FEATURE

Kavli IPMU Professor Masahiro Takada
Research Area : Cosmology

Cosmology and Statistics

1. Introduction

Observational cosmology is an exciting research field. High-precision cosmology data sets, as represented by the cosmic microwave background (CMB) experiments, have allowed us to address the most fundamental questions of the Universe, such as dark matter, dark energy, and the age of the Universe. 21st century cosmology requires both theory and observation and has now grown to be an "experimental" science.

Since some articles in the previous IPMU News touched on these advances in cosmology, we would like to discuss a different topic in this article, namely, statistical aspects of cosmology. Cosmological theory and methodology involve various statistical concepts. Cosmological data sets may look complicated at first glance. For instance, the fundamental observational quantity for CMB is the black-body radiation temperature of CMB photons observed in each direction in the celestial sky. In the case of a galaxy survey, the spatial distribution of galaxies is the fundamental observational quantity. How can these cosmological data sets be quantified and compared with theory? What are the premises and assumptions made in cosmological analysis and what are the limitations? The goal of this article is to discuss these questions.

2. Cosmological Principle, Ergodic Hypothesis and Two-Point Correlation

Suppose that $F(\boldsymbol{\theta})$ is a fluctuation field obtained from a cosmological data set. In the case of CMB, it is the CMB temperature fluctuation field, defined as $F(\boldsymbol{\theta}) \equiv [T(\boldsymbol{\theta}) - \bar{T}]/\bar{T}$, where $T(\boldsymbol{\theta})$ is the CMB temperature in the direction $\boldsymbol{\theta}$ and $\bar{T}$ is the average temperature over the sky (see the upper-left plot of Fig. 1). Although a similar field can be defined in three-dimensional space, let's consider the two-dimensional field in the following. The field $F(\boldsymbol{\theta})$ is observed with a finite angular resolution in actual observation, and would therefore be given in discretized pixels. The Planck satellite observed the CMB sky with angular resolution of about 5 arcmin, providing the CMB temperature fluctuation field in about 5 millions pixels. This is an enormous data set.

However, cosmology theory cannot reproduce the $F(\boldsymbol{\theta})$ field as is observed. More precisely, to build a model to reproduce the observed $F(\boldsymbol{\theta})$, one would need to introduce too many model parameters that can be as many as the degrees of freedom of the data. This is obviously not interesting, and we do not want to make such a fruitless effort. For this, a cosmological analysis usually employs the cosmological principle:

- the Universe is homogeneous and isotropic in a statistical sense.

In simpler terms, it says that there is no *spatial* direction or position in the Universe, or we on the Earth are not located at a spatial position in the Universe. Thus, this principle would be regarded as a democratic concept.

If we accept the cosmological principles, we can

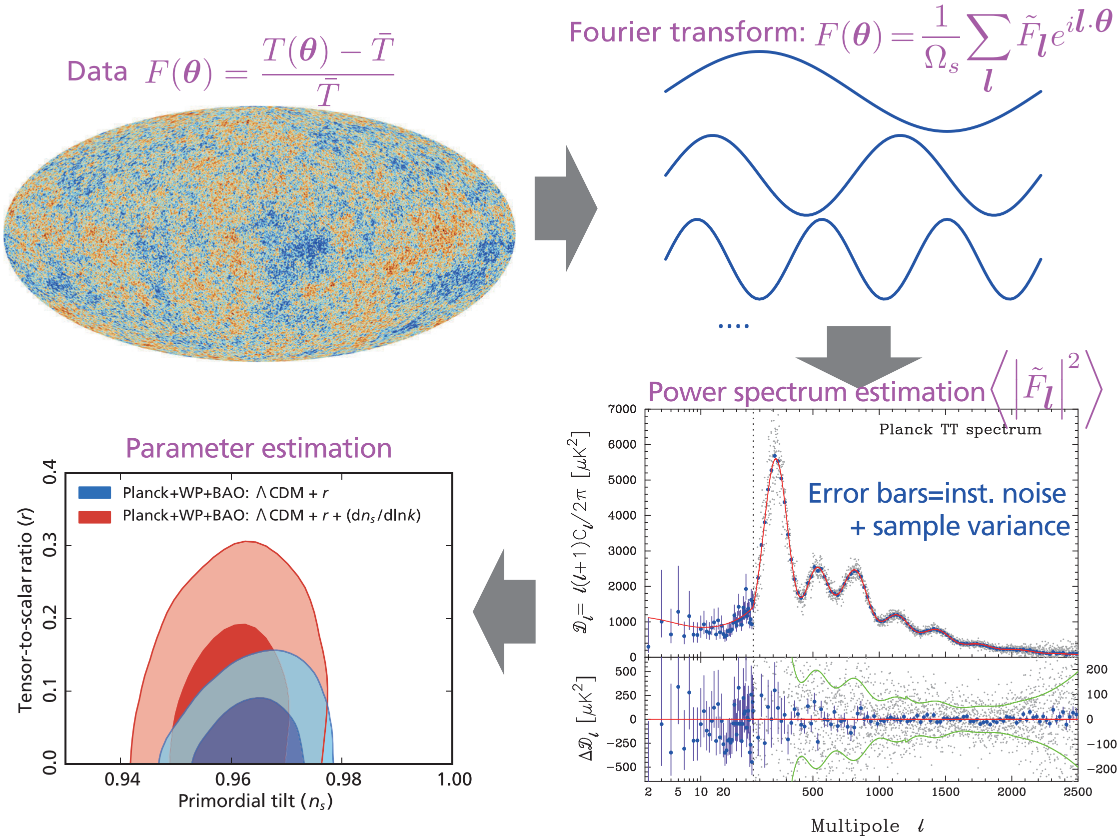

Figure 1. A flowchart of cosmological analysis. *Upper-left plot*: A cosmological fluctuation field $F(\boldsymbol{\theta})$. In this example, the temperature fluctuation field of the cosmic microwave background (CMB), taken from the Planck data. *Upper-right plot*: A Fourier decomposition of the field. *Lower-right plot*: A power spectrum estimation from the Fourier coefficients of the field, $\tilde{F}_{\boldsymbol{l}}$. The gray points (in the upper panel) show the measured value at each wavenumber bin (more exactly multipole bin in the spherical harmonics expansion). The blue points are the average among the several wavenumber bins. The error bars account for statistical uncertainties in each bin arising from both the sample variance due to a finite number of the Fourier modes and observational effects such as the instrumental noise. The red curve is the best-fit theoretical model. *Lower- left plot*: An example of parameter estimation, obtained by comparing the measurement with the model prediction. The confidence region of each parameter is obtained by properly propagating the measurement errors into parameters.

consider that the observed field $F(\boldsymbol{\theta})$ is a representative sample of the parent population of $\{F(\boldsymbol{\theta})\}$ that exists in a vast universe. Again in the CMB case, we assume that even if an independent observer who is located somewhere far away from us in the Universe observed the CMB, he would see a similar field of the CMB fluctuations as what we observed. We need to statistically quantify the extent of similarity, however; i.e., how typical or representative the field we observed is compared to the expectation, as we will discuss below.

A Fourier decomposition is useful to quantify the observed field $F(\boldsymbol{\theta})$. That is, we decompose the field into orthogonal functions of different wavenumbers:

$$F(\boldsymbol{\theta}) = \frac{1}{\Omega_s}\sum_{\boldsymbol{l}} \tilde{F}_{\boldsymbol{l}} e^{i\boldsymbol{l}\cdot\boldsymbol{\theta}} \quad (1)$$

Here we ignored the curvature of the celestial sphere for simplicity and assumed that the sky is a two-dimensional flat space. If Ω_s is an area of the observed region, the fundamental wavenumber of the Fourier decomposition $l_f \equiv 2\pi/\Omega_s^{1/2}$ ($\Omega_s^{1/2}$ is a one-side length of the observed area). Then the Fourier modes are given as $\boldsymbol{l} = l_f(n_x, n_y)$ ($n_x, n_y = \pm1, \pm2, \ldots$). The Fourier coefficient $\tilde{F}_{\boldsymbol{l}}$ is a quantity to describe how much the Fourier mode of wavenumber $\boldsymbol{l}$ contributes to the observed field.

The Fourier coefficient is generally expressed as $\tilde{F}_{\boldsymbol{l}} = |\tilde{F}_{\boldsymbol{l}}|\, e^{i\varphi_{\boldsymbol{l}}}$, carrying two degrees of freedom: the amplitude and the phase. Due to the statistical isotropy of the cosmological principle, the phase unlikely carries useful information.[1] Hence, as

[1] If there is any correlation between the phases of different Fourier modes, it easily causes an anisotropic or direction-aligned pattern of the field.

a useful quantity to characterize the amplitude information of the Fourier mode, we can define the power spectrum estimator as

$$\hat{P}_F(l) \equiv \frac{1}{N_{\rm mode}(l)\Omega_s} \sum_{|\boldsymbol{l}'|\in l} \left|\tilde{F}_{\boldsymbol{l}'}\right|^2 \qquad (2)$$

where $\Sigma_{|\boldsymbol{l}'|\in l}$ denotes a summation over all the Fourier modes satisfying $|\boldsymbol{l}'| \in l$ to within the bin width. $N_{\rm mode}(l)$ is the number of the Fourier modes in the summation, and is approximated for the mode $l \gg l_f$ as

$$N_{\rm mode}(l) \equiv \sum_{|\boldsymbol{l}'|\in l} \simeq \frac{2\pi l \Delta l}{(2\pi)^2/\Omega_s} \qquad (3)$$

where Δl is the bin width around the mode l, which an observer has to choose. Here $2\pi l \Delta l$ is the area of Fourier space used for the power spectrum estimation, while $(2\pi)^2/\Omega_s = l_f^2$ is the area of the fundamental mode element. Thus, the power spectrum estimation allows for a significant data compression, from the vast information content of two-dimensional data to the one-dimensional scalar quantity, $\hat{P}_F(l)$. By ignoring the phase information of the Fourier coefficients and assuming that the Fourier modes with the same length $l = |\boldsymbol{l}|$ are equivalent to each other due to the cosmological principle, we estimate the power spectrum from the average of the Fourier coefficients $|\tilde{F}_{\boldsymbol{l}}|^2$. This would be the simplest statistical quantity to characterize the amplitude of the Fourier coefficients.

What is the main difference between laboratory physics and cosmological experiments? In most cases a physics experiment can be repeatedly done at a laboratory. Then, the expectation value and the statistical error can be estimated from the average and variance of the independent experiment events (realizations). On the other hand, we cannot repeat a cosmological experiment: there is one observation region or there is only one universe even if the all-sky survey is done. Thus, in a cosmological analysis we approximate the ensemble average of independent realizations by the "sample average"; for the above example, the power spectrum is estimated from the average over the Fourier modes of the same length, $|\tilde{F}_{\boldsymbol{l}}|^2$, which are all drawn from the same realization (one survey region). This is sometimes called the "ergodic hypothesis" in cosmology. A cosmologist has to take into account statistical uncertainty and limitation arising from this hypothesis.

3. Primordial Gaussian Random Fluctuations and Sample Variance

In the preceding section we introduced the power spectrum that is a standard statistical quantity in cosmology. In fact, there is a case that the power spectrum carries *all* the statistical information of data. You may wonder, "is there such a convenient case for us?", but the answer is indeed yes! As we will see below, our Universe appears to be simple and beautiful in a "statistical" sense.

The inflationary scenario provides a plausible mechanism to explain the origin of cosmological fluctuation fields such as the CMB temperature fluctuations and large-scale structure. It is a scenario that the universe underwent an exponential expansion at the beginning of hot Big Bang universe. Since the universe itself is tiny during the inflation era, one needs to consider a quantization of a field that causes an inflationary expansion, which is often called *inflaton*. Due to the uncertainty principle of quantum mechanics, the inflaton inevitably has quantum fluctuations. Then the inflationary scenario predicts that the quantum fluctuations are stretched out by the exponential expansion to macroscopic scales, leaving classical fluctuations on horizon scales (the scales beyond causal contact). In quantum field theory, different wavenumber modes ($\boldsymbol{k}$) correspond to different quantum states. Furthermore, it is assumed in a standard inflation model that an interaction involved in the inflation field is small, and different quantum states are independent from each other. That is, if we denote the primordial classical fluctuation field by $\tilde{\zeta}_{\boldsymbol{k}}$ (often called the primordial curvature perturbation), the following condition is satisfied:

$$\langle \tilde{\zeta}_{\boldsymbol{k}} \tilde{\zeta}_{\boldsymbol{k}'} \rangle \equiv P_\zeta(k)(2\pi)^3 \delta_D^3(\boldsymbol{k}+\boldsymbol{k}') \qquad (4)$$

where $P_\zeta(k)$ is the primordial power spectrum. $\delta_D^3(\boldsymbol{k}+\boldsymbol{k}')$ is the three-dimensional Dirac delta function imposing that the fields of different wavenumbers are independent. In addition, we assumed that the primordial field is statistically isotropic, leading the power spectrum P_ζ to depend only on the length $k=|\boldsymbol{k}|$, which is valid as long as an inflation expansion is isotropic.

Thus, the inflationary scenario naturally predicts the generation of isotropic classical perturbations across the whole universe originating from random, free quantum fluctuations. More precisely, we expect that the phase of the primordial fluctuation field, $\tilde{\zeta}_{\boldsymbol{k}}$, is a random zero-mean variable, and only the amplitude contains the physical information, where the typical amplitude is given by the primordial power spectrum $P_\zeta(k)$. This means that the primordial field $\zeta(\boldsymbol{x})$ is a *Gaussian random field*. The Gaussian field has simple statistical properties. The even order correlation functions[2] are given by a product of the two-point correlation function, which is the inverse Fourier transform of the power spectrum. The odd order correlation functions[3] are all vanishing. That is, *the statistical properties of a Gaussian random field are fully described by its power spectrum*.

One problem of cosmic structure formation is solving how the perturbations of each component of radiation, baryon and dark matter have evolved in an expanding universe that has undergone from the radiation-, matter-, and dark-energy dominated era, given the initial conditions such as those set by the inflation models. As long as the amplitude of the perturbations is much smaller than unity, we can use *the linearized perturbation theory* to solve the dynamical evolution of multi-component system based on the linearized Einstein equations and the linearized Boltzmann equations. The nice thing is that different Fourier modes evolve *independently* in the linear regime. This means that the perturbations in the linear regime preserve statistical properties of the primordial perturbations. In fact the observed CMB temperature fluctuation field, which is well in the linear regime, is found to be in remarkable agreement with a Gaussian random field, which is considered as one of the pieces of strong evidence of the inflationary scenario.

Hence, a power spectrum estimation from the cosmological data can be regarded as an appropriate approach, motivated by the inflationary scenario. As we stated in the previous section, we need to take into account statistical uncertainty in the power spectrum estimation arising due to a finite number of the Fourier modes. This uncertainty is called the *sample variance*. The statistical uncertainty in the power spectrum estimation is given by the covariance matrix, and can be analytically computed for a Gaussian field:

$$\mathrm{Cov}[\hat{P}_F(l),\hat{P}_F(l')] \equiv \left\langle \hat{P}_F(l)\hat{P}_F(l')\right\rangle - \left\langle \hat{P}_F(l)\right\rangle\left\langle \hat{P}_F(l')\right\rangle = \frac{2}{N_{\mathrm{mode}}(l)}\,\delta^K_{ll'}\,P_F(l)^2 \qquad (5)$$

Here $\delta^K_{ll'}$ is a Kronecker-type delta function, defined in that $\delta^K_{ll'}=1$ if $l=l'$ to within the bin width, otherwise $\delta^K_{ll'}=0$. Thus the covariance matrix for a Gaussian field has only diagonal elements.[4] In other words, the power spectra of different bins are independent from each other. For a non-Gaussian field, the covariance matrix has additional contribution that arises from the 4-point correlation function that cannot be expressed by a combination of the power spectra, causing correlations between the power spectra of different bins. In actual observation, other statistical noise such as instrument noise also contributes to the covariance, but we ignore the contribution for simplicity in the following.

The covariance elements at each l bin give a 1σ-distribution of the measurement values around the expectation value, when the power spectrum at the l bin is estimated from each of the survey realizations for the fixed area $\boldsymbol{\Omega}_s$. Hence an expected statistical significance of the power spectrum

[2] For instance, the 4-point correlation function is given by $\langle\zeta(\boldsymbol{x}_1)\zeta(\boldsymbol{x}_2)\zeta(\boldsymbol{x}_3)\zeta(\boldsymbol{x}_4)\rangle$.
[3] For instance, the 3-point correlation function is given by $\langle\zeta(\boldsymbol{x}_1)\zeta(\boldsymbol{x}_2)\zeta(\boldsymbol{x}_3)\rangle$.
[4] The prefactor 2 on the r.h.s arises from the fact $\tilde{F}_{\boldsymbol{l}}=\tilde{F}^*_{-\boldsymbol{l}}$ due to real condition of the fluctuation field $F(\boldsymbol{\theta})$.

estimation at each l bin is

$$\left[\frac{P_F(l)}{\sigma(P_F(l))}\right]^2 = \frac{N_{\rm mode}(l)}{2} \qquad (6)$$

where $\sigma(P_F(l)) = [{\rm Cov}(\hat{P}_F(l), \hat{P}_F(l))]^{1/2}$. For a Gaussian case the statistical significance does not depend on the power spectrum, but rather depends only on the number of Fourier modes around the bin l, $N_{\rm mode}(l)$. Because of $N_{\rm mode}(l) \propto \Omega_s l \Delta l$, the statistical significance is greater for a larger area, a higher wavenumber bin l, and a wider bin width Δl.

The lower-right panel of Fig. 1 shows the CMB temperature power spectrum measured from the Planck data. The plot clearly shows that the measurement accuracy is increasingly higher at higher wavenumber bin l (more exactly the higher multipoles that are the order in spherical harmonics expansion). The error bars around blue data points denote the statistical uncertainties due to the sample variance as well as the instrumental noise. Again we would like to emphasize that an observer has to model the sample variance.

Once the statistical uncertainties in the power spectrum measurement are given, one can compare the measurement with theory:

$$\hat{P}_F(l) \longleftrightarrow P^{\rm model}[l; P_\zeta(k), \Omega_{\rm m0}h^2, \Omega_{\rm b0}h^2, \Omega_{\rm de}, ...] \qquad (7)$$

Here we assumed on the r.h.s that the model power spectrum, $P_F^{\rm model}(l)$, is given as a function of the primordial power spectrum and other cosmological parameters that specify the cosmic expansion history. This assumption is valid as long as the fluctuations are in the linear regime. First, we can study whether a model of interest can give a good fit to the measurement to within the error bars. Next, by propagating the measurement errors into parameter estimation, we can estimate a confidence region for each parameter. The lower-left panel shows such an example. Strikingly, the Planck team successfully attained the ultimate statistical precision allowed by the sample variance up to $l \sim 2000$, as it uses the all-sky data and the instrument noise is negligible up to the multipoles. Namely, since the power spectrum contains all the statistical information for a Gaussian field, they were able to use all the CMB temperature information in the Universe in order to extract the cosmological information.

What we have so far described is summarized by the following three remarks. (1) We measure the power spectrum from a cosmological data set assuming the cosmological principle that the Universe is statistically isotropic and homogeneous. (2) Assuming the ergodic hypothesis that an observed region is a representative sample randomly drawn from the parent populations, an observer needs to "model" sample variance effects in the power spectrum estimation. (3) If the cosmological fluctuation field is a Gaussian random field as predicted by the inflationary scenario, the power spectrum contains all the statistical information of the field. If any of these three breaks down, the power spectrum is no longer the optimal quantity to extract all the statistical information.

4. Large-Scale Structure Formation: Nonlinearity of Gravity

While we have so far mainly discussed the linear Gaussian field such as the CMB field, in this section we consider the cosmological data sets available from galaxy surveys that are aimed at exploring the accelerating universe. The dynamical evolution of fluctuations after the CMB epoch is driven mainly by the spatial inhomogeneities of dark matter distribution. The current standard scenario is that the seed perturbations have grown due to gravitational instability and then formed the present-day, cosmic hierarchical structures containing nonlinear structures such as galaxies, clusters of galaxies and large-scale structure seen through the distribution of galaxies. This scenario is the cold dark matter dominated structure formation model (hereafter simply CDM model). The CDM model predicts a bottom-up structure formation: small-scale nonlinear structures are first formed

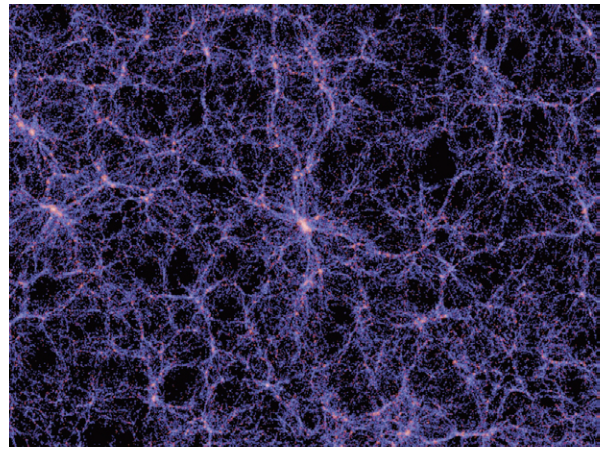

Figure 2. An example of N-body simulations for the cold dark matter (CDM) dominated the structure formation model. Even if starting from the initial Gaussian condition, the spatial distribution of dark matter, as represented by N-body particles, display rich non-Gaussian features as a consequence of complex nonlinear gravitational instability. Dark matter halos appear at the spatially small places where dark matter particles particularly cluster. Very massive halos of cluster scales, with masses $\gtrsim 10^{15} M_\odot$, tend to appear at the intersection of cosmic webs/filaments (the obvious example is the halo at the center of the plot). The amount of dark matter contained in halos with masses greater than galactic scales $10^{12} M_\odot$ is a few 10 percent of the total matter in the universe. On the other hand, about 70 percent of the spatial volume is occupied by the spatially big, under-density regions – the so-called voids. Reflecting such asymmetry, the dark matter distribution becomes to have non-vanishing values for all the *n*-point correlation functions.

and then merge in a continuous hierarchy to form larger-scale structures.

We assume that CDM particles have negligible thermal velocity (cold) and interact with other particles only via gravity. We can show that, on scales well within horizon scale and from a spatially coarse-grained viewpoint, the time evolution of dark matter clustering follows the pressureless and irrotational fluid equations in an expanding universe:

$$\begin{aligned} &\frac{\partial \delta_m}{\partial t} + \frac{1}{a}\nabla\,[(1+\delta_m)\,v_m] = 0 \\ &\frac{\partial v_m}{\partial t} + \frac{\dot{a}}{a} v_m + \frac{1}{a}\,(v_m \cdot \nabla)\, v_m = -\frac{1}{a}\nabla\phi \\ &\nabla^2 \phi = 4\pi G \bar{\rho}_m\, a^2 \delta_m \end{aligned} \tag{8}$$

Here $a(t)$ is the scale factor that is an increasing function with cosmic expansion, $\delta_m(\boldsymbol{x}) \equiv [\rho_m(\boldsymbol{x}) - \bar{\rho}_m]/\rho_m$ is the mass density fluctuation field, $v_m(\boldsymbol{x})$ is the peculiar velocity field, and $\phi(\boldsymbol{x})$ is the gravitational potential field. In a completely isotropic and homogeneous universe, $\delta_m = |v_m| = 0$ everywhere. We can study the dynamical evolution of the dark matter fields by solving these equations, starting from the initial conditions constrained by the CMB observations. As is obvious from the above equations, the dark matter fields linearly evolve as long as the amplitudes of the perturbations are much smaller than unity, i.e. $|\delta_m| = |v_m| \ll 1$ in units of $c = 1$ for speed of light. Once the nonlinear terms $\delta_m v_m$ and $(v_m \cdot \nabla) v_m$ become non-negligible compared to the linear terms as time goes by, however, the dark matter fields enter into the nonlinear regime. The nonlinear evolution induces a mode coupling between different Fourier modes, leading to a complex evolution of the dark matter fields. Thus, even if the initial fields are Gaussian, the nonlinearity of gravity induces non-Gaussian features in the spatial distribution of dark matter. The degrees of non-Gaussianity become greater at smaller scales and at lower redshift.

Thus, the power spectrum no longer describes the full information of dark matter distribution in the late-time universe. In fact, N-body simulation based studies for the CDM model, as illustrated in Fig. 2, have shown that the all *n*-point correlation functions of the dark matter distribution generally become non-vanishing.

Having described the non-Gaussian features of large-scale structure, one question arises. As described in the previous sections, the power spectrum describes the *full* information contained in the linear, Gaussian fluctuation field in the early universe, which gives the initial conditions of structure formation. On the other hand, the dark matter fluctuation field displays non-Gaussian features, and has non-vanishing values for all higher-order correlation functions beyond the power spectrum. Since we cannot extract any extra information beyond the initial conditions in analogy to the second law of thermodynamics, we can consider that some of the initial Gaussian

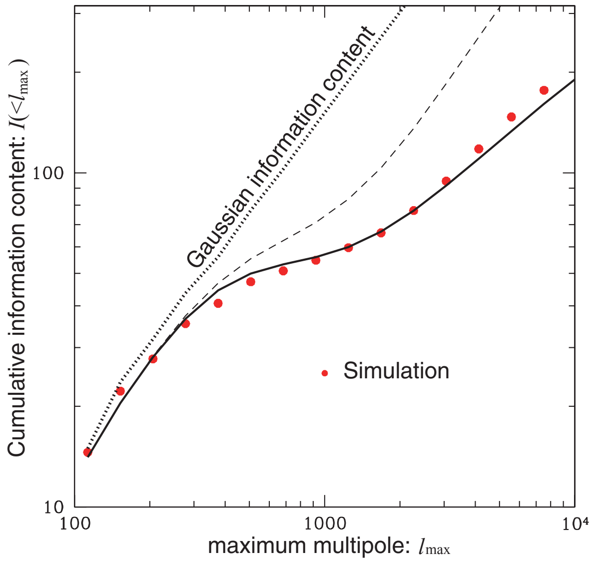

Figure 3. The cumulative information content contained in the weak lensing power spectrum. The weak lensing field is the projected field of the mass density fluctuation field along the line-of-sight, between an observer and source galaxies at redshift z_s = 1. Here we assume Ω_s = 1400 sq. degrees for the survey area. The "cumulative" information content is obtained by summing the signal-to-noise ratio of the power spectrum measurement (e.g., see Eq. 6) from the minimum wavenumber $l_{\rm min}$ = 72 up to a certain maximum wavenumber $l_{\rm max}$ as denoted by the x-axis. The dotted curve shows the information content if the weak lensing field is Gaussian as expected from the initial linear fluctuation field – the maximum information content. In this case, the information content follows a simple scaling given by $I(< l_{\rm max}) \propto l_{\rm max}\Omega_s^{1/2}$. The circle points show the result obtained by using mock catalogs of the weak lensing field that are constructed by using N-body simulations of nonlinear large-scale structure for the CDM model as in Fig. 2. In this case we fully take into account correlations between the power spectra of different bins, due to the non-Gaussianity of the weak lensing field. The solid curve shows the analytical model, showing remarkably nice agreement with the simulations. The dashed curve is the analytical result, if we ignore the effect of super-survey modes on the covariance (see text for details).

information leak into the higher-order correlation functions as a consequence of the nonlinear structure formation. Namely, the question is

- Can we recover the initial Gaussian information content from the present-day dark matter distribution by combining the statistical quantities?

This is indeed an unresolved problem in the field of cosmology. If the dynamical system is time-reversible, we can back to the initial conditions, and therefore should be able to recover the initial information content from observables of the current data (the final state). In the bottom-up CDM model the density fluctuations at sufficiently large scales are still in the linear regime, preserve the Gaussianity of the initial field, and therefore we can recover the initial information. On the other hand, the density fluctuations at very small scales are in the strongly nonlinear regime ($|\delta_m| \gg 1$) and the scales may have already lost the initial memory. For example, consider dark matter particles that are bound in a dark matter halo. Such particles may have a complex orbit scattering or oscillating around the halo center during their evolution history, so the snapshot information alone of the final state would not allow us to perfectly back their distribution to their initial positions in a time-reversible way. At the intermediate scales, which are in the weakly nonlinear regime, we may be able to recover "most" of the initial information. It is still not clear, however, to what extent we can recover the Gaussian information from the available observables from the scales. In fact, upcoming galaxies surveys aim at achieving a high-precision measurement of galaxy clustering or weak lensing observables in the weakly nonlinear regime, in order to use the measurements to do cosmology. Hence this question is very important for cosmologists.

In recent years we have been addressing the above question based on both N-body simulations and analytical model of structure formation. Fig. 3 shows one of our results. Here, we studied how much information the weak lensing power spectrum in the late-time universe carries, by using mock catalogs of weak lensing observables that were constructed using N-body simulation outputs of CDM structure formation. At small wavenumber $l \sim 100$, the weak lensing power spectrum recovers almost all the Gaussian information. However, at the larger wavenumber $l \gtrsim$ a few 100, the information content of the power spectrum is significantly reduced compared to the Gaussian information, more than a factor of 2 at $l \sim 1000$, from around which multipoles we expect to extract most useful cosmological simulation for upcoming surveys such as HSC. Thus the plot clearly shows that the power spectrum alone cannot extract all the information contained in the weak lensing field, and more than half of the Gaussian information has gone somewhere! We have indeed found in other work that the 3-point correlation function does add the information. However, the combined information of the 2- and 3-point correlation functions is not still as sufficient as the initial Gaussian information. This implies that the 4-point correlation function is important, which is even more complicated to measure.

Through these studies, we are also finding something very interesting we did not expect. We found that most of the information lost is caused by Fourier modes that are comparable with or beyond a survey region – *super survey modes*. Obviously super survey modes are not observable. This super-survey mode is regarded as a constant offset in the mean density inside the survey region compared to the cosmic mean (the ensemble averaged mean). However, we cannot know whether a modulation in the mean density is positive or negative. Suppose that a survey region is embedded into a slightly coherent, over-density region. In this case, the survey region is considered as a slightly positive curvature universe. Due to the nonlinear nature of gravity, the positive super-survey mode becomes coupled with all the Fourier modes inside the survey volume once the nonlinear structure formation evolves. That is, the time evolution of all sub-volume Fourier modes is enhanced compared to what we naively expect. We found that the mode-coupling between the super-survey mode and the Fourier modes inside the survey region causes a significant contribution to the sample variance in the power spectrum measurement. We then succeeded in formulating this effect in a unified form that can be applied to any large-scale structure probes. The solid curve in Fig. 3 shows the analytical model prediction for the information content including the effect of the super-sample mode, showing a remarkably nice agreement with the simulation results. Our finding opens up a new and interesting possibility: if we can properly take into account the effect of super-survey mode when comparing the measured power spectrum with theory, we may be able to infer the existence of the super survey mode, which lies in very, very large length scales, for upcoming wide-area surveys. This is potentially very interesting, and we are planning to further explore this effect.

5. Future Prospects

Upcoming galaxy surveys such as the SuMIRe galaxy survey are increasingly expensive, both in terms of time and cost. We want to attain the full potential of these surveys in order to tackle the most fundamental, yet profound problems such as the nature of dark energy. In this article we have described the hypotheses, assumptions and procedures that are often used when measuring statistical quantities from a cosmological data set and then using those to estimate cosmological parameters. The CMB experiment attained the maximum success in terms of the extracted information content, largely because the CMB field is Gaussian, the simplest field. For galaxy surveys, on the other hand, it is not yet clear which observables are most optimal to extract the maximum information. Given these facts, an interdisciplinary field between cosmology and statistics will be more and more strengthened. If you have any new idea, you should work on this new subject!

References
M. Takada and S. Bridle, New J. Phys. 9, 446 (2007)
M. Takada and B. Jain, MNRAS 395, 2065 (2009)
M. Sato et al., Astrophys. J. 701, 945 (2009)
I. Kayo, M. Takada, and B. Jain, MNRAS 429, 344 (2013)
M. Takada and W. Hu, Phys. Rev. D 87, 123504 (2013)
Y. Li, W. Hu, and M. Takada, Phys. Rev. D 89, 083519 (2014)
M. Takada and D. N. Spergel, MNRAS 441, 2456 (2014)