

# Statistics and Dynamics in the Large-scale Structure of the Universe

**Takahiko Matsubara**

Department of Physics & Astrophysics, Nagoya University, Chikusa, Nagoya 464-8602, JAPAN

E-mail: [taka@nagoya-u.jp](mailto:taka@nagoya-u.jp)

**Abstract.** In cosmology, observations and theories are related to each other by statistics in most cases. Especially, statistical methods play central roles in analyzing fluctuations in the universe, which are seeds of the present structure of the universe. The confrontation of the statistics and dynamics is one of the key methods to unveil the structure and evolution of the universe. I will review some of the major statistical methods in cosmology, in connection with linear and nonlinear dynamics of the large-scale structure of the universe. The present status of analyses of the observational data such as the Sloan Digital Sky Survey, and the future prospects to constrain the nature of exotic components of the universe such as the dark energy will be presented.

## 1. Introduction: History of the Universe

We now have a rough idea how this whole universe evolves with time and even how it began. The very beginning of the universe, however, is still quite uncertain. Some people say the universe is created from “nothing” by some sort of quantum gravity effect.

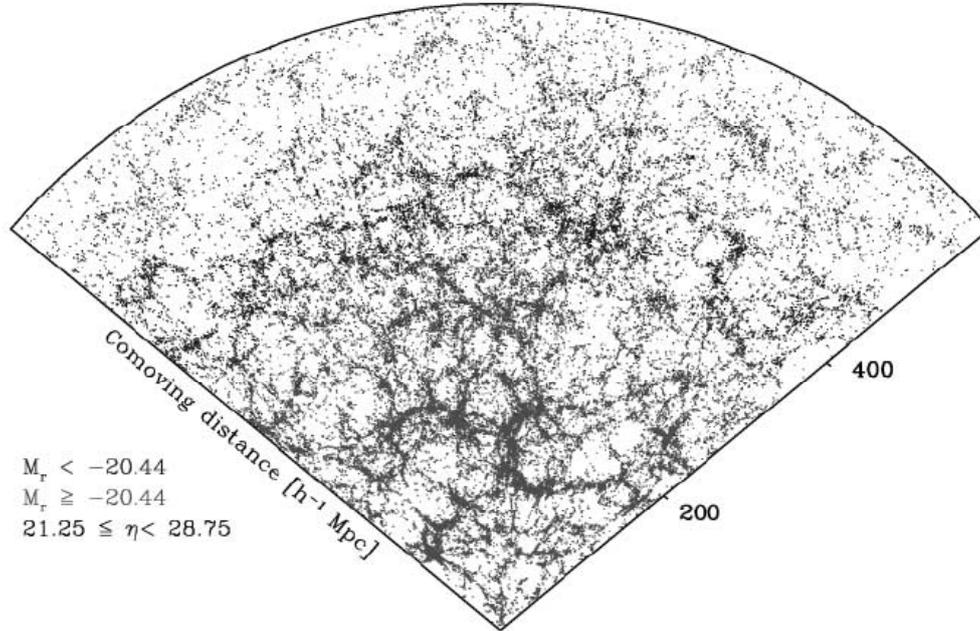
Soon after the beginning of the universe, the inflationary phase of the universe is believed to take place. The inflation stretched the whole space, and made the universe very smooth and made the geometry very flat. After that, the universe is filled with the state of quark soup, which is the plasma of quarks and gluons and many other particles.

After one second from the beginning of the universe, the nuclei, like hydrogen and helium, are synthesized. After four hundred thousand years, the universe suddenly became transparent with respect to photons. We can observe the state of the universe by photons that were emitted at this epoch.

These relic photons are called “cosmic microwave background radiation”. They were observed by Penzias & Wilson in 1960’s, and recently WMAP team observed the exquisite details of the temperature fluctuations of the CMB (Bennet et al. 2003). These fluctuations are the seeds of the present structure of the universe. All the complexity of this universe has its origin in this fluctuations in some sense.

After around one billion years or so, the first galaxies were formed. When and how the first galaxies were formed are currently one of the most exiting questions in cosmology. We still don’t know much about the origin of various species of galaxies. The age of the present universe is about fourteen billion years old.

Figure 1 shows the large-scale structure of the universe, which is probed by spatial distribution of galaxies (Park et al. 2005). In this figure, our galaxies is sitting at the bottom corner, and



**Figure 1.** The large-scale structure probed by the Sloan Digital Sky Survey.

the galaxies are represented by small dots. This structure is sometimes called as "the cosmic web". The distribution of galaxies is quite complex and there are many features, such as voids, filaments and walls, and so on. It is not trivial how this complexity emerges. How one can analyze this complex structure is the main topic of this review.

I would like to summarize our present knowledge on the composition of our universe. This is one of the remarkable outcomes in recent cosmology. One of the surprising fact is that the ordinary matters like atoms are only 5% of the whole universe. One-thirds of the rest is the dark matter, which is an unknown form of matter. The existence of the dark matter is only known by gravitational interaction. Two-thirds are mysterious dark energy, which is unknown form of energy. The dark energy accelerates the expansion of the universe. Most of the ordinary matters are free hydrogen and helium, they are 4.2%. Stars are 0.3%, neutrinos are 0.1%, photons are 0.001%, and planets are only  $10^{-6}\%$ .

We don't know what exactly the dark matter and dark energy are. This is a big question in present cosmology. The true nature of these components stand still in the darkness.

## 2. Large-scale Structure of the Universe

Now we are going to focus on the large-scale structure of the universe. Initial density perturbations are the origin of the large-scale structure. The initial perturbations are probably generated by quantum fluctuations in inflationary phase of the universe. The initial fluctuations are very small and expected to be represented by a random Gaussian field.

Even though the initial density fluctuations are small, the gravitational force amplifies the density fluctuations. The gravitational evolution is nonlinear, and the density field becomes non-Gaussian even though the initial field is Gaussian.

Depending on each cosmological models, the history of forming the large-scale structure is quite different. Some models can't reproduce the filamentary structure which is actually

observed in our universe as seen in Figure 1. But we shouldn't rely on our visual impression. To distinguish the true cosmological models from others, we should quantitatively analyze the large-scale structure of the universe. Quantifying the large-scale structure of the universe is a key to the mystery of our universe.

### 3. The Two-point Correlation Function

Totsuji & Kihara (1969) introduced the correlation function of galaxies to quantify the spatial distribution of galaxies. This statistic quantifies the clustering pattern of galaxies as a function of scales. It is defined as follows. We consider the small cells  $dV_1$  and  $dV_2$  separated by  $r$ . We write the probability of having galaxies in both cells as

$$dP = \bar{n}^2 dV_1 dV_2 [1 + \xi(r)], \quad (1)$$

where  $\bar{n}$  is the mean number density of galaxies. If the galaxy distribution is completely random, and Poissonian, the probability is just given by only the first term in the square bracket. When the clustering is present, the probability deviates from this and the correlation function  $\xi(r)$  is defined by the remaining term. This function depends on the separation  $r$ . When the clustering is strong on scale  $r$ , the value of the correlation function is large. When the galaxies are anti-correlated and galaxies doesn't prefer to exist on particular scales, then the correlation function will be negative on that scales.

Intuitively, when the correlation function is large on a scale  $r$ , then the galaxies prefer to form clusters of size  $r$ . When the correlation function is small, galaxies are not clustered on scale  $r$ .

Totsuji & Kihara suggested that the galaxy correlation function is given by approximately a power-law function of  $\xi(r) = (r/4.7)^{-1.8}$  by analyzing the Shane-Wirtanen catalog, which is a 2-dimensionally projected catalog of galaxies.

Davis & Peebles (1983) calculated the correlation function by the CfA redshift survey, which is a 3-dimensional catalog of galaxies. They also found that the correlation function is given by approximately a power-law function  $\xi(r) = (r/5.4)^{-1.8}$  on scales of  $0.1h^{-1}\text{Mpc} \leq r \leq 10h^{-1}\text{Mpc}$ . The amplitude of the correlation function varies from sample to sample, but the power-law slope ( $-1.8$ ) does not change from sample to sample.

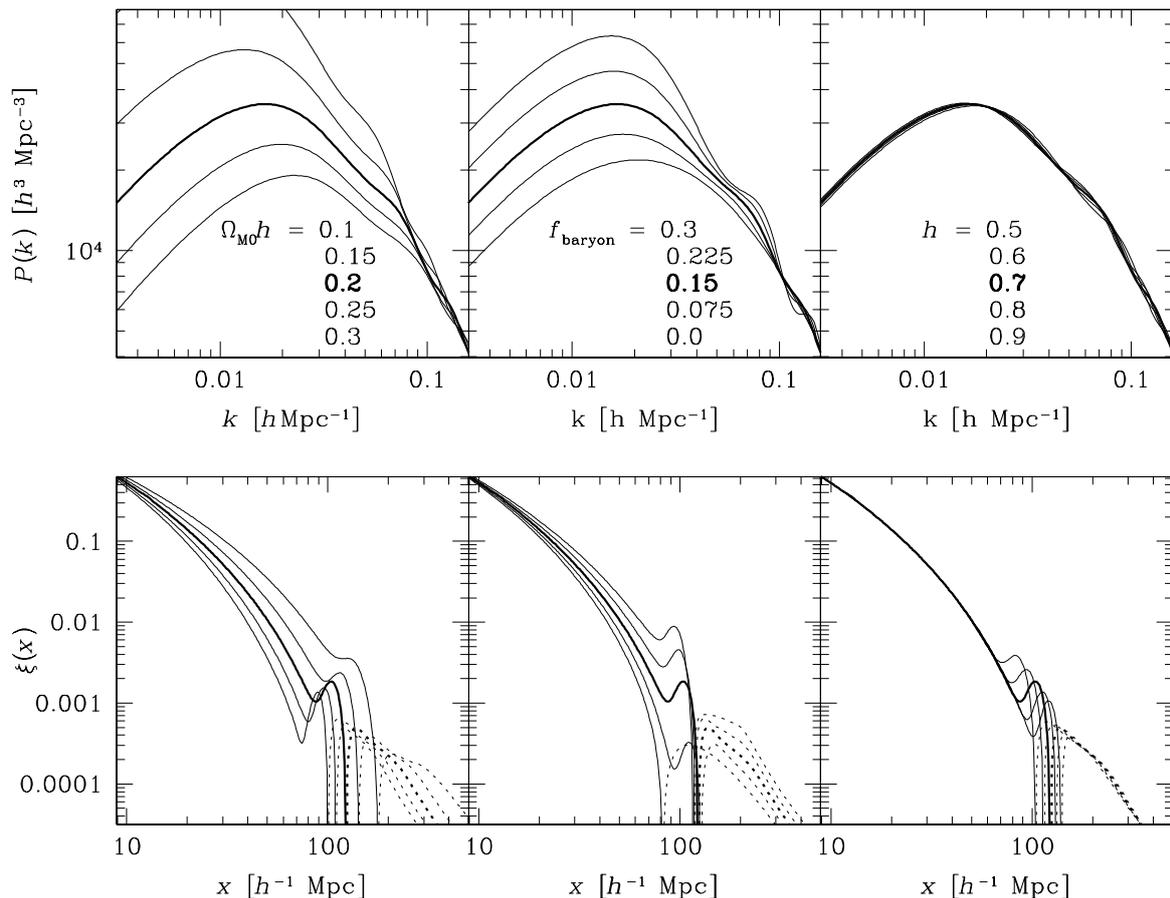
### 4. The Power Spectrum

Another statistic that is equally useful in cosmology is the power spectrum. The power spectrum  $P(k)$  is just a Fourier counterpart of the correlation function:

$$P(k) = \int d^3k e^{-i\mathbf{k}\cdot\mathbf{r}} \xi(r). \quad (2)$$

It represents the amplitude of the density fluctuations as a function of wave number,  $k$ . Theoretically, it is more straightforward to predict the evolution of the Fourier coefficients of density fields in a given cosmological models, than the evolution of real-space density fields. The power spectrum depends on the amount and nature of dark matter, baryons, and also depends on properties of the initial density fluctuations.

Figure 2 shows the effects of changing cosmological parameters on the power spectrum and the correlation function, assuming the linear perturbation theory (Matsubara 2004). The normalization is somehow arbitrary in cosmology, the fluctuations on  $8h^{-1}\text{Mpc}$  is fixed in this figure. In the left panels the effect of the dark matter is shown, in the center panels the effect of the baryons, and the right panels the effect of the hubble constant. Each parameters affect the correlation function quite differently, so that we can determine each parameters separately by accurately determine the correlation function or the power spectrum by galaxy surveys.



**Figure 2.** Parameter dependences of the power spectrum (top panels) and the correlation function (bottom panels). Variations from a fiducial model of  $\Omega_M = 0.2$ ,  $\Omega_b/\Omega_M = 0.15$  and  $h = 0.7$  are plotted.

The Sloan Digital Sky Survey, or SDSS, is the largest survey which is ever made in the human history. A dedicated 2.5m telescope is build only for this survey at Apache Point, New Mexico. This telescope is build for measuring the three-dimensional positions of galaxies. The survey started in the year 2000, and it will continue until 2008. By the end of this survey we will have 3-dimensional positions of order one million galaxies and one hundred thousand quasars. We already have large-fraction of data which can immediately be analyzed.

The Figure 1 shows the galaxy distribution that SDSS observed so far. The correlation function of SDSS galaxies is analyzed in, e.g., Abazajian et al. (2005). Because the number of galaxies are much larger than other surveys, the errors of the correlation function is very small. It is impressive. As we can see in the Figure 1 of Abazajian et al. (2005), the correlation function has approximately a power-law form, but in detail, it deviates from the power-law. There is a shoulder around the scale of 5 Mpc or so.

The existence of the shoulder in the correlation function involves the linear and nonlinear dynamics. Recently, this kind of behavior of the correlation function is modeled by the "Halo Occupation Distribution". In this model, the probability  $P(M|N)$  for a halo of mass  $M$  to harbor  $N$  galaxies is considered. This probability is modeled by assumptions like the radial distribution of galaxies in each halos. In this model, the correlation function is decomposed into two parts, i.e., 1-halo term and 2-halo term. The 1-halo term is the contribution from

the correlations among the same members of a halo. the 2-halo term is the contribution from the cross-halo pairs of galaxies. Once the model parameters are fixed, the resulting correlation function is uniquely predicted.

Comparison between the HOD model and the SDSS galaxies is given by Zehavi et al. (2004). In Figure 3 of Zehavi et al. (2004), one can see that the existence of the shoulder in the correlation function corresponds to the transition from 1-halo to 2-halo term.

## 5. Statistics for the Non-Gaussianity

The correlation function or the power spectrum I explained so far is so-called two-point statistics. These two-point statistics are known to be able to probe only Gaussian feature of distribution. That means, the two-point statistics cannot tell whether a given distribution is Gaussian or not. So one needs other statistics to search for non-Gaussianity. As I described above the non-Gaussianity is inevitable in the large-scale structure because of gravitational evolution, even when initial perturbation is Gaussian. Below I will explain how the Gaussianity can be statistically tested. There are many ways to test the Gaussianity, and I choose the representative statistics, like higher-order correlation function and Topological analyses.

### 5.1. Three-point correlation function

A the straightforward extension of the correlation analysis is to introduce the higher-order correlation function. They are just calculated by considering galaxy triplets or quadruplets instead of galaxy pairs as in calculating the two-point correlation function. We consider the probability of having galaxies in all cells  $dV_1$ ,  $dV_2$  and  $dV_3$ . The mutual distance among cells are  $r_{12}$ ,  $r_{23}$ , and  $r_{31}$ . This probability is written as

$$dP = \bar{n}^3 dV_1 dV_2 dV_3 [1 + \xi(r_{12}) + \xi(r_{23}) + \xi(r_{31}) + \zeta(r_{12}, r_{23}, r_{31})]. \quad (3)$$

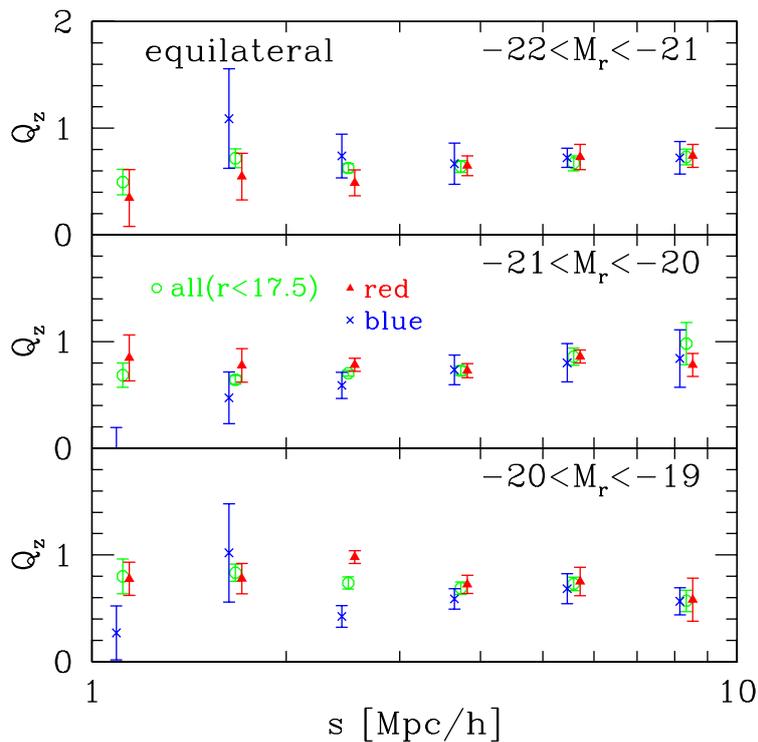
There are contributions from two-point correlation function here, and excluding these contributions, the three-point correlation function is defined by  $\zeta(r_{12}, r_{23}, r_{31})$ . In the large-scale structure, the order of the three-point correlation function on certain scales is approximately of order square of the value of two-point correlation functions of corresponding scales. Therefore it is common to define the normalized three-point correlation function Q defined by

$$Q(r_{12}, r_{23}, r_{31}) = \frac{\zeta(r_{12}, r_{23}, r_{31})}{\xi(r_{12})\xi(r_{23}) + r_{23}\xi(r_{31}) + r_{31}\xi(r_{12})} \quad (4)$$

Figure 3 is an example of the behavior of Q, evaluated from the SDSS galaxies (Kayo et al. 2004). The equilateral configuration of the triangle is considered in this Figure. Each panel and each line style corresponds to different species of galaxies. Quite irrespective to different kinds of galaxies, the value Q is nearly constant. One can distinguish the cosmological models using this statistics. It is important that the higher-order statistics is independent to the two-point correlation function.

### 5.2. Topological Analysis and Minkowski Functionals

Evaluating the N-point correlation functions from data becomes difficult when the order N is large. If all the hierarchy of the higher-order correlation functions are accurately determined, then the statistical information is complete. But it is practically impossible to evaluate all the N-point correlation function. Thus we need to explore other non-Gaussian statistics in order to deal with information from the higher-order statistics in practice. The topological analysis is one of such statistics. Minkowski functionals are some kind of extensions of the topological analysis. In this analysis the isodensity contours of galaxy distributions are considered.



**Figure 3.** Normalized three-point correlation function. Equilateral configuration is exemplified. Dependences on the absolute magnitude of the r-band ( $M_r$ ) and on colors are shown.

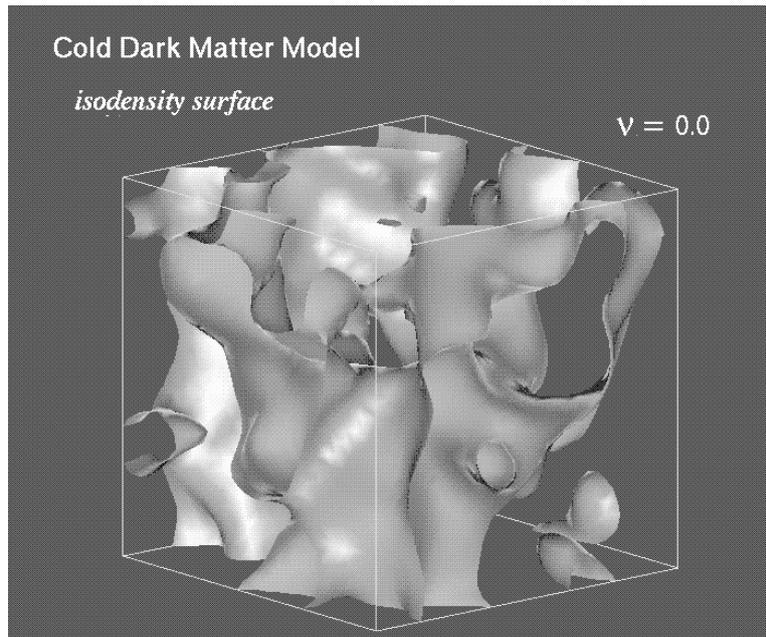
Figure 4 is the isodensity contours calculated from numerical simulation. Isodensity contours are the surface where the number density of galaxies has a certain value.

In the topological analysis the quantity called genus is usually analyzed. The genus is defined by a number of isolated contours minus the number of handles or holes. For example, the genus of isodensity surfaces in Figure 5 is 1, because the number of isolated contours is 3 and number of handles is 2. The Minkowski functionals are series of statistical quantities calculated from the isodensity surface. They are volume inside the isodensity contours, surface area, mean curvature, and Euler number. The Euler number is mathematically equivalent to the genus, and thus the genus statistic is included in Minkowski functionals.

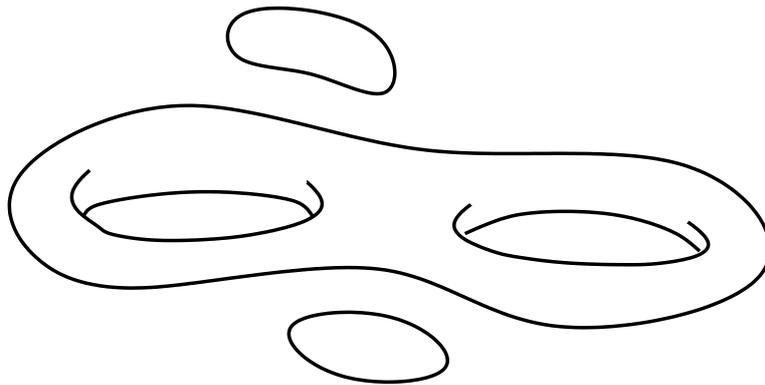
The important property of the topological analysis is that the genus as a function of the density threshold has a universal shape for Gaussian field. So if the shape of genus is deviates from Gaussian prediction, it means that the distribution is non-Gaussian. The relation between non-Gaussianity of density field and the deviation in the genus statistics was explored only by numerical method, but sometime ago I found the analytical method to track the non-Gaussianity in the topological analysis (Matsubara. Figure 6 shows the analytical prediction of the Minkowski functionals in weakly non-Gaussian fields, generated by the gravitational evolution.

## 6. Summary

- The large-scale structure of the universe is a useful probe of the history of the universe.
- The most popular statistics to analyze the large-scale structure is the two-point correlation function and the power spectrum.
- Gravitational nonlinear evolution results in the non-Gaussian distributions of galaxies even



**Figure 4.** Isodensity surface of a density field generated by a cosmological  $N$ -body simulation. The surface of the mean density is plotted in this figure.

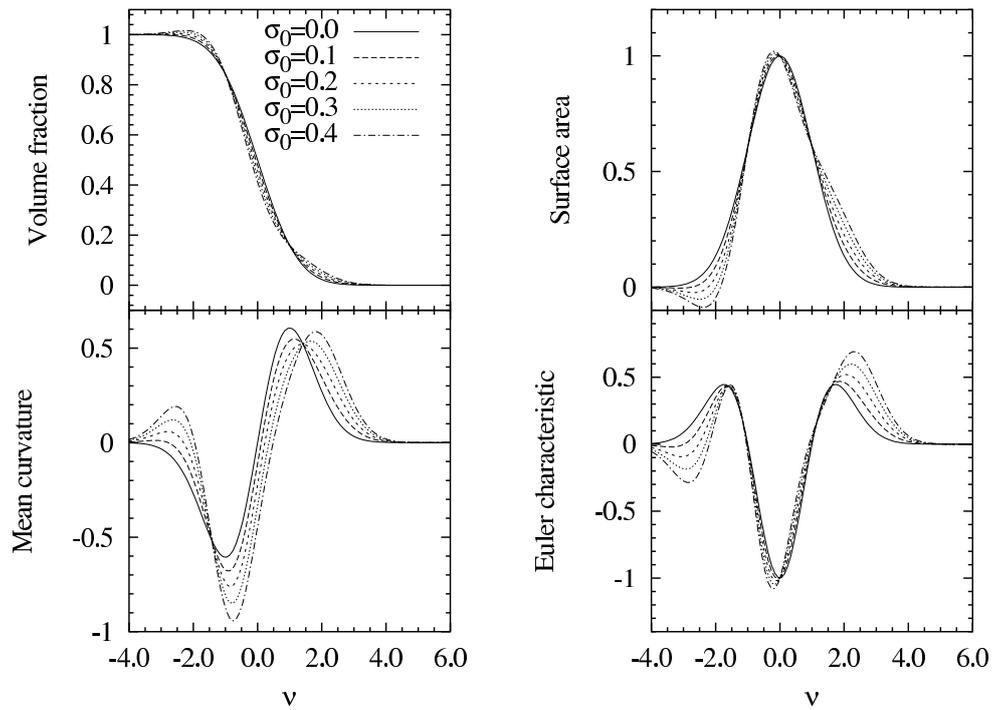


**Figure 5.** The genus of such isodensity contours is  $3 - 2 = 1$ .

if the initial perturbation is Gaussian.

- it is important to have other statistics to probe the non-Gaussian features in the Large-scale structure.

- [1] Abazajian, K., et al. 2005, *Astrophys. J.*, **625**, 613
- [2] Bennett, C. L., et al. 2003, *Astrophys. J. Suppl.*, **148**, 1
- [3] Kayo, I., et al. 2004, *Publ. Astron. Soc. Japan*, **56**, 415
- [4] Matsubara, T. 2004, *Astrophys. J.*, **615**, 573
- [5] Park, C., et al. 2005, *The Astrophys. J.*, **633**, 11
- [6] Totsuji, H., & Kihara, T. 1969, *Publ. Astron. Soc. Japan*, **21**, 221
- [7] Zehavi, I., et al. 2004, *Astrophys. J.*, **608**, 16
- [8] Strite S and Morkoc H 1992 *J. Vac. Sci. Technol. B* **10** 1237
- [9] Jain S C, Willander M, Narayan J and van Overstraeten R 2000 *J. Appl. Phys.* **87** 965



**Figure 6.** Analytical prediction of the Minkowski functionals in weakly non-Gaussian fields.

- [10] Nakamura S, Senoh M, Nagahama S, Iwase N, Yamada T, Matsushita T, Kiyoku H and Sugimoto Y 1996 *Japan. J. Appl. Phys.* **35** L74
- [11] Akasaki I, Sota S, Sakai H, Tanaka T, Koike M and Amano H 1996 *Electron. Lett.* **32** 1105
- [12] O'Leary S K, Foutz B E, Shur M S, Bhapkar U V and Eastman L F 1998 *J. Appl. Phys.* **83** 826
- [13] Jenkins D W and Dow J D 1989 *Phys. Rev. B* **39** 3317