficients are Lipschitz continuous to deformation, meaning that similar fields differing by a small deformation (including a small dilation) are also similar in the representation formed by scattering coefficients (Mallat 2012), and therefore the scattering characterization is a stable one. Fourier coefficients (without binning), in contrast, are not stable to deformation at high frequencies.

One key idea of the scattering transform is to generate 'first-order' statistics, in contrast to higher order moments, which multiply an increasing number of field intensities and cause instability to outliers. Being a linear operator, the wavelet convolution certainly keeps the 'first-order' property. However, for a homogeneous random field, convolution alone cannot extract information beyond the mean of the original field $\langle I \rangle$, because the expected value operator commutes with all linear transformations. Extracting more information requires non-linear operations. For example, in N-point functions, the multiplication of field intensities plays the role. The scattering transform, on the other hand, employs the *modulus* operation, which is a natural choice to preserve the desired property of working with first-order statistics (Mallat 2012).

As the modulus is taken in the real space and is non-linear, its behaviour in Fourier space is not simple. Nevertheless, we collect some intuitive understandings and present them in Appendix B for interested readers.

2.4 Information extraction beyond the power spectrum

There are a number of similarities between the power spectrum and each single iteration of the scattering transform. Indeed, the power spectrum can be defined using the formalism of the first-order scattering coefficients $S_1 = \langle | I_0 \star \psi | \rangle$:

$$P(\boldsymbol{k}) \propto \langle |I_0 \star \psi'|^2 \rangle \text{ with } \psi' = \mathrm{e}^{-i\boldsymbol{k}\cdot\boldsymbol{x}} . \tag{10}$$

The differences between the two estimators S_1 and $P(k)$ are the choice of convolution kernels (wavelets ψ or Fourier modes ψ') and that of the norm (L^1 versus L^2). Therefore, the first-order scattering coefficients have similarity to the power spectrum. Both of them characterize the strength of fluctuations (or clustering) as a function of scale.

However, in the case of the power spectrum, the convolution kernel ($\psi' = \mathrm{e}^{-i\boldsymbol{k}\cdot\boldsymbol{x}}$) is completely de-localized in real space. Thus, the power spectrum's version of I_1 fields ($|I_0 \star \psi'|^2$) lose all spatial information. In contrast, the use of localized wavelets in the scattering transform allows I_1 to preserve spatial information, as shown in Figs 1 and 3. According to the analogy with the power spectrum, the mean of an I_1 field characterizes the average amplitude of Fourier modes selected by the wavelets, whereas the spatial distribution of fluctuations in I_1, missing in the power spectrum analogue, in turn encodes the phase interaction between those Fourier modes. This information can be extracted by applying the scattering operations once again, $I_1 \rightarrow I_2 = |I_1 \star \psi_2| = ||I_0 \star \psi_1| \star \psi_2|$, and then measuring the mean of I_2, i.e. second-order scattering coefficients S_2.

According to the power spectrum analogy, S_2 coefficients resemble the power spectrum of I_1 fields and measure clustering properties on I_1. Because I_1 fields highlight the regions where fluctuations around a scale are stronger, the second-order coefficients can be understood as measuring the clustering of structures highlighted in I_1, i.e. the 'clustering of (clustered) structures'.

This leads to an interesting intuition: we need two points to describe the scale of one structure and an additional two points for another one. Therefore, the second-order scattering coefficients S_2, which measure the clustering of sized structures, include information up to about 4-point. In general, an nth-order scattering coefficient S_n will contain information up to about 2^n-point function of the input field. By this 'hierarchical clustering' design, the scattering-

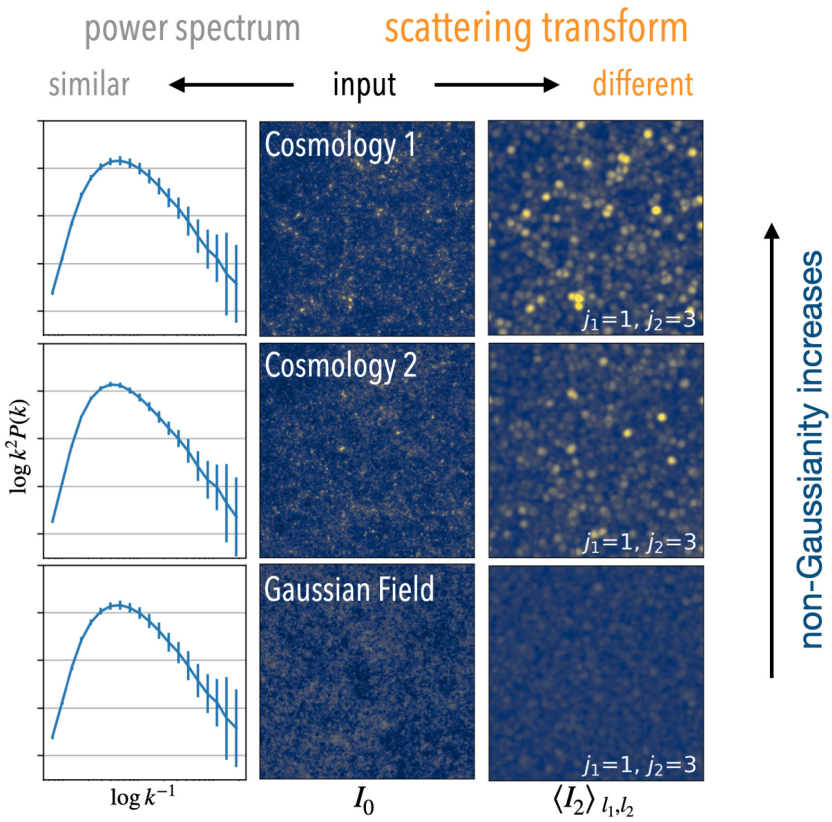

Figure 3. The scattering transform of three fields (I_0) with indistinguishable power spectra. Row 1 shows a realization of convergence maps in cosmology (Ω_m, σ_8) = (0.292, 0.835), row 2 shows cosmology (Ω_m, σ_8) = (0.566, 0.520), and row 3 is for a Gaussian random field with the same (2D) power spectrum as row 1. It can be seen by eye that the average intensity of the second-order scattering fields (the last column), which corresponds to an s_2 coefficient and measures the clustering strength of structures highlighted by I_1, is significantly different from each other, while their power spectra (the first column) are indistinguishable.

transform expansion quickly includes information from higher order statistics.

However, it should be noted that there is still a fundamental difference between the scattering transform and N-point functions. There are mainly two difficulties associated with N-point functions to characterize a random field: the failure to describe distribution tails and the huge number of configurations. The first difficulty, related to the multiplication of multiple random variables, leads to high variances and also prevents the extraction of information from fields whose pdf has a tail (Carron 2011). The scattering transform, which uses modulus and does not enhance the tail, can significantly alleviate this problem. We will discuss it further in another paper (Cheng et al. in preparation). The second difficulty may be alleviated by an efficient binning. For example, the hierarchical wavelet transform used in the scattering transform *is* a binning strategy that can also be applied to N-point functions (see Appendix B).

3 APPLICATION IN WEAK LENSING COSMOLOGY

We now show that the scattering transform can be a powerful tool in observational cosmology to extract non-Gaussian information from the matter density field. To illustrate this point, we consider an application with two-dimensional fields: we show how well cosmological parameters can be constrained using the scattering coefficients of weak lensing convergence maps $\kappa(\vec{\theta})$ or, equivalently, measurements of cosmic shear. Being projections of the density field along the line of sight, these maps present an appreciable level of non-Gaussianities on scales smaller than a few degrees, reflecting the non-linear growth of matter fluctuations. For the necessary background on cosmology with gravitational lensing, we refer the reader to reviews (Kilbinger 2015; Mandelbaum 2018).

We explore the use of our reduced scattering coefficients on simulated weak lensing convergence maps to infer Ω_m and σ_8 and compare their performance with that of the power spectrum. We also compare our results with that of a state-of-the-art CNNs by Ribli et al. (2019b) and peak count statistics.

3.1 Simulated convergence maps

We use mock convergence maps in the 'Dark Matter' dataset generated by the Columbia Lensing team[3] and described in Zorrilla Matilla et al. (2016) and Gupta et al. (2018). The maps are produced through ray-tracing to redshift $z = 1$ in the output of dark matter-only N-body simulations for a set of ΛCDM cosmologies. Each simulation is run in a 240 h^{-1} Mpc box with 512^3 particles. The cosmologies differ only in two parameters: the present matter density relative to the critical density Ω_m, and a normalization of the power spectrum σ_8. Other cosmological parameters are fixed: baryon density $\Omega_b = 0.046$, Hubble constant $h = 0.72$, scalar spectral index $n_s = 0.96$, effective number of relativistic degrees of freedom $n_{eff} = 3.04$, and neutrino masses $m_\nu = 0.0$. The dark energy density is set so that the universe is spatially flat, i.e. $\Omega_\Lambda = 1 - \Omega_m$. For each cosmology, 512

[3] http://columbialensing.org

convergence maps with 3.5 × 3.5 deg^2 field of view are generated from the simulations, allowing us to sample cosmic variance. The corresponding scales are well suited to probing the non-Gaussianities of the convergence field (Kilbinger 2015). These maps were also used by Ribli et al. (2019b). To compare our results to Ribli et al. (2019b), we use the same resolution as theirs, down-sampling the original 1024^2 pixel maps to a 512^2 resolution with 0.41 arcmin per pixel.

3.2 Galaxy shape noise and smoothing

In practice, convergence or shear estimates are obtained from measurements of galaxy shapes, with a level of noise that depends on the galaxy ellipticity distribution and their number density on the sky. To first order, background galaxies used for shear measurements have a wide range of redshifts and are not correlated. The noise can be well approximated as Gaussian white noise. Its contribution to the convergence maps can be modelled (van Waerbeke 2000) as

$$\sigma^2_{\rm noise} = \frac{\sigma^2_\epsilon}{2 n_{\rm g} A_{\rm pix}}\,, \tag{11}$$

where σ^2_ϵ is the intrinsic variance of ellipticity of galaxies, which is taken to be 0.4^2, $n_{\rm g}$ is the number density of background galaxies, $A_{\rm pix}$ is the area per pixel, which is 0.1682 arcmin^2. For some existing and on-going surveys such as CFHTLenS, KiDS,[4] and DES,[5] $n_{\rm g}$ is around 10 arcmin^{-2} (Kilbinger et al. 2013; Abbott et al. 2018); for some upcoming surveys we expect substantially higher densities: $n_{\rm g} \sim 25$ arcmin^{-2} for the survey at Vera C. Rubin Observatory (LSST) ,[6] $n_{\rm g} > 30$ arcmin^{-2} for *Euclid*,[7] and $n_{\rm g} \sim 50$–75 arcmin^{-2} for the planned survey with Nancy Grace Roman Space Telescope (*WFIRST*).[8]

After adding noise, we also smooth the maps. As the power of Gaussian white noise is distributed mostly at high frequencies, smoothing the convergence maps can help to increase the signal-to-noise of specific estimators. By default, we do not smooth the noiseless maps, and we perform a $\sigma = 1$ arcmin (2.44 pixel) Gaussian smoothing on noisy maps.

3.3 Statistical descriptors

Scattering coefficients: For each 3.5 × 3.5 deg^2 convergence field in each cosmology, we apply the scattering transform up to second order using the 'kymatio' PYTHON package[9] (Andreux et al. 2020) and then calculate the reduced coefficients (s_0, s_1, s_2) as defined in Section 2.2. To probe the available range of scales, we set $J = 8$ and $L = 4$ in the scattering transform, i.e. we use 8 scales spaced logarithmically with central wavelengths between 1.2 and 75 arcmin and 4 azimuthal orientations, resulting in 32 different wavelets used in total.

By default, the 'kymatio' package only calculate the second-order coefficients with $j_2 > j_1$, because the coefficients with $j_2 \leq j_1$ is mainly determined by the property of wavelets but not the input field, as illustrated by the upper-right sketch of Fig. 4. Intuitively, this is because structures of a particular size, say j_1, do not have meaningful clustering at scales smaller than their own size. A mathematical reasoning for this property can also be found in Appendix B. To demonstrate these coefficients' behaviour, we modified the 'kymatio' code to calculate them, and show them together with the coefficients with $j_2 > j_1$ in Fig. 4. Nevertheless, we checked that they do not contribute to constraining cosmological parameters, and therefore in our inference analysis, we only use second-order coefficients with $j_2 > j_1$, which yields an even more compact set of $1 + 8 + 28 = 37$ scattering coefficients used for our cosmological inference.

Power spectrum: For the same set of input fields, we also compute the power spectrum and peak count statistics using the publicly available 'LensTools' PYTHON package[10] (Petri 2016). The power spectrum is calculated within 20 bins in the range $100 \leq l \leq 37\,500$ (corresponding to 0.58–216 arcmin) with logarithmic spacing, following the setting adopted in Ribli et al. (2019b).

Peak count: In our analysis, a peak is defined as a pixel with higher convergence (κ) than its eight neighbours. Then, peaks are binned by their κ values and counted in each bin. We adopt a binning similar to that in Liu et al. (2015a). We use 20 bins in total, including 18 bins linearly spaced between $\kappa = -0.02$ and 0.12, one bin for peaks below –0.02, and one bin above 0.12. For reference, $\kappa = 0.12$ corresponds to a significance of peak $\nu \equiv \kappa/\sigma_{\rm noise}$ around 7 when $n_{\rm g} = 30$. Although using more bins for very high peaks ($\kappa > 0.12$) may enhance the constraining power of the peak count method, we do not use them in this study, because the count distribution of these rare peaks can no longer be approximated by Gaussian distribution (see e.g. Lin & Kilbinger 2015).

To obtain constraints on the cosmological parameters, we use the Fisher inference framework (Fisher 1935; Tegmark, Taylor & Heavens 1997), in which we assume the probability distribution of statistical descriptors is a multivariate Gaussian distribution for a given cosmology. The mean vector and covariance matrix of this Gaussian distribution are dependent on cosmological parameters and estimated from the 512 realizations of each cosmology in simulations. Details of our cosmological inference framework are described in Appendix C. Because s_1, s_2, and power spectra must be positive for a non-trivial field, we consider their logarithm to better satisfy a multivariate Gaussian likelihood. To perform the cosmological inference analysis with the three methods introduced above, we use

(i) 37 scattering coefficients.
(ii) 20 power spectrum coefficients.
(iii) 20 peak count coefficients.

4 RESULTS

In this section, we examine the distribution and cosmological sensitivity of scattering coefficients, and present their constraining power for two cosmological parameters, $\Omega_{\rm m}$ and σ_8. We show that the scattering coefficients provide substantially more information than the power spectrum and is on a par with CNN.

4.1 Cosmological sensitivity of the scattering coefficients

In Fig. 4, we present the distributions of reduce scattering transform in the noiseless case together with the power spectrum. In the first row, we show the values for a fiducial cosmology that has the Planck cosmology of $\Omega_{\rm m} = 0.309$ and $\sigma_8 = 0.816$ (Planck Collaboration et al. 2016). The expected values of these descriptors are estimated by

[4] http://kids.strw.leidenuniv.nl
[5] https://www.darkenergysurvey.org
[6] https://www.lsst.org
[7] https://sci.esa.int/euclid
[8] https://roman.gsfc.nasa.gov
[9] https://www.kymat.io
[10] https://lenstools.readthedocs.io

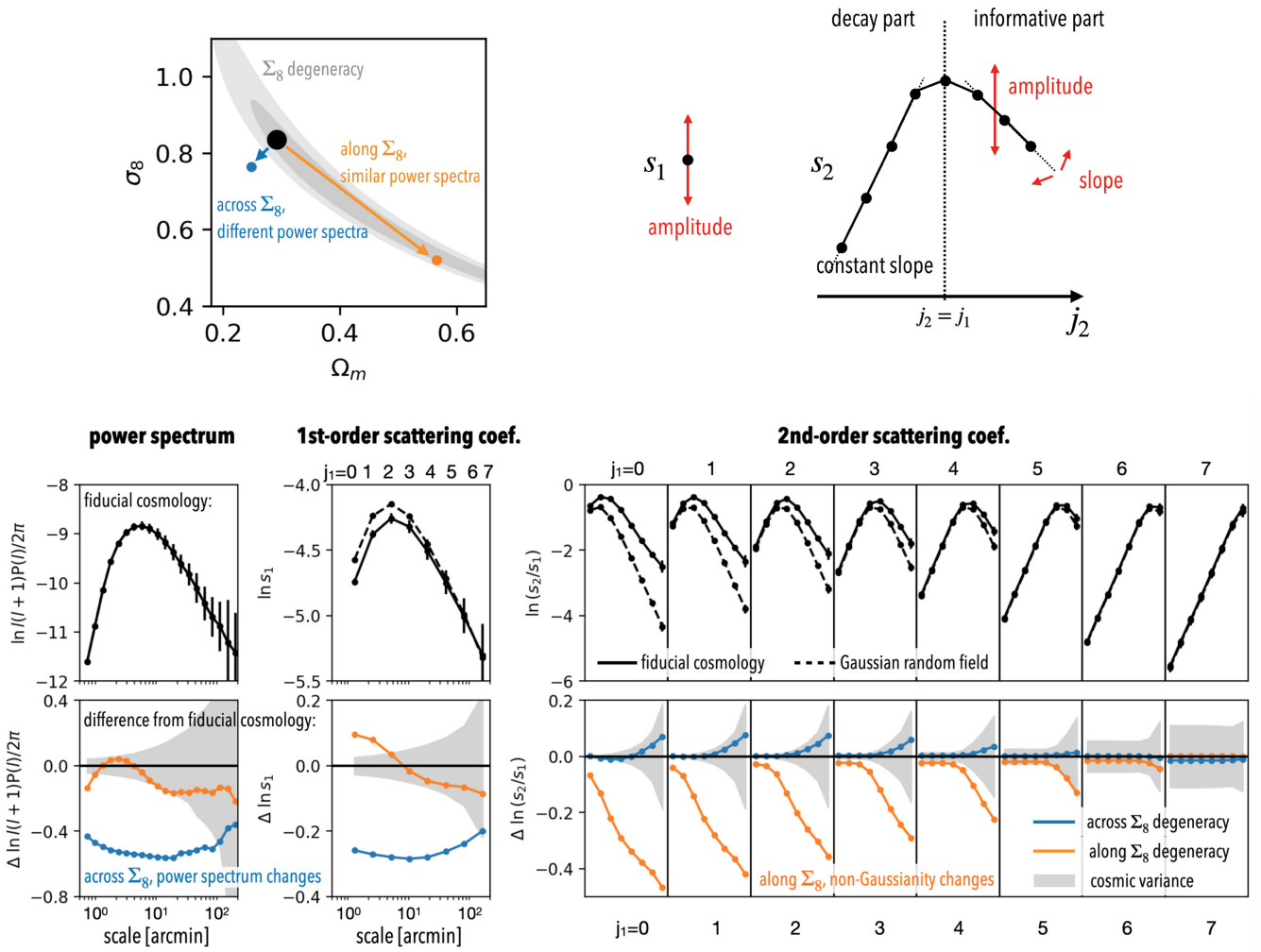

Figure 4. *Upper-left panel*: The fiducial cosmology (black) and two other cosmologies on the (Ω_m, σ_8) plane. *Upper-right panel*: Illustration of reduced scattering coefficients $s_1(j_1)$ and $s_2(j_1, j_2)$ for a single j_1 scale. *Lower panel*: The power spectrum and scattering coefficients for the three cosmologies in noiseless case. The first row presents coefficients of the fiducial cosmology and of Gaussian random fields with the same power spectrum, and the second row shows changes of coefficients (Δ coef.) when we move from the fiducial cosmology to the other two. Error bars and grey shaded regions show cosmic variance, i.e. the variability among realizations. The first-order scattering coefficients behave similarly to the power spectrum, while the second-order scattering coefficients can break the Σ_8 degeneracy, along which non-Gaussianity of weak lensing field changes.

averaging over different realizations of a given cosmology. Error bars, which are the sample standard deviations of realizations, represent the cosmic variance in this noiseless case. We can see the similarity between the power spectrum and s_1 coefficients, as they have similar physical meanings (Section 2.4). We can also see the different behaviours of s_2 coefficients for $j_2 < j_1$ and $j_2 > j_1$, as discussed in Section 3.3.

Then, we investigate the cosmological sensitivity of the power spectrum and scattering coefficients. The power spectrum is known to be mostly sensitive to one combination of the cosmological parameters, namely

$$\Sigma_8 \equiv \sigma_8 \left(\frac{\Omega_m}{0.3}\right)^a , \tag{12}$$

with a around 0.6 (see e.g. Kilbinger 2015), but can hardly distinguish cosmologies with the same Σ_8, as illustrated in the upper-left panel of Fig. 4. Breaking this degeneracy requires the extraction of non-Gaussian information from lensing maps.

In the second row of Fig. 4, we show the response of coefficients as cosmological parameters move along (orange curves) and across (blue curves) the Σ_8 degeneracy. Grey areas indicate cosmic variance of the fiducial cosmology. As expected, the first-order scattering coefficients show a cosmological sensitivity similar to that of the power spectrum, because both of them measure the strength of fluctuations as a function of scale.

The second-order scattering coefficients, on the other hand, characterize the spatial distribution of sized fluctuations. To make the second-order scattering coefficients less correlated with the first-order ones, here we present de-correlated second-order coefficients s_2/s_1, as each $s_2(j_1, j_2)$ is proportional to the corresponding $s_1(j_1)$ according to their definitions (Bruna et al. 2015). These s_2/s_1 exhibit particularly high sensitivity to cosmological change along the Σ_8 degeneracy. In addition, they are indifferent to the other direction of cosmological change, which means they provide a piece of information roughly orthogonal to that carried by the first-order coefficients s_1 or the power spectrum. In noisy cases, though the information from s_2/s_1 is not orthogonal to s_1 anymore, we have

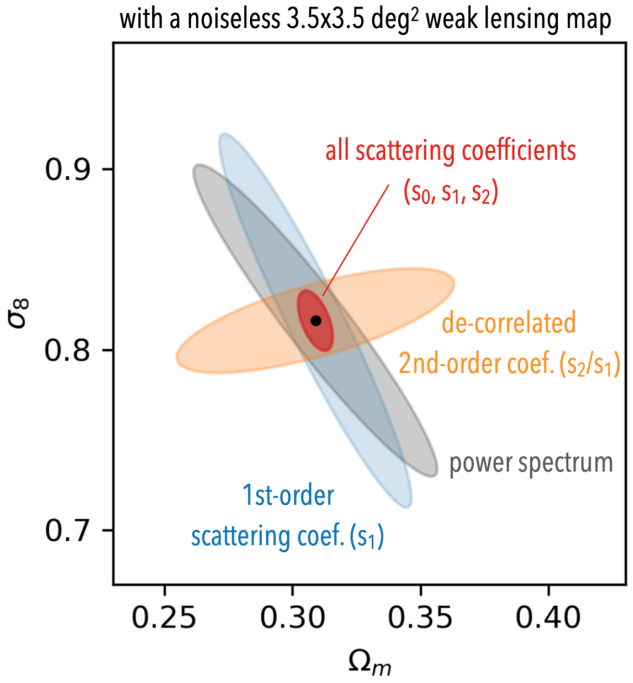

Figure 5. The 1σ Fisher forecast of cosmological parameters from a noiseless convergence map (3.5×3.5 deg^2, with 0.41 arcmin per pixel resolution). The de-correlated first-order scattering coefficients s_2/s_1 provide critical information to break the Σ_8 degeneracy along which the power spectrum cannot distinguish, therefore drastically improve the constraint.

checked that they still provide substantial sensitivity along the Σ_8 degeneracy. Due to this additional sensitivity, the scattering transform can be used to better constrain cosmological parameters than the power spectrum.

4.2 Constraining cosmological parameters

We now present the cosmological constraints set by the scattering coefficients measured from a single 3.5×3.5 deg^2 field. For reference, we note that LSST will generate about 2,000 times more data, leading to constraints about 40 times tighter than the numbers presented below. In this study, we only probe the constraints on $\Omega_{\rm m}$ and σ_8 and leave the work of using scattering coefficients to constrain the dark energy equation of state parameter w or neutrino mass m_ν to future study. Cosmological inference is just another aspect of the cosmological sensitivity problem examined in the previous subsection. The Fisher inference formalism we use in this study is described in Appendix C.

We first present results in the noiseless case. In Fig. 5, we demonstrate the 1σ Fisher forecast of $\Omega_{\rm m}$ and σ_8 using all scattering coefficients (red ellipse) and power spectrum (grey ellipse). The scattering coefficients provide a dramatically tighter constraint than the power spectrum. We also show a break-down of this constraining power into contributions from first-order (blue ellipse) and second-order (orange ellipse) coefficients alone. As expected, the first-order coefficients (s_1) and power spectrum set similar constraints. The slight difference of ellipse orientation originates from the difference between the L^1 and L^2 norms used by the scattering transform and the power spectrum. The de-correlated second-order scattering coefficients (s_2/s_1) provide a strong constraint along the Σ_8 degeneracy, consistent with our cosmological sensitivity discussion in Section 4.1.

The zeroth-order coefficient s_0 is the mean of the 3.5×3.5 deg^2 field. While its expectation value over the sky is zero, it does carry relevant information on those scales by capturing larger scale modulations of the convergence field. We also note that it has strong correlations with other scattering coefficients (and the power spectrum), which is a sign of being in the non-linear regime of cosmology (see e.g. Li, Dodelson & Croft 2020). Therefore, although the expected value of s_0 is identically zero in all cosmology, combining s_0 with other coefficients helps to substantially tighten the constraints on cosmological parameters. However, this piece of information may not scale as fast with the increasing field of view as the small-scale information, because in real data each patch of 3.5×3.5 deg^2 fields on the sky are not independent. The mass sheet degeneracy (see e.g. Bradač, Lombardi & Schneider 2004) is another problem for using s_0, though the s_0 of small patches may be obtained by inheriting the zero-point solution of the whole survey. We find that including s_0 only improves the constraint of Σ_8, consistent with the understanding that it is a leakage of larger scale fluctuation. Similar improvement is also found when combining s_0 with the power spectrum.

To be more quantitative, we compare different methods using the reciprocal of the area of their 1σ Fisher forecast ellipses on the ($\Omega_{\rm m}$, σ_8) plane as the figure of merit (FoM). In the noiseless case, combining all scattering coefficients (s_0, s_1, s_2) leads to a constraint that is 14 times tighter than that of the power spectrum, 5 times tighter than peak count statistics, and 3.3 times tighter than the joint constraint from power spectrum and peak count.

We then compare the performance of the scattering transform to a state-of-the-art CNN analysis by Ribli et al. (2019b). To perform a meaningful comparison, we follow Ribli et al. (2019b) to use noiseless convergence maps smoothed with a $\sigma = 1$ arcmin Gaussian filter. Interestingly, we find that the scattering coefficients extract a similar amount of cosmological information to the CNN trained in Ribli et al. (2019b). The corresponding figures of merit are shown in Table 1.[11]

We now consider convergence fields in the presence of galaxy shape noise. As the noise level increases, small-scale structures, which carry plenty of cosmological information, get erased. As a result, the constraining power of the scattering coefficients (as well as other methods) degrades. In Fig. 6, we show the Fisher forecast of $\Omega_{\rm m}$ and σ_8 from a 3.5×3.5 deg^2 convergence map under three noise levels, using the scattering coefficients and the power spectrum. We also show the posterior constraints from CNNs trained by Ribli et al. (2019b) on the same simulations. The figures of merit for these methods, together with the peak count method, are listed in Table 1. Again, we find that the scattering transform not only outperforms the power spectrum and peak count, but also provides cosmological constraints on a par with the state-of-the-art CNNs.

To summarize, we have demonstrated the power of the scattering transform for cosmological parameter inference with weak lensing data. For simplicity, we focused on the convergence field but a similar analysis can also be performed on the shear field. In Fig. 7, we present quantitative comparisons between the four techniques discussed in our study. It shows the high performance of the scattering transform over a wide range of noise levels. We therefore advocate using this new estimator in the analysis of existing and upcoming weak lensing surveys, in observational cosmology, and

[11] We note that Ribli et al. (2019b) do include the field mean information in their CNN training. So, a fair comparison would be $s_1 + s_2$ versus power spectrum, and $s_0 + s_1 + s_2$ versus CNN.

Table 1. Comparison of the constraining power for (Ω_m, σ_8) between different methods, with a single 3.5 × 3.5 deg^2 convergence map. The FoM is defined as the reciprocal of the 1σ confident area based on Fisher matrix (or the 68 per cent posterior contour, in parentheses) on the (Ω_m, σ_8) plane. The convergence maps are smoothed with $\sigma = 1$ arcmin Gaussian filter except for the case shown in the last column with no smoothing.

Methods	Ω_m–σ_8 FoM				
	$n_g = 10$ arcmin^{-2}	$n_g = 30$ arcmin^{-2}	$n_g = 100$ arcmin^{-2}	Noiseless	Noiseless (no smoothing)
Scattering transform: $s_0 + s_1 + s_2$	50	140	329	1053	3367
Scattering transform: $s_0 + s_1$	21	55	133	492	565
Scattering transform: $s_1 + s_2$	39	91	181	446	1720
Power spectrum P(l)	20	40	67	104	253
Peak count	30	89	162	170	667
CNN (Ribli et al. 2019b)	(44)	(121)	(292)	(1201)	(-)

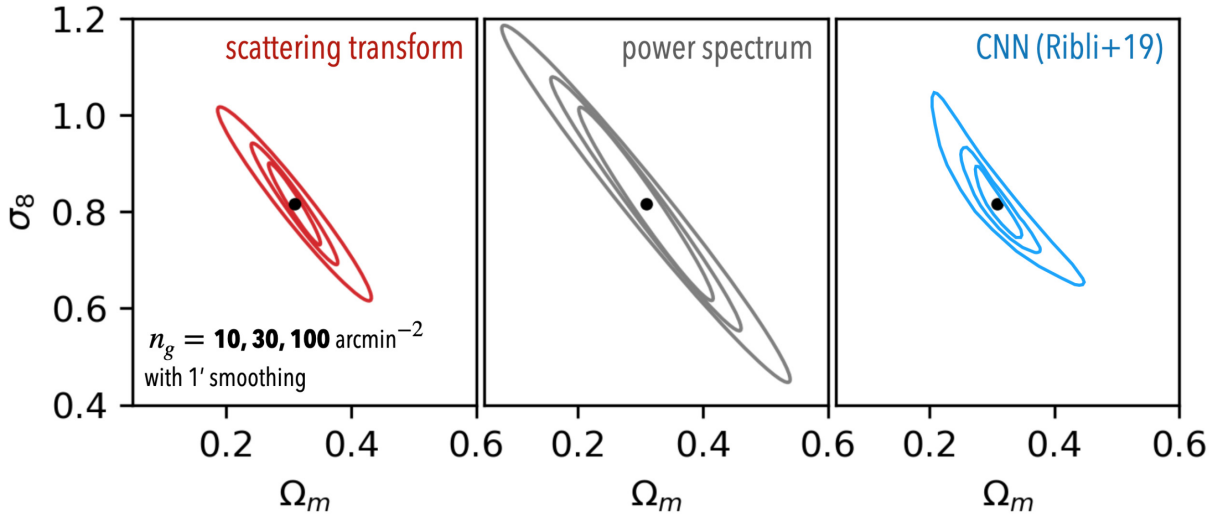

Figure 6. Comparison of different estimators in noisy cases. Ellipses are 1σ Fisher forecast (or posterior, in the CNN case) of cosmological parameters (Ω_m and σ_8) from noisy 3.5 × 3.5 deg^2 convergence maps smoothed with $\sigma = 1$ arcmin Gaussian filter. The scattering coefficients have comparable performance as a state-of-the-art CNN (Ribli et al. 2019b) at all noise levels, and 3–5 times better than the power spectrum depending on the noise level.

more generally, in the analysis of stochastic fields encountered in physics.

5 DISCUSSION

5.1 Inference for non-Gaussian fields

In physics, many inference problems concern estimating physical parameters from realizations of random fields. Ideally, one would like to use the likelihood function of the field itself, but this is often out of reach except for several simple cases such as some Gaussian random fields. Therefore, for the inference problem to be feasible, a statistical representation of the data is often used. Statistical descriptors reduce the dimensionality of the data vector and they tend to Gaussianize according to the central limit theorem. Both of these properties help to regularize the likelihood. However, it is still challenging to find a proper representation because in general a random field can be random in too many different ways. In these cases, a useful characterization must be one that makes use of known properties of the field.

Viewed in this direction, traditional statistical approaches with their own representation framework may or may not suit the properties of particular fields. For example, the peak count statistic used in weak lensing cosmology suits the presence of distinct haloes in convergence maps. N-point functions, closely related to perturbation theory and convenient for analytical prediction, represent the field with a series expansion, which makes them good descriptors for fields slightly deviating from a Gaussian one. A highly non-Gaussian field, however, requires using larger N. As the number of coefficients and the complexity of configurations increase rapidly with N, N-point functions quickly become an inefficient and non-robust representation of the input field. On the other hand, CNNs try to learn the field properties and search for informative representation through a training optimization.

Fortunately, the non-Gaussian fields that originate from physical interactions do often have common properties. Such fields typically display localized, coherent structures in multiscales, and smaller structures often act as building blocks of larger structures. These properties can be used as the 'domain knowledge' to guide our design or choice of the statistical representation in a general sense. As we will explain in the next section, the design and operations of the scattering transform lead to an efficient and robust representation for such fields, because they are tailored for these properties.

5.2 Attractive properties of the scattering transform

Efficiency: All the three elements (wavelet convolution, modulus, and the hierarchical design) play essential roles to make the scattering transform efficient. The use of wavelets balances the resolution in real and frequency domain. As a result, the scattering transform can capture localized information from a large range of scales with only a few coefficients, at each order. After selecting structures of scale j_n in

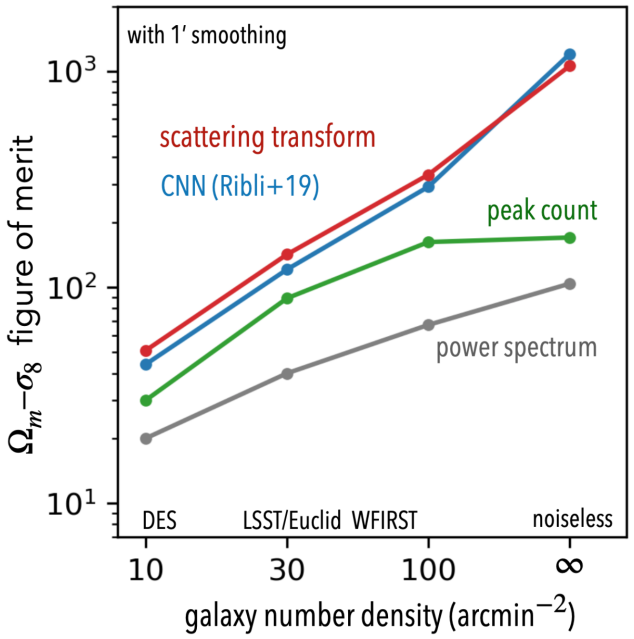

Figure 7. Noise dependence of the (Ω_m, σ_8) constraints with different methods. The FoM is defined as the 1σ confident area on the (Ω_m, σ_8) plane. Note that the CNN result (Ribli et al. 2019b) is reported in terms of posterior, while others are Fisher forecast. For noisy cases, the difference between scattering transform and CNN is not intrinsic but due to the difference between posterior and Fisher forecast. Peak count's performance does not increase as fast because it is more sensitive to smoothing scale than the other methods.

one order, the scattering transform then selects structures 'assembled' by these j_n-scale structures in the next order. This hierarchical design allows the nth order scattering coefficients to quickly access configurations described by about 2^n points. Moreover, the 'low-order' non-linear operator, modulus, helps to collect information even beyond the access of 2^n point functions. We will discuss it further in another paper (Cheng et al. in preparation).

These strategies concentrate relevant information to a limited set of statistical descriptors, which is desirable in terms of compactness of the representation and the signal-to-noise ratio of each estimator. For example, in our case, the scattering transform compresses weak lensing information into 37 coefficients, a number that is much smaller than typical bi-spectrum descriptors, while achieving CNN-like constraint on cosmological parameters.

Robustness: All scattering coefficients are 'first-order' statistics in the sense that they are proportional to the input field, and it is proved that the scattering representation is non-expansive, i.e. the distance between two vectors in the scattering representation never exceeds their distance in the original pixel-based representation (Mallat 2012; Bruna & Mallat 2013). Therefore, it does not amplify the process variability. This is in contrast to the N-point correlation function approach, which requires multiplying an increasing number of field fluctuations and causes high variability. As a result, the scattering coefficients are low variance descriptors and insensitive to outliers.

The locality of wavelets, which is related to their logarithmic spacing and widths in frequency space, also introduces stability to deformations (Mallat 2012), which is a desired property of robust descriptors that classical N-point functions do not have.

Interpretability: As discussed in Section 2.4, the scattering coefficients have a simple and intuitive interpretation. They describe clustering properties of the field in the following way:

The first-order scattering coefficients are similar to a coarsely binned power spectrum, which characterize the clustering strength at different scales j_1. As the scattering transform uses an L^1 norm as opposed to an L^2 norm, the ratio between s_1 coefficients and the power spectrum provides a measure of sparsity of the field. This explains why in Fig. 5 the constraints from first-order coefficients and the power spectrum are slightly different, and just combining these two can also provide a stronger constraint on cosmology than using power spectrum alone.

The second-order scattering coefficients characterize the clustering strength of j_1-scale structures separated by j_2-scales. In other words, these coefficients characterize the clustering of structures selected over a given frequency range, or the 'clustering of clustering'. Their departure from their Gaussian counterparts is a robust measure of the strength of non-Gaussianities. The nth-order scattering coefficients, though not shown explicitly in this study, can in turn be understood as the strengths of nth-order hierarchy of clustering of the field at all different combinations of scales.

5.3 Comparison to CNNs

The scattering transform and CNNs share a number of properties. Both of them have hierarchical layers with localized convolution kernels and use a simple non-expansive non-linear operation. Although CNNs are usually trained to directly map a field to physical parameters, their inside can be considered as composed of a convolutional part that extracts spatial features and a second part that learns the mapping from these features to physical parameters. Both parts are trainable and trained together. The scattering transform, on the other hand, uses preset wavelets as convolutional kernels and just a few layers (in our case two layers). So it can be viewed as a non-trainable mini-CNN playing the role of the first part of trainable CNNs. In the scattering transform's approach, the second part of trainable CNNs is supplanted by using traditional regression techniques.

The trainable kernels make CNNs more flexible and may lead to a higher performance for finer classification problems such as classifying different types of rabbits, but in the mean time this overparametrization defines a much more brittle statistical model (Szegedy et al. 2013; Bruna & Mallat 2019). Our results imply that compared to CNNs, the scattering transform has enough expressiveness to characterize the matter density field in the cosmological context while holding provable stability properties. Indeed, as shown by Ribli et al. (2019b), a CNN trained on convergence maps internally generates kernels similar to (azimuthally averaged) Morlet wavelets. Our results also imply that much of the power of CNNs may be detached from its trainable nature.

Overparametrized models tend to overfit, i.e. to 'remember' single realizations instead of comprehending the overall property of the whole training set. Thus the overparametrized CNNs require a large number of simulations as training set to alleviate the overfitting problem. In contrast, the scattering transform uses preset kernels, thus has no parametrization in the kernels. In addition, the choice of CNN architecture can modify the results substantially, as can be seen in the comparison between results of Ribli et al. (2019b) and Gupta et al. (2018). As such, CNNs usually require much, and often ad hoc, fine-tuning. The scattering transform, on the other hand, is not subject to these sources of variability. It requires the use of simulations only to probe the cosmic variance of the descriptors. Without learning the

kernels, the scattering transform also significantly save calculation time.

Another view on the overfitting problem is given by the framework of maximum-entropy regularized estimation, which looks for the most 'non-committal' statistical model under the constraints of a 'feature vector' of sufficient statistics (Jaynes 1957). There is thus a tension in the design of such vector of sufficient statistics (Bruna & Mallat 2019): On the one hand, the features should be descriptive enough so that they introduce enough constraints, i.e. typical samples from the estimated model should also be typical in the true distribution; On the other hand, one would like the features to be efficiently estimated from the available samples, so that the corresponding statistical model is robust under resampling. In other words, typical samples from the true distribution should remain typical under the estimated statistical model.

Finally, when applied to observational data, the scattering transform holds another advantage over CNNs, namely the possibility to investigate systematic effects. As traditional statistics, the scattering coefficients can be used to derive not only the best-fitting cosmological parameters, but also an evaluation of the goodness of fit and therefore a sanity check of the result. In contrast, although the internal machinery of CNNs can be roughly divided into a feature extraction part and a regression one, the CNNs are trained as a whole to learn a direct mapping from the data to the physical parameters. Due to the overparametrization nature, outputs from intermediate layers (i.e. the intermediate abstraction of CNN) do not typically have good statistical properties. Therefore, when using CNN, it is challenging to check for systematic error in real data.

5.4 Relation to peak count method

The non-linear gravitational evolution of density fluctuations in the universe gives rise to haloes, which are virialized systems locally bound by gravity. As highlighted by Ribli et al. (2019b) in their fig. 10, a substantial amount of non-Gaussian cosmological information can be extracted from these features. The peak count method directly captures information in the abundance of haloes. However, it does not characterize the spatial information including profiles and positions of these haloes, which is also sensitive to cosmological parameters. The scattering transform implicitly extracts a comprehensive information of the abundance, profile, and distribution of haloes by first highlighting structures of particular scales and then characterizing their clustering at other scales, as described in Section 2.4. In the limit of small j_1 and large j_2, the second-order scattering coefficients can be understood as a measure of the 'two-halo term' in the halo model at scale j_2, weighted by the halo response to the first wavelet with scale j_1. This response is related to halo profiles. In general, the scattering transform provides a non-parametric description of the one-halo, two-halo, and transitional regime where haloes overlap and form larger haloes.

6 CONCLUSION

Characterizing arbitrary non-Gaussian fields is challenging as the dimensionality of their description can be arbitrarily high. The subset of fields relevant in physics, however, tends to be more constrained as they typically display localized, coherent structures. In the cosmological context, the matter density field presents another characteristic property, namely hierarchical clustering. An efficient statistical descriptor of the cosmological density field would ideally make use of these properties.

In this paper, we advocate the use of the scattering transform (Mallat 2012; Bruna & Mallat 2013), which generates statistics designed to extract information from complex fields with provable stability properties. It involves operations similar to those found in CNNs: it uses wavelet convolution, which is particularly suitable for characterizing localized structures; it uses modulus as the non-linear operation; and it iterates these operations. However, in contrast to CNNs, the scattering transform does *not* require training. It generates a compact set of robust coefficients, which forms a representation of the input field and can be used as efficient summary statistics for non-Gaussian information.

We applied the scattering transform to a parameter inference problem in the context of weak lensing cosmology. For simplicity, we focused on the convergence field but a similar analysis can also be performed on the shear field. We used simulated convergence maps generated by ray-tracing *N*-body simulation results (Zorrilla Matilla et al. 2016; Gupta et al. 2018) and measured their scattering coefficients to infer the cosmological parameters $\Omega_{\rm m}$ and σ_8. On maps with and without galaxy shape noise, the scattering transform outperforms the power spectrum and peak counts, and is on a par with the state-of-the-art CNNs.

As described in Section 5.2, the scattering transform possesses a series of attractive properties for parameter estimation. It is efficient, robust, and interpretable. Obtained by iteratively applying wavelet convolution and modulus and finally taking the expectation value, the scattering coefficients can be interpreted as the strength of a hierarchy of clustering at various combinations of scales. Different from *N*-point functions, all scattering coefficients have the welcome property that they remain proportional to the input field, thus avoid instability problems and extract much more information when the field distribution has a long tail. Similar to classic statistical estimators, the scattering transform requires no training or tuning and offers the possibility to investigate systematic errors potentially present with real data.

In this paper we demonstrated applications of the scattering transform in weak lensing data. Using it with existing and upcoming surveys (see e.g. DES, LSST, *Euclid*, *WFIRST*) can be of great interest to improve constraints and provide consistency checks. Based on its properties and design, the scattering transform can also be an attractive approach for many other applications: in observational cosmology, astrophysics, and beyond.

ACKNOWLEDGEMENTS

We thank the anonymous referee, Jean-François Cardoso, Yi-Kuan Chiang, and Zuhui Fan for useful comments. We also thank Dezső Ribli for discussions. We thank the Columbia Lensing group (http://columbialensing.org) for making their suite of simulated maps available, and NSF for supporting the creation of those maps through grant AST-1210877 and XSEDE allocation AST-140041. YST is supported by the NASA Hubble Fellowship grant *HST*-HF2-51425.001 awarded by the Space Telescope Science Institute. This work is partially supported by the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091, Samsung Electronics, and the Institute for Advanced Study. SC thanks Siyu Yao for her constant encouragement and inspiration.

DATA AVAILABILITY

The data underlying this article were accessed from the Columbia Lensing group (http://columbialensing.org). The derived data gen-

erated in this research will be shared on reasonable request to the corresponding author.

REFERENCES

Abbott T. M. C. et al., 2018, ApJS, 239, 18
Allys E., Levrier F., Zhang S., Colling C., Regaldo-Saint Blancard B., Boulanger F., Hennebelle P., Mallat S., 2019, A&A, 629, A115
Allys E., Marchand T., Cardoso J. F., Villaescusa-Navarro F., Ho S., Mallat S., 2020, preprint (arXiv:2006.06298)
Andén J., Mallat S., 2014, IEEE Trans. Signal. Process., 62, 4114
Andreux M. et al., 2020, J. Mach. Learn. Res., 21, 1
Bernardeau F., Mellier Y., van Waerbeke L., 2002, A&A, 389, L28
Bradač M., Lombardi M., Schneider P., 2004, A&A, 424, 13
Bruna J., Mallat S., 2013, IEEE Trans. Pattern. Anal., 35, 1872
Bruna J., Mallat S., 2019, Math. Stat. Learn., 1, 257
Bruna J., Mallat S., Bacry E. J.-F. M., 2015, Ann. Stat., 43, 323
Carron J., 2011, ApJ, 738, 86
Carron J., Szapudi I., 2013, MNRAS, 434, 2961
Eickenberg M., Exarchakis G., Hirn M., Mallat S., Thiry L., 2018, J. Chem. Phys., 148, 241732
Fisher R. A., 1935, J. R. Stat. Soc., 98, 39
Fu L. et al., 2014, MNRAS, 441, 2725
Gama F., Ribeiro A., Bruna J., 2018, preprint (arXiv:1806.08829)
Giblin B. et al., 2018, MNRAS, 480, 5529
Gupta A., Matilla J. M. Z., Hsu D., Haiman Z., 2018, Phys. Rev. D, 97, 103515
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hikage C. et al., 2003, PASJ, 55, 911
Hirn M., Mallat S., Poilvert N., 2017, Multiscale Model. Simul., 15, 827
Jain B., Van Waerbeke L., 2000, ApJ, 530, L1
Jaynes E. T., 1957, Phys. Rev., 106, 620
Kilbinger M., 2015, Rep. Prog. Phys., 78, 086901
Kilbinger M. et al., 2013, MNRAS, 430, 2200
Kratochvil J. M., Haiman Z., May M., 2010, Phys. Rev. D, 81, 043519
Kratochvil J. M., Lim E. A., Wang S., Haiman Z., May M., Huffenberger K., 2012, Phys. Rev. D, 85, 103513
Lecun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278
Li P., Dodelson S., Croft R. A. C., 2020, Phys. Rev. D, 101, 083510
Lin C.-A., Kilbinger M., 2015, A&A, 583, A70
Liu J., Petri A., Haiman Z., Hui L., Kratochvil J. M., May M., 2015a, Phys. Rev. D, 91, 063507
Liu X. et al., 2015b, MNRAS, 450, 2888
Mallat S., 2010, Recursive interferometric representations, Proc. 18th European Signal Processing Conference. IEEE, Denmark, p. 716
Mallat S., 2012, Commun. Pure Appl. Math., 65, 1331
Mandelbaum R., 2018, ARA&A, 56, 393
Marian L., Smith R. E., Bernstein G. M., 2009, ApJ, 698, L33
Mecke K. R., Buchert T., Wagner H., 1994, A&A, 288, 697
Neyrinck M. C., Szapudi I., Szalay A. S., 2011, ApJ, 731, 116
Petri A., 2016, Astron. Comput., 17, 73
Pisani A. et al., 2019, BAAS, 51, 40
Planck Collaboration XIII, 2016, A&A, 594, A13
Ribli D., Pataki B. Á., Csabai I., 2019a, Nat. Astron., 3, 93
Ribli D., Pataki B. Á., Zorrilla Matilla J. M., Hsu D., Haiman Z., Csabai I., 2019b, MNRAS, 490, 1843
Sefusatti E., Crocce M., Pueblas S., Scoccimarro R., 2006, Phys. Rev. D, 74, 023522
Semboloni E., Schrabback T., van Waerbeke L., Vafaei S., Hartlap J., Hilbert S., 2011, MNRAS, 410, 143
Shirasaki M., Yoshida N., 2014, ApJ, 786, 43
Sifre L., Mallat S., 2013, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 1233
Simpson F., James J. B., Heavens A. F., Heymans C., 2011, Phys. Rev. Lett., 107, 271301
Sinz P., Swift M. W., Brumwell X., Liu J., Kim K. J., Qi Y., Hirn M., 2020, J. Chem. Phys., 153, 084109
Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R., 2013, preprint (arXiv:1312.6199)
Takada M., Jain B., 2003, ApJ, 583, L49
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
van Waerbeke L., 2000, MNRAS, 313, 524
Welling M., 2005, in Cowell R. G., Ghahramani Z., eds, Robust Higher Order Statistics, Society for Artificial Intelligence and Statistics, New Jersey (USA), p. 405
Zorrilla Matilla J. M., Haiman Z., Hsu D., Gupta A., Petri A., 2016, Phys. Rev. D, 94, 083506

APPENDIX A: MORLET WAVELETS

Wavelets are localized oscillations in real space and band-pass filters in Fourier space. If we simply use a Gaussian envelope to modulate a plane wave, then we obtain a Gabor function

$$G(\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}/2} \mathrm{e}^{i \boldsymbol{k}_0 \cdot \boldsymbol{x}} , \tag{A1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix describing the size and shape of the Gaussian envelope, and $\boldsymbol{k}_0$ determines the frequency of the modulated oscillation. To keep maximum symmetry, usually $\boldsymbol{\Sigma}$ is selected to have only 1 eigen-value different from the others, and $\boldsymbol{k}_0$ to be along that eigen direction. Thus, we denote the eigen value along $\boldsymbol{k}_0$ by σ^2 and the other eigen value by σ^2/s^2. The parameter s is also the ratio of transverse to radial width of the wavelet in Fourier space.

The Fourier transform of a Gabor function is simply a Gaussian filter centred at $\boldsymbol{k}_0$,

$$\tilde{G}(\boldsymbol{k}) = \mathrm{e}^{-(\boldsymbol{k}-\boldsymbol{k}_0)^T \boldsymbol{\Sigma} (\boldsymbol{k}-\boldsymbol{k}_0)/2} . \tag{A2}$$

Wider envelope in real space makes narrower filter in Fourier space. Note that the product $k_0\sigma$ determines the number of oscillations within $\pm\pi \approx 3$ standard deviation of the Gaussian envelope and allows for a trade off between spatial and frequency resolution.

Unfortunately, a Gaussian profile in Fourier space does not go to zero at 0 frequency. This contradicts the admissibility of wavelet which requires wavelets to strictly be band-pass filters, not low-pass filters. Therefore, a small correction is required. A simple solution is to introduce an offset, β, before the Gaussian modulation. In Fourier space this is equivalent to subtracting another Gaussian profile centred at 0 to cancel out the 0-frequency contribution. Families of wavelets created in this way are called Morlet wavelets. Formally

$$\psi(\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}/2} \left(\mathrm{e}^{i \boldsymbol{k}_0 \cdot \boldsymbol{x}} - \beta \right) , \tag{A3}$$

where $\beta = \mathrm{e}^{-\boldsymbol{k}_0^T \boldsymbol{\Sigma} \boldsymbol{k}_0/2}$ is determined by the admissibility criterion. Its Fourier transform is

$$\tilde{\psi}(\boldsymbol{k}) = \tilde{G}(\boldsymbol{k}) - \beta \mathrm{e}^{-\boldsymbol{k}^T \boldsymbol{\Sigma} \boldsymbol{k}/2} . \tag{A4}$$

In our study, which is a 2D case, we follow the settings used in the 'kymatio' package mentioned in Section 3.3

$$\begin{aligned} \sigma &= 0.8 \times 2^j \\ k_0 &= \frac{3\pi}{4 \times 2^j} \\ s &= 4/L , \end{aligned} \tag{A5}$$

where σ is in unit of pixels, j is an integer starting from 0, and k_0 is always between 0 and 1. This choice allow a family of Morlet wavelets best covers the whole Fourier space with a dyadic sequence of scales (2^j). Examples of the Morlet wavelets we use are shown in Fig. 2. Within the wavelet envelope, there are about 2 cycles of oscillations, because $k_0\sigma \approx 2$.

APPENDIX B: SCATTERING TRANSFORM IN FOURIER SPACE

It is enlightening to collect some intuition of the scattering transform in the Fourier domain. In general, as a non-linear operator, a modulus in real space will mix Fourier modes and scatter information among different frequencies. In particular, taking the modulus of $I\star\psi$, where ψ has a single peak in Fourier space, will re-express I's information around ψ's frequency in forms of lower frequencies. In other words, the typical frequency of $|I\star\psi|$ is lower than $I\star\psi$.

Intuitively, this is because the modulus is converting complex-valued oscillations into its local strength, namely its envelope. Formally, this can be revealed by first writing $|I\star\psi|$ as $\sqrt{(I\star\psi)(I\star\psi)^*}$, where * stands for complex conjugate, and then Taylor expanding the square root in terms of $(I\star\psi)(I\star\psi)^* - C$, where C is the mean of $(I\star\psi)(I\star\psi)^*$ over all pixels (Mallat 2010). The leading term of the Taylor expansion is proportional to $(I\star\psi)(I\star\psi)^* - C$ itself, which corresponds to $I\star\psi$'s autocorrelation in Fourier space. When the power spectrum of I is a smooth function, the frequency distribution of $I\star\psi$ is similar to ψ. For the Morlet wavelets used in the scattering transform, the central wavenumber of the wavelet ψ is roughly k_0 (as defined in Appendix A), and its half-width in Fourier space around $1/\sigma$. So, its autocorrelation will have an half-width around $\sqrt{2}/\sigma$ and a centroid at 0. As $\sqrt{2}/\sigma \approx 0.75k_0 < k_0$ (equation A5), this means that the typical frequency of $|I\star\psi|$ is lower than $I\star\psi$. Therefore, the core operation $I \rightarrow |I\star\psi|$ re-expresses high-frequency information of I_n in terms of lower frequency modes including the 0-frequency component in the next-order fields I_{n+1}. As the 0-frequency component is translation invariant, it can be directly used as a statistical descriptor of the original field.

Writing the modulus $|x|$ as $\sqrt{|x|^2} = \sqrt{x\cdot x^*}$ brings an interesting question: what happens if we replace each modulus by modulus squared? It can be shown that, in this case, the nth-order scattering coefficients will exactly become some averaged 2^n-point-spectra weighted (binned) by wavelets. Nevertheless, they are not equivalent to any degenerate case of 2^n-point functions in either real or Fourier domain. For example, at the second order, these 'pseudo' scattering coefficients become $\iiint \tilde{I}_0(\boldsymbol{k}_1)\tilde{I}_0(-\boldsymbol{k}_1' - \boldsymbol{k}_2)\tilde{I}_0(\boldsymbol{k}_1')\tilde{I}_0(-\boldsymbol{k}_1' + \boldsymbol{k}_2)\cdot W\cdot \mathrm{d}\boldsymbol{k}_1\mathrm{d}\boldsymbol{k}_1'\mathrm{d}\boldsymbol{k}_2$, where the weight is determined by the wavelets: $W = \tilde{\psi}_1(\boldsymbol{k}_1)\tilde{\psi}_1(\boldsymbol{k}_1+\boldsymbol{k}_2)\tilde{\psi}_1(-\boldsymbol{k}_1')\tilde{\psi}_1(-\boldsymbol{k}_1'+\boldsymbol{k}_2)\tilde{\psi}_2{}^2(\boldsymbol{k}_2)$, and the tilde sign denotes Fourier conjugate. Although these 'pseudo' coefficients may help us understand the connection between scattering transform and N-point functions in terms of how they organize spatial configurations, the genuine scattering transform is fundamentally different from N-point functions, because it generates 'first-order' estimators, which alleviates the problem of classic moments described in Carron (2011) when dealing with tailed probability distribution. Indeed, we find that the constraining power of genuine scattering coefficients is about 4 times stronger than these 'pseudo' ones (in the noiseless, unsmoothed case). We will discuss this further in another paper (Cheng et al. in preparation).

APPENDIX C: COSMOLOGICAL INFERENCE FRAMEWORK

In this appendix, we describe the Fisher forecast formalism used to infer the cosmological parameters in this study. According to the Cramér–Rao inequality, the variance of any unbiased estimator $\hat{\boldsymbol{\theta}}$ for model parameters $\boldsymbol{\theta}$ cannot be smaller than the inverse of the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ of the model

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}) \geq \mathbf{I}(\boldsymbol{\theta})^{-1}. \tag{C1}$$

Elements of the Fisher matrix is defined as

$$I_{m,n}(\boldsymbol{\theta}) \equiv \left\langle \frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_n}\right\rangle, \tag{C2}$$

where $\boldsymbol{x}$ is the observable, p is the likelihood function, and $\langle\cdot\rangle$ is the expectation over $\boldsymbol{x}$. In our cosmological case, $\boldsymbol{\theta}$ represents cosmological parameters, $\boldsymbol{\theta} = (\Omega_\mathrm{m}, \sigma_8)$, and $\boldsymbol{x}$ represents the statistical descriptors such as the scattering coefficients. The function $p(\boldsymbol{x}|\boldsymbol{\theta})$ is called the likelihood of $\boldsymbol{\theta}$ when $\boldsymbol{x}$ is fixed, and is called the probability density function (PDF) of $\boldsymbol{x}$ when $\boldsymbol{\theta}$ is fixed.

In our study, we assume that given any cosmology $\boldsymbol{\theta}$, the PDF of statistical descriptors $\boldsymbol{x}$ is Gaussian

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \frac{1}{\sqrt{|\mathbf{C}|}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})], \tag{C3}$$

where $\mathbf{C}(\boldsymbol{\theta})$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$ are the mean and covariance matrix depending on the cosmological parameters $\boldsymbol{\theta}$. Thus, elements of the Fisher matrix can be written as

$$I_{m,n} = \frac{\partial\boldsymbol{\mu}^T}{\partial\theta_m}\mathbf{C}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_n} + \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\theta_m}\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\theta_n}), \tag{C4}$$

where the first and second items describe the information from cosmological dependence of $\boldsymbol{\mu}$ and $\mathbf{C}$, respectively. To obtain these items for arbitrary cosmology, we first calculate the sample mean and covariance matrix of the 512 realizations of each cosmology in the simulations (Section 3.1). The sample mean is an unbiased estimator of the real mean vector, but to unbiasly estimate the inverse of covariance matrix, $\mathbf{C}^{-1}$, a correction factor is needed (Hartlap, Simon & Schneider 2007):

$$\widehat{\mathbf{C}^{-1}} = \frac{N-D-2}{N-1}\widehat{\mathbf{C}}^{-1}, \tag{C5}$$

where $\widehat{\mathbf{C}^{-1}}$ is the unbiased estimator in the inverse, N is the number of independent sample used for the estimation, D is the dimension of each data vector, and $\widehat{\mathbf{C}}$ is the sample covariance before Bessel's correction. Then, with a further assumption that $\boldsymbol{\mu}$ and $\mathbf{C}$ have smooth cosmological dependence, we use third-order polynomials to fit for the cosmological dependence of $\boldsymbol{\mu}$'s elements and use 2nd-order polynomials for $\mathbf{C}$'s elements.

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