# Statistical properties of cosmological correlation functions

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Philipp Wilking
aus
Wissen (Sieg)

Bonn, November 2014

Dieser Forschungsbericht wurde als Dissertation von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem Hochschulschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

| | |
|---|---|
| 1. Gutachter: | Prof. Dr. Peter Schneider |
| 2. Gutachter: | Prof. Dr. Cristiano Porciani |
| Tag der Promotion: | 25.03.2015 |
| Erscheinungsjahr: | 2015 |

# Abstract

Correlation functions are an omnipresent tool in astrophysics, and they are routinely used to study phenomena as diverse as the large-scale structure of the Universe, time-dependent pulsar signals, and the cosmic microwave background. In many cases, measured correlation functions are analyzed in the framework of Bayesian statistics, which requires knowledge about the likelihood of the data. In the case of correlation functions, this probability distribution is usually approximated as a multivariate Gaussian, which is not necessarily good approximation – hence, this work aims at finding a better description.

To this end, we exploit fundamental mathematical constraints on correlation functions, which we use to construct a quasi-Gaussian likelihood. We explain how to compute the constraints, in particular for multi-dimensional random fields, where this can only be done numerically, check the quality of the quasi-Gaussian approximation, and compare it to alternative approaches – most importantly, we test the new-found description of the likelihood in a toy-model Bayesian analysis.

Finally, we compute correlation functions from the Millennium Simulation and show that they obey the constraints. By studying statistical properties of the measured correlation functions, we present further indications for the validity of the quasi-Gaussian approach.

# Contents


**1 Introduction** **1**

**2 Cosmology and statistics** **5**
2.1 The standard model . . . 5
2.1.1 World models . . . 5
2.1.2 A brief history of the Universe . . . 7
2.1.3 Cosmological parameters . . . 9
2.1.4 Problems of the standard model . . . 11
2.2 Statistical methods in cosmology . . . 12
2.2.1 Random fields . . . 12
2.2.2 Two-point statistics . . . 13
2.2.3 Higher-order statistics . . . 15
2.2.4 Bayesian methods . . . 16

**3 Constraints on correlation functions** **21**
3.1 Analytical results . . . 21
3.2 Numerical methods . . . 23
3.3 Comparison of the constraints for one-dimensional fields . . . 25

**4 A new approximation for the probability distribution of correlation functions** **31**
4.1 Simulations . . . 31
4.2 Analytical work . . . 32
4.3 The quasi-Gaussian approach . . . 32
4.3.1 Quality of the quasi-Gaussian approximation . . . 34
4.4 Analytical calculation of mean and covariance matrix . . . 36
4.4.1 Calculations in $\xi$-space . . . 36
4.4.2 Transformation of mean and covariance matrix to $y$-space . . . 41
4.5 Alternative approaches . . . 43
4.5.1 A copula approach . . . 43
4.5.2 Box-Cox transformations . . . 45

**5 Performance of the quasi-Gaussian approximation in a likelihood analysis** **47**
5.1 Setup . . . 47
5.2 Components of the quasi-Gaussian likelihood . . . 48
5.3 Results . . . 50

**6 Constrained correlation functions from the Millennium Simulation** **53**
6.1 $N$-body simulations . . . 53
6.2 The Millennium Simulation . . . 54
6.3 Computing correlation functions . . . 56
6.4 Testing the constraints . . . 60
6.5 Quality of the Gaussian approximation in $\xi$ and $y$-space . . . 66

**7 Conclusion** **71**
7.1 Summary . . . 71
7.2 Outlook . . . 72

**A Acknowledgments** **75**

**B Bibliography** **77**

**C List of figures** **85**

**D List of tables** **87**

**E List of acronyms** **89**

CHAPTER 1

# Introduction

In the past decades, there have been remarkable advances in cosmology, finally culminating in the emergence of a cosmological standard model, which is widely accepted despite the fact that it still faces some challenges on small scales and, even more noteworthy, states that our Universe consists largely of "dark" components, which are not yet well understood. Due to the unique situation of cosmology as a science that deals with the Universe as a whole, this feat has only been possible with the help of sophisticated tools – especially considering that, as Martin Rees put it, "our brains are made for survival in the African savanna, not for cosmology"[1]. Among the most powerful "weaponry" of a modern astronomer are statistical methods, mainly owing to the fact that large parts of astronomical data are statistical in nature, be it the distribution of galaxies in the sky, cosmic shear signals from gravitational lensing by large-scale structure (LSS) in the Universe, or time-series measurements of pulsars – in fact, one of the most ground-breaking recent astronomical results, namely the determination of cosmological parameters from the cosmic microwave background (CMB) with unprecedented accuracy by the Planck Collaboration *et al.* (2014b), would not have been possible without refined statistical tools. With upcoming surveys and instruments like Euclid becoming more and more ambitious, it is to be expected that the requirements to the astronomers' statistical tool set will rise even higher.

It is striking, however, that despite the high level of sophistication of this tool set, even in the very simplest of mathematical setups, not everything is understood – in particular, there are still gaps in our knowledge about something as (seemingly) basic as two-point statistics of one-dimensional Gaussian random fields, which is the key motivation for this thesis.

The two-point correlation function $\xi$ is a very common tool in cosmology (in fact, a remarkable percentage of astronomical literature already deals with higher-order statistics), and, whenever correlation function measurements are used in a Bayesian framework in order to determine cosmological parameters, the probability distribution function (PDF) of the correlation function is needed. Usually, this likelihood is assumed to be Gaussian; examples include an analysis of the CMB correlation function by Seljak and Bertschinger (1993) or common methods of baryon acoustic oscillations (BAO) detection (see e.g. Labatie *et al.* 2012a).

However, the use of Gaussian likelihoods does not necessarily provide the level of precision required from statistical tools that are used to analyze state-of-the-art astronomical data. For

[1] From an interview published in ZEITmagazin 31/08, translated from German.

one, objections against the use of Gaussian likelihoods as a "safe default" in cases where knowledge of the exact form of the likelihood is lacking have been raised: For example in the case of power spectrum estimators, Carron (2013) shows that using Gaussian likelihoods can assign too much information to the data and thus violate the Cramér-Rao inequality. Along the same lines, Sun *et al.* (2013) argue that the use of Gaussian likelihoods in power spectra analyses can have significant impact on the bias and uncertainty of estimated parameters.

More importantly, there are indications that, specifically in the case of correlation functions, the assumption of a Gaussian likelihood is problematic: Using independent component analysis, Hartlap *et al.* (2009) detect a significant non-Gaussianity in the likelihood in the case of a cosmic shear study. As a more fundamental argument, the proof of the existence of constraints on correlation functions (i.e. on the range of allowed values correlation functions can take) by Schneider and Hartlap (2009) from purely mathematical arguments shows that the likelihood of the correlation function cannot be truly Gaussian, since a Gaussian distribution would require infinite support.

Of course, non-Gaussian likelihood functions have been studied before: For the case of the cosmic shear power spectrum, Sato *et al.* (2011) use a Gaussian copula to construct a more accurate likelihood function. The use of a copula requires knowledge of the univariate distribution, for which the authors find a good approximation from numerical simulations. However, in the case of correlation functions, work aiming for a description of the non-Gaussian likelihood is sparse in the literature, and thus, a better description of the correlation function likelihood is sorely needed in order to "catch up" and allow for more precise constraints on cosmological parameters from correlation function measurements.

Several steps have been taken in this direction: Clearly, the most fundamental approach is to try and calculate the probability distribution of correlation functions, $p(\xi)$, analytically – to do so, Keitel and Schneider (2011) use a Fourier mode expansion of a Gaussian random field in order to obtain the characteristic function of $p(\xi)$. From it, they calculate the uni- and bivariate probability distribution of $\xi$ through a Fourier transform. Since this calculation turns out to be very tedious, an analytical computation of higher-variate PDFs (which would be required for cosmological parameter estimation) is probably not feasible.

A different approach, namely using a coordinate transformation suggested in Schneider and Hartlap (2009) (which makes use of the constraints derived in the aforementioned work) in order to Gaussianize the correlation function and thus obtain a new, "quasi-Gaussian" likelihood function for the correlation functions of one-dimensional Gaussian random fields, will be one of the pinnacles of this PhD thesis – some results of the quasi-Gaussian approach have previously been published in Wilking and Schneider (2013).

This work is structured as follows: In Sect. 2, we briefly introduce the cosmological standard model. In particular, we stress the connection between cosmology and statistics and highlight numerous statistical methods that are being used in astronomy. Sect. 3 deals with the constraints on correlation functions, summarizing both analytical and numerical results previously obtained and comparing them. The main focus of Sect. 4 are attempts to find a new approximation for the probability distribution of correlation functions – we describe analytical as well as numerical work in that direction, culminating in the quasi-Gaussian approximation, and also mention alternative approaches: In particular, we try to obtain the multivariate $p(\xi)$ by coupling the analytical univariate results with a Gaussian copula. Also, the strategy of transforming a random variable in order to obtain a well-known probability distribution suggests a comparison to similar attempts, e.g. using Box-Cox methods in order to find an optimal variable transformation to Gaussianity (see e.g. Joachimi *et al.* 2011). We will argue, however, that

both approaches do not seem to yield satisfactory results in our case.

Thus, in Sect. 5, we stick to the previously derived quasi-Gaussian likelihood and test its performance in a toy-model likelihood analysis, indicating its possible impact on cosmological parameter estimation. Sect. 6 summarizes our attempts to test our results in a more astrophysical context: To this end, we compute correlation functions from a cosmological $N$-body simulation and test the relevance of the constraints, as well as the Gaussianity of $p(\xi)$. We conclude with a short summary and outlook in Sect. 7.

CHAPTER 2

# Cosmology and statistics

In this chapter, we will give an overview of the theoretical background needed for this thesis. Since this work is thematically situated in the intersection of cosmology and statistics, we will give a brief overview of the prerequisites from both fields, starting with a summary of our current cosmological model in Sect. 2.1 and then highlighting the connection to statistics and introduce required statistical concepts in Sect. 2.2.

## 2.1 The standard model

As a science, cosmology has evolved significantly over the past decades, not only due to huge datasets and surveys produced by more and more sophisticated instruments, but also thanks to fruitful interactions with related scientific fields such as mathematics, computer science, and physics. These developments have given rise to a standard model of cosmology, the so-called ΛCDM model. In this chapter, we will briefly summarize this model and its theoretical foundations, tackle the problem of how to mathematically describe the Universe as a whole (Sect. 2.1.1) as well as outline our understanding of the history of the Universe (Sect. 2.1.2). In particular, we will explain how to obtain the parameters of the model using different methods and give their current best values in Sect. 2.1.3. More detailed introductions to cosmology can be found in numerous textbooks, e.g. Schneider (2006) or Peacock (1999).

Finally, although the standard model is widely accepted as current state-of-the-art, it is not devoid of problems – we will highlight some in Sect. 2.1.4.

### 2.1.1 World models

When describing the Universe as a whole, i.e. on cosmological scales, gravity is the only relevant interaction, hence we have to work within the framework of General Relativity (GR). Thus, one of the most fundamental equations in cosmology is Einstein's field equation:

$$G_{\mu\nu} = -\frac{8\pi G}{c^4} T_{\mu\nu} - \Lambda g_{\mu\nu}. \tag{2.1}$$

Einstein's theory aims for a fully covariant formulation and thus operates in a (four-dimensional) space-time, thus the most fundamental quantities are tensors with indices (e.g. $\mu$, $\nu$) running from zero to three. In particular, $G_{\mu\nu}$ denotes the Einstein tensor, which is a function of the

metric tensor $g_{\mu\nu}$ and describes the geometry of space-time. On the other hand, the stress-energy tensor $T_{\mu\nu}$ encodes the matter and energy content of the Universe; the additional term in Eq. (2.1) includes the famous cosmological constant $\Lambda$, which Einstein originally introduced to allow for a static Universe, later discarding it. In modern cosmology, it has again found its place as a description of the Dark Energy component of the Universe.

While the mathematical framework of GR is quite complex, it is often helpful to operate in a more intuitive picture: Namely, the field equation describes how matter and energy bend space-time, thus influencing the paths (so-called geodesics) along which matter (and light) move in space-time.

In order to solve the field equation, one usually assumes the cosmological principle to hold: It states that the Universe is isotropic and homogeneous on large scales, a claim that is well justified e.g. from CMB measurements, though obviously not true on small scales. Under these assumptions, one exact solution of the field equation is the so-called Friedmann-Lemaître-Robertson-Walker (FLRW) metric (Robertson 1935; Walker 1937), which describes an expanding Universe. In the dynamics of expansion, the spatial location $\boldsymbol{r}$ is given as $\boldsymbol{r}(t) = a(t)\,\boldsymbol{x}$, where $\boldsymbol{x}$ denotes the comoving coordinate, and the scale factor $a(t)$ (by definition, it is equal to one at present time, $a(t_0) = 1$) incorporates the global expansion.

Using these definitions and plugging in the FLRW metric as well as the corresponding tensors, Eq. (2.1) yields the famous Friedmann equations:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda}{3}, \tag{2.2}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda}{3}. \tag{2.3}$$

Here, $p$ and $\rho$ denote pressure and density, respectively, and are related by the equation of state; $K$ is the curvature of the Universe, which can be used as a classification of cosmological models: $K < 0$ corresponds to an open, $K > 0$ to a closed Universe. Specializing the limiting case $K = 0$ to today (i.e. $t = t_0$) yields the so-called critical density $\rho_{\mathrm{cr}}$, i.e. the total energy density today corresponding to $K = 0$. It can be used to define the density parameters $\Omega_i = \rho_i/\rho_{\mathrm{cr}}$ of the different components of the Universe (denoted by indices), namely **r**adiation, **m**atter, and $\Lambda$, interpreted as Dark Energy (DE). The total energy density is denoted by $\Omega_0 \equiv \sum_i \Omega_i$, where the case $\Omega_0 = 1$ corresponds to a spatially flat Universe.

In this formulation, the Friedmann equations can be solved for the various components of the Universe, i.e. their densities with different scale factor dependence, which can be written in the form $\rho_i = \rho_{0i}\; a^{-\nu_i}$, where $\rho_{0i}$ denotes the current density (recall $a(t_0) = 1$). This power-law ansatz is motivated by the first law of thermodynamics and the components' equations of state, which relate pressure and density – for example, in the case of (pressure-less) matter, $p = 0$, resulting in $\nu_{\mathrm{m}} = 3$. In this context, Eq. (2.3) is usually rewritten using the Hubble parameter $H(t) = \dot{a}/a$ (whose value today is the Hubble constant $H_0 \equiv H(t_0)$), yielding

$$H^2 = H_0\left(\frac{\Omega_{\mathrm{r}}}{a^4} + \frac{\Omega_{\mathrm{m}}}{a^3} + \frac{1-\Omega_0}{a^2} + \Omega_\Lambda\right). \tag{2.4}$$

One important consequence of the expanding Universe is cosmological redshift: Assuming a distant source emits a photon of wavelength $\lambda_{\mathrm{e}}$, this wavelength is stretched due to cosmic expansion, resulting in an observed wavelength $\lambda_{\mathrm{obs}}$ which is higher by a factor equal to the scale factor (correspondingly, the observed frequency is lower than the emitted one). Thus, one

defines the redshift $z$ as

$$(1+z) \equiv \frac{\nu_\mathrm{e}}{\nu_\mathrm{obs}} = \frac{\lambda_\mathrm{obs}}{\lambda_\mathrm{e}} = \frac{1}{a(t)}. \tag{2.5}$$

Assuming the emitted wavelength is known (and shifts due to the local Doppler effect can be neglected), $z$ can be used as a direct proxy for the distance of the source, depending, of course, on the underlying cosmological model whose parameters determine the scale factor. Thus, measuring redshifts (ideally by taking spectra; however, redshift determination by photometric methods is also possible, though less accurate) is an important tool for distance estimation of far-away sources, which remains one of the biggest challenges in astronomy. In this context, it is important to note that the concept of distance in cosmology merits some discussion, since in an expanding (and potentially curved) Universe, distance is not a straightforward concept: For example, assuming the angular diameter of a source at a given redshift is known, it is possible to determine the (comoving) distance from the Friedmann equations, which of course requires an underlying cosmological model. In an alternative scenario, we might know the luminosity of a source, allowing us to measure its flux and determine the distance from these quantities. In Euclidean space, the two measurements would yield the same result, but in general, angular diameter distance and luminosity distance are not equal.

### 2.1.2 A brief history of the Universe

The evolution of our Universe is in fact a thermal history, since it can be expressed in terms of temperature or energy, which decrease as the Universe expands from its singular, hot initial state to its current form. It can be shown from the Friedmann equations that there must have been a point in the past where the scale factor $a$ was very close to zero, formally corresponding to an "infinitesimally small" Universe. This initial state is called the Big Bang, and while we can trace back the history of the cosmos to very early times, the Big Bang itself is still not completely understood – in particular due to the fact that the interpolation $t \to 0$ involves energy scales at which known physics breaks down, and a Grand Unified Theory (GUT), i.e. a theory that describes both electroweak and strong interactions in a generalized framework, would be required. While the GUT scale is still out of reach for today's particle accelerators such as the Large Hadron Collider (LHC), future experiments might shed light on the GUT epoch.

Still, we have a fairly good idea about the very early stages of the Universe, namely about the era of inflation – a very fast exponential expansion of the Universe that happened immediately after the Big Bang. Inflation is in fact not considered part of the standard model, but rather an extension of it that has been introduced to solve problems in the standard expansion history, namely the flatness problem (the fact that the Universe is almost flat today, requiring a fine-tuning at early time to guarantee $\Omega_0 \approx 1$) as well as the horizon problem, i.e. the homogeneity of the Universe even between regions that would never have been in causal contact without an early rapid expansion. While inflation serves to solve these problems, its exact picture is still being debated and a huge variety of inflation models exist (see Martin *et al.* 2013 for an exhaustive compilation as well as a more thorough introduction to inflation). As a final remark on inflation, it is important to note that until recently, there was no direct proof for inflation. However, measurements of the CMB polarization by the BICEP2 experiment (BICEP2 Collaboration *et al.* 2014) show a clear signal originating from inflationary gravitational waves, thus providing not only the first direct experimental evidence of the era of inflation, but also the first detection of gravitational waves.

In the description of the (standard) expansion history of the Universe, the equilibria and decoupling processes between the different matter components play a crucial role: In the early Universe, high temperatures allow particles to be in (thermodynamic) equilibrium, since the reaction rates for processes that convert particles (e.g. $e^+e^- \leftrightarrow \gamma\gamma$ for electrons and photons) are high enough. As the Universe cools down and reaction rates decrease, the various species decouple successively, namely neutrinos (at a temperature of 1.4 MeV) and, later, protons and neutrons (at 0.7 MeV). Once the temperature drops below 0.5 MeV, electron-positron pairs can no longer be created, all remaining anti-matter is annihilated, and finally, at around 0.1 MeV, the first atomic nuclei (most importantly $^4$He) form – notably, the evolution described so far covers "only" the first few minutes after the Big Bang.

When studying the creation of atomic nuclei in the early Universe, usually called Big Bang nucleosynthesis (BBN), it is possible to compute number densities and mass fractions of different nuclei. In fact, the standard model predicts a helium abundance of 25 %, which can be used as a rigorous test of the model.

After BBN, the Universe experiences a long, less eventful era. At this point, the main matter components are protons, helium nuclei, electrons, photons, neutrinos, and Dark Matter (DM), believed to consist of weakly interacting massive particles (WIMPs). Below temperatures of about 1 eV, nuclei (mostly protons) and electrons can form atoms; this era is called recombination (although it actually happens for the first time, since up to that point, all baryonic matter was ionized). Shortly after the Universe becomes neutral, it also becomes transparent, since matters and photons decouple, allowing the latter to stream freely. This marks the origin of the CMB, which today allows us a glance at the Universe during the epoch of recombination, i.e. about 380 000 years after the Big Bang, corresponding to a redshift of about 1100. Due to the ongoing expansion of the Universe, the CMB has since cooled down from a few thousand to 2.7 K, and it has been studied in great detail using satellite missions such as the Cosmic Background Explorer (COBE), the Wilkinson Microwave Anisotropy Probe (WMAP), and Planck. While it shows an almost perfect black body spectrum, there are small anisotropies imposed on it, which constitute one major source for determining cosmological parameters – we will give results from CMB experiments in Sect. 2.1.3.

It is important to mention that, despite recombination, the Universe today is essentially completely ionized: The so-called Gunn-Peterson test, i.e. the search for a drop in the spectrum of very distant quasi-stellar objects (QSOs) (which would result from redshifted Ly$\alpha$ lines of neutral hydrogen along the line of sight) yields that practically all hydrogen in the intergalactic medium (IGM) is ionized. While the epoch of reionization is not yet understood in detail, it has most likely been caused by the first stars or active galactic nuclei (AGN).

So far, we have only considered a homogeneous Universe – however, if we want to understand how the structures we observe today have formed, we obviously have to consider deviations from homogeneity. Mathematically, these can be expressed in terms of the density contrast $\delta$, which describes how the density at a given location differs from the average density of the Universe:

$$\delta(\boldsymbol{x}, t) = \frac{\rho(\boldsymbol{x}, t) - \hat{\rho}(t)}{\hat{\rho}(t)}. \tag{2.6}$$

As mentioned before, the seeds of structure formation are already visible in our earliest images of the Universe, namely in the CMB anisotropies, which are tiny perturbations of order $10^{-5}$ originally caused by quantum fluctuations. To describe their evolution, one commonly uses perturbation theory, so one has to solve the fluid equations. Assuming that the perturbations

are small and linear theory holds, this can be done analytically, and for example the scale factor dependence of $\delta$ for radiation and matter during different epochs can be computed.

However, as perturbations become larger (most notably on small scales), non-linear structure growth sets in. In this case, one has to resort to numerical methods (we will go into details on numerical simulations of structure formation in Sect. 6.1 and also show some exemplary output of such a simulation, see Fig. 6.1). In large $N$-body simulations, one tries to reproduce observable structures in the cosmos, such as galaxies, galaxy clusters, and the filaments and voids of the large-scale structure (LSS). One striking result from these simulations, which seems to agree quite well with observations, is that the mass profiles of DM halos of very different sizes can be approximated by a universal law, the so-called Navarro–Frenk–White (NFW) profile (Navarro *et al.* 1996).

At this point, one connection to statistics becomes apparent: Obviously, we cannot expect the results of numerical simulations to look exactly like our Universe, but rather to share some of its (statistical) properties. Thus, in order to describe and compare these results, we have to use statistical methods such as correlation functions, power spectra, and related statistical quantities. In this context, the density contrast $\delta$ is understood as a realization of a random field, as we will describe in more detail in Sect. 2.2.

### 2.1.3 Cosmological parameters

When describing our state of knowledge about the Universe as a whole, one usually tries to summarize our current cosmological model in a few numbers, the so-called cosmological parameters. Their respective values have drastic influence on the expansion behavior and geometry of the Universe; as an example, Fig. 2.1 shows a classification of cosmological models in the $\Omega_\mathrm{m} - \Omega_\Lambda$-plane.

While numerous fundamentally different methods of obtaining cosmological parameters exist (see for example Lahav and Liddle 2014 for a brief review), they all agree that the dominating components in the Universe today are Dark Energy and Cold Dark Matter (hence the name ΛCDM model). In the following, we will list some methods of measuring cosmological parameters: One of the most powerful tools of modern cosmology, the CMB angular power spectrum of fluctuations, can constrain practically all cosmological parameters. In addition to that, measuring the expansion rate of the Universe allows us to determine the Hubble constant $H_0$ directly, as done most famously by the Hubble Key Project (Freedman *et al.* 2001), but also by time-delay measurements in strong gravitational lensing (Suyu *et al.* 2013). Type Ia supernovae (SNe) are almost "standard candles", so by applying a correction based on the shape of their light-curves, it is possible to obtain their absolute magnitudes, which can then be used as a distance indicator. Finally, X-ray observations of galaxy clusters as well as galaxy redshift surveys allow us to determine the baryon and matter content of the Universe. In conclusion, it is important to stress that not only have all cosmological parameters been measured by independent methods, but more remarkably, these distinct measurements establish a consistent framework.

One difficulty in the description of a cosmological model is the definition of a minimal set of parameters, since they are, in part, interdependent. Assuming a spatially flat Universe (which is well justified by measurements), the minimum number of base parameters is usually taken to be six; as an example, Table 2.1 shows one of the most precise (and most recent) sets of measurements, namely from the Planck CMB satellite. Here, the first six rows show the base parameters, which were chosen to be the baryon and CDM density parameters today

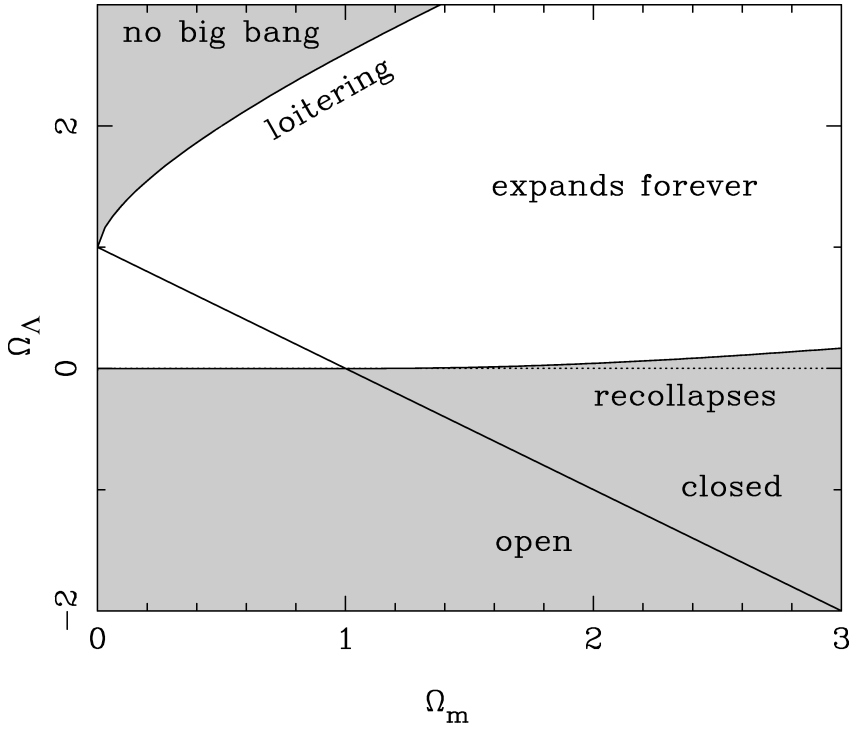


Figure 2.1: Different cosmological models classified by their expansion behavior and spatial geometry. The diagonal line marks flat models (adapted from Peacock 1999).

(multiplied by $h^2$, where $h$ denotes the Hubble constant in units of 100 km/s/Mpc), the angular size of the sound horizon at recombination $\theta_{\mathrm{MC}}$, the optical depth after reionization, the slope of the primordial power spectrum (assumed to be a power-law, $P(k) \propto k^{n_s}$), and the log power of the primordial curvature perturbations. The lower six rows show derived parameters, namely the DE and matter density parameters, the root mean square (RMS) matter fluctuations at $8h^{-1}$ Mpc, the reionization redshift, the Hubble constant, and the age of the Universe.

It is important to mention that the measurements from Planck are in some tension with other measurements of the Hubble constant, including previous CMB measurements by WMAP (Hinshaw *et al.* 2013) – a large compilation of cosmological parameters from many sources from 1990 to 2010 (collected by Croft and Dailey 2011) as well as a comparison to Planck results using median statistics can be found in Crandall and Ratra (2013). There are attempts to explain the tension between Planck and other results: In the case of the Hubble constant, Schneider and Sluse (2013) show that measurements from strong lensing time-delays are subject to a potential bias, which can be large enough to explain the discrepancy between the $H_0$ values from strong lensing and Planck. As a further example, Li *et al.* (2014) use a different calibration of light-curve fitting parameters in Type Ia SNs distance estimation in order to improve compatibility between the measurements. On a more theoretical note, Wyman *et al.* (2014) propose an additional massive sterile neutrino species – however, despite those attempts, it is not yet clear how to explain the discrepancies between Planck and other measurements.

Table 2.1: Cosmological parameters for the best-fit ΛCDM model, adapted from Planck Collaboration *et al.* (2014a,b) (see text for an explanation of the parameters). The columns show the best-fit values and $1\sigma$ limits obtained when using only the Planck temperature maps plus lensing; for the rightmost columns, also external data sets (WMAP polarization [WP] at low multipoles $\ell$, high-$\ell$ experiments, as well as BAO) were used (note that the value for $\Omega_\mathrm{m}$ in the rightmost columns only makes use of Planck maps and WMAP polarization).

| | Planck (CMB+lensing) | | Planck+WP(+highL+BAO) | |
|---|---|---|---|---|
| Parameter | Best fit | 68 % limits | Best fit | 68 % limits |
| $\Omega_\mathrm{b}h^2$ . . . . . | 0.022242 | $0.02217 \pm 0.00033$ | 0.022161 | $0.02214 \pm 0.00024$ |
| $\Omega_\mathrm{c}h^2$ . . . . . | 0.11805 | $0.1186 \pm 0.0031$ | 0.11889 | $0.1187 \pm 0.0017$ |
| $100\theta_\mathrm{MC}$ . . . . . | 1.04150 | $1.04141 \pm 0.00067$ | 1.04148 | $1.04147 \pm 0.00056$ |
| $\tau$ . . . . . | 0.0949 | $0.089 \pm 0.032$ | 0.0952 | $0.092 \pm 0.013$ |
| $n_\mathrm{s}$ . . . . . | 0.9675 | $0.9635 \pm 0.0094$ | 0.9611 | $0.9608 \pm 0.0054$ |
| $\ln(10^{10}A_\mathrm{s})$ . . . . . | 3.098 | $3.085 \pm 0.057$ | 3.0973 | $3.091 \pm 0.025$ |
| $\Omega_\Lambda$ . . . . . | 0.6964 | $0.693 \pm 0.019$ | 0.6914 | $0.692 \pm 0.010$ |
| $\Omega_\mathrm{m}$ . . . . . | 0.3036 | $0.307 \pm 0.019$ | 0.3183 | $0.315^{+0.016}_{-0.018}$ |
| $\sigma_8$ . . . . . | 0.8285 | $0.823 \pm 0.018$ | 0.8288 | $0.826 \pm 0.012$ |
| $z_\mathrm{re}$ . . . . . | 11.45 | $10.8^{+3.1}_{-2.5}$ | 11.52 | $11.3 \pm 1.1$ |
| $H_0$ . . . . . | 68.14 | $67.9 \pm 1.5$ | 67.77 | $67.80 \pm 0.77$ |
| Age/Gyr . . . . . | 13.784 | $13.796 \pm 0.058$ | 13.7965 | $13.798 \pm 0.037$ |

### 2.1.4 Problems of the standard model

Apart from the aforementioned issues regarding the values of some cosmological parameters determined from Planck data, the ΛCDM model faces several additional challenges. Most of them concern the immediate vicinity of our Milky Way: For example, some properties of dwarf galaxies in the Local Group, in particular their anisotropic spatial distribution and low number, are not in good agreement with expectations from simulations. To explain this, it has been suggested to modify the laws of gravity (see for example Kroupa *et al.* 2012), e.g. by applying Modified Newtonian Dynamics (MOND), but by now, relativistic extensions of MOND exist as well. Some of these theories seem to be quite successful in explaining observations on small scales and, in parts, also on large scales, but there, they also require the introduction of an additional DM component.

However, since most of these challenges of the standard model arise either on small scales (where ΛCDM simulations run into problems, since more complex physics than simple gravitational interactions become relevant) or rely on low-number statistics (in extreme cases being limited to very few or even single objects, e.g. galaxy clusters considered too heavy to be compatible with ΛCDM), the general consensus seems to be that most likely, these issues can be remedied without a major change of paradigm, in particular since GR is one of the most thoroughly tested theories in modern physics.

Nonetheless, it is important to keep in mind that fundamental parts of the cosmological standard model are still unexplained, in particular the nature of DM and DE, which are still very much science in progress: As mentioned before, the search for DM candidates, for example

at the LHC, is still ongoing, and numerous other experiments attempting direct and indirect detection of DM particles exist (see Arrenberg *et al.* 2013 for a review on complementary DM searches). One interesting recent result in this context is the discovery of a new, weak line in the X-ray spectra of two DM-dominated objects (namely the Andromeda galaxy and the Perseus galaxy cluster) by Boyarsky *et al.* (2014), which might be a possible signature of DM decay, but requires further studies.

The nature of DE poses an even greater mystery: While we do have some knowledge about it (for example, weak gravitational lensing, galaxy cluster studies, and BAO already allow us to constrain the DE equation of state; see e.g. Davis 2014 for a qualitative review of different ways DE has been measured), its origin is still unclear. Even worse, quantum-mechanical predictions of the DE density are off by many orders of magnitudes, indicating that our theoretical understanding in this respect is still very incomplete. Future missions, in particular X-ray observatories such as eROSITA, will hopefully deliver insights into the nature of DE.

## 2.2 Statistical methods in cosmology

In Sect. 1 and 2.1.2, we already mentioned some examples of applied statistics in astrophysics – in this chapter, we will give a more detailed introduction to key statistical concepts and introduce some notation (for a more detailed introduction to statistics and probability theory from a mathematical point of view, we refer to Hogg *et al.* 2005). We will mostly concentrate on basic quantities relevant for this work since giving a complete overview of the field is not possible in this framework – in recent years, astrostatistics has become a viable science of its own right, with a very vivid community and even attempts to establish a dedicated journal and a separate PhD.

The key motivation for a statistical approach to astrophysics is twofold: On the one hand, the LSS of the Universe is inherently statistic because of its quantum origins. Furthermore, due to the nature of many astronomical observations (e.g. probing a patch of sky which is assumed to be representative, but random), astronomical data often comes in the form of random fields. Thus, this is the first concept we will introduce in Sect. 2.2.1; in the following sections, we will deal with the challenge of how to characterize a random field, since the field itself is seldom directly observable. Therefore, we will give an overview of two-point statistics, but also briefly venture into higher-order statistics, which are important tools nowadays, though we do not make use of them in the main body of this work.

### 2.2.1 Random fields

For most practical purposes, it is sufficient to have a graphic understanding of the concept of a random field, i.e. to think of it simply as spatial data: At different locations, a certain quantity is measured, and this set of measurements (which are stochastic in nature) forms a random field. To give a mathematically sound characterization, however, a random field has to be defined via a mapping from an index space $\mathcal{I}$ (e.g. $\mathcal{I} = \mathbb{R}^3$ for a three-dimensional Cartesian coordinate system) to a target space $\mathcal{T}$ (in the case of a scalar field, for example, $\mathcal{T} = \mathcal{R}$). This mapping $g : \mathcal{I} \to \mathcal{T}$ is one realization of a random field, which in turn is fully characterized by its probability. Thus, the most fundamental way to define a random field is to understand it as a mapping from a realization $g$ to its probability, $g \to \mathbb{R}_0^+$. However, for the sake of readability, one usually does not distinguish between the mapping and a realization and simply refers to $g$

as "the random field"; we will also follow this convention. In the following, we introduce some notation.

As a first important concept, we define the Fourier transform of a random field $g(\boldsymbol{x})$ as

$$\tilde{g}(\boldsymbol{k}) = \int \mathrm{d}^n x \; g(\boldsymbol{x}) \; \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{2.7}$$

In practice, especially in the case of numerical simulations of random fields, it is necessary to evaluate the field at discrete grid points, i.e. simulate it in an $n$-dimensional periodic cube with side length $L$ in real space (which is equivalent to introducing a grid in $\boldsymbol{k}$-space such that the wave number can only take integer multiples of the smallest possible wave number $\Delta k = 2\pi/L$). In order to span the whole space $\mathbb{R}^n$, we can then choose $L$ large enough and introduce periodic boundary conditions. The discretized form of the inverse of Eq. (2.7) reads

$$g(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \tilde{g}_{\boldsymbol{k}} \; \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \tag{2.8}$$

with the Fourier coefficients

$$\tilde{g}_{\boldsymbol{k}} = \left(\frac{\Delta k}{2\pi}\right)^n \tilde{g}(\boldsymbol{k}). \tag{2.9}$$

In this work, we will mostly be dealing with Gaussian random fields, i.e. fields where the Fourier components $\tilde{g}_{\boldsymbol{k}}$ are statistically independent and follow Gaussian distributions. Gaussian random fields play a crucial role in cosmology, as we will see in Sect. 2.2.3, and have some convenient mathematical properties: In particular, a Gaussian field is fully described by its power spectrum (which we will define in the next section); additionally, assuming a continuous power spectrum, a Gaussian field is always ergodic: This means that, whenever an ensemble average of a field-related quantity needs to be taken, it can be replaced by a spatial average. Obviously, this property is vital in many astronomical applications where only one realization of the random field is available – this can then be remedied by observing different areas of the sky.

### 2.2.2 Two-point statistics

The simplest way of analyzing random fields is two-point statistics, i.e. using correlation functions and power spectra. The two-point correlation function of the field is defined as $\xi(\boldsymbol{x}, \boldsymbol{y}) = \langle g(\boldsymbol{x}) \; g^*(\boldsymbol{y}) \rangle$ ; it is the Fourier transform of the power spectrum:

$$\xi(|\boldsymbol{x}|) = \int \frac{\mathrm{d}^n k}{(2\pi)^n} P(|\boldsymbol{k}|) \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}) = \int \frac{\mathrm{d}k}{2\pi} k^{n-1} P(k) \; Z_n(kx), \tag{2.10}$$

where the function $Z_n(\eta)$ is obtained from integrating the exponential over the direction of $\boldsymbol{k}$. In particular, $Z_2(\eta) = J_0(\eta)$ and $Z_3(\eta) = j_0(\eta)$, where $J_0(\eta)$ denotes the Bessel function of the first kind of zero order and $j_0(\eta)$ is the spherical Bessel function of zero order. In the one-dimensional case, Eq. (2.10) becomes a simple cosine transform,

$$\xi(x) = \int_0^\infty \frac{\mathrm{d}k}{\pi} \; P(|k|) \cos(kx). \tag{2.11}$$

As mentioned before, the central property of a Gaussian random field is that it is entirely specified by its power spectrum, which is given as a function of $\sigma(|\boldsymbol{k}|)$:

$$P(|\boldsymbol{k}|) = \left(\frac{2\pi}{\Delta k}\right)^n \left\langle |\tilde{g}_{\boldsymbol{k}}|^2 \right\rangle = \left(\frac{2\pi}{\Delta k}\right)^n \sigma^2(|\boldsymbol{k}|). \tag{2.12}$$

The definition of the correlation function as an ensemble average of the underlying random field is mainly constructed from a theoretical point of view, since the random field is the mathematically fundamental quantity in statistics. However, the field itself cannot be measured directly in many practical applications and thus, a different approach is required. In the case of galaxy surveys, one defines the correlation function via the probability d$p$ of finding a galaxy in a volume element d$V$ at distance $r$ from another galaxy:

$$\mathrm{d}p = \bar{n}\,[1 + \xi\,(r)]\;\mathrm{d}V, \tag{2.13}$$

where $\bar{n}$ denotes the average galaxy density. Thus, in order to measure $\xi$, one has to make use of pair counting methods: Namely, one counts the number $dd(s)$ of galaxy pairs for different separations $s$ in the data catalog and in a random catalog ($rr(s)$) with uniform galaxy distribution, as well as the cross-correlation $dr(s)$. This procedure shows an advantage of using correlation functions instead of power spectra in the statistical analysis of galaxy surveys, namely, it is quite easy to deal with complicated survey geometries (e.g. due to masks) by simply applying the same mask to the random and the data catalog. Since the catalogs can differ in size (in particular, it is advisable to generate a large random catalog in order to avoid introducing unnecessary noise), these count rates have to be normalized by the total number of galaxy pairs, yielding

$$DD(s) = \frac{dd(s)}{1/2\; n_\mathrm{d}\,(n_\mathrm{d} - 1)} \tag{2.14}$$

$$RR(s) = \frac{rr(s)}{1/2\; n_\mathrm{r}\,(n_\mathrm{r} - 1)} \tag{2.15}$$

$$DR(s) = \frac{dr(s)}{n_\mathrm{d} n_\mathrm{r}}, \tag{2.16}$$

where $n_\mathrm{d}$ and $n_\mathrm{r}$ denote the number of galaxies in the data and random catalog, respectively.

From these count rates, the correlation function can be computed – however, different estimators have been used in the literature (see Table 2.2 for a compilation of the most common ones). Various studies have been conducted in order to gauge the performance of the various estimators (see for example Kerscher *et al.* 2000 or Vargas-Magaña *et al.* 2013 and references therein), and in general, the Landy-Szalay (LS) estimator (Landy and Szalay 1993) is considered to show the best properties (in particular, low bias and variance). Under some simplifying assumptions (namely infinitely large random catalogs, large survey volumes, and vanishing correlation function), it can even be shown analytically to have minimal variance.

It is worth mentioning that, although the LS estimator is quite well established, the search for more sophisticated estimators is still ongoing: For example, Vargas-Magaña *et al.* (2013) show that the LS estimator does not reach the Poisson noise limit in the case of a non-vanishing correlation function and propose a way to find an optimized estimator. The iterative procedure they set up is based on the observation that all commonly used estimators (i.e. the ones listed in Table 2.2) are linear combinations of ratios of pair counts – thus, they suggest to use all pair

Table 2.2: Different correlation function estimators (adapted from Vargas-Magaña *et al.* 2013).

| | |
|---|---|
| $\hat{\xi}_{\mathrm{PH}} = \frac{DD}{RR} - 1$ | Peebles and Hauser (1974) |
| $\hat{\xi}_{\mathrm{Hew}} = \frac{DD - DR}{RR}$ | Hewett (1982) |
| $\hat{\xi}_{\mathrm{DP}} = \frac{DD}{DR} - 1$ | Davis and Peebles (1983) |
| $\hat{\xi}_{\mathrm{H}} = \frac{DD \times RR}{DR^2} - 1$ | Hamilton (1993) |
| $\hat{\xi}_{\mathrm{LS}} = \frac{DD - 2DR + RR}{RR}$ | Landy and Szalay (1993) |

count ratios up to second order (i.e. terms of the form $DD/RR$, $DR/RR$, ..., $DR \times RR/DD^2$, $RR^2/DD^2$, ...) and construct an estimator (called Vargas-Magaña (VM) estimator in the following) as a weighted sum of all these ratios. The most crucial step is to optimize the coefficients in the summation, which they do using a $\chi^2$ minimization over a large set of mock catalogs, finally yielding a correlation function estimator that has, by construction, a lower variance than the LS estimator, but might be biased. This bias can, in principle, be calculated and accounted for, however, it does show a cosmology dependence. In several tests, Vargas-Magaña *et al.* (2013) show that the optimized estimator can yield a $20 - 25\%$ improvement in the error bars of $\xi$ for Sloan Digital Sky Survey (SDSS) geometry and a $\Lambda$CDM correlation function (see Ahn *et al.* 2014 for the latest SDSS data release and details on the survey).

A different approach to find an optimized correlation function estimator is performed in Jones Baxter and Rozo (2013): Assuming a Gaussian likelihood for the observables (i.e. the count rates $D$, $R$, $DR$, ...), they optimize the correlation function estimator in such a way that it maximizes the likelihood. The resulting estimator can be shown to converge to the LS estimator in the limit of infinite survey volume. While the authors stress the speed and universality of their approach when compared to the VM estimator (although the latter yields a more drastic improvement in the error bars of $\xi$ in some cases), the assumption of multivariate Gaussianity for the count rates does not seem well motivated.

In summary, the level of sophistication in the methods used to optimize correlation function estimators may yield improvements in the estimator properties under certain conditions, however, it makes a careful choice of estimator even more important.

### 2.2.3 Higher-order statistics

In this section, we will briefly present extensions of the statistical methods discussed so far, namely higher-order (or $n$-point) statistics – an extensive introduction to this topic from both a theoretical and an observational point of view can be found in the review on LSS and perturbation theory by Bernardeau *et al.* (2002). Higher-order correlation functions and their Fourier counterparts (in the lowest orders, bi- and trispectra) are of high relevance in astronomy: For example, the CMB bispectrum has first been measured by the COBE satellite (see Komatsu *et al.* 2002), and the earliest measurements of the three-point correlation function from galaxy catalogs date back as far as Peebles (1975).

Mathematically, $n$-point correlation functions are defined analogously to the two-point case discussed in Sect. 2.2.2. However, since for example an estimator of the three-point function contains products of three count rates (e.g. $DDD$), the spatial arrangement of the three points can have a strong influence on the amplitude and shape of the correlation function – in this case, one has to distinguish between different triangle configurations, e.g. squished or equilateral ones.

The most relevant property of higher-order correlation functions for cosmological applications is the fact that they are zero for Gaussian random fields – as explained in Sect. 2.2.2, a Gaussian field is completely determined by its power spectrum (or two-point correlation function), and thus, $n$-point signals can be used as a probe of non-Gaussianity. An example that is of particular importance for cosmology is the quest to put constraints on the so-called primordial non-Gaussianity, i.e. on non-Gaussian components in the density field of the early Universe. Since the level of non-Gaussianity strongly depends on the exact characteristics of inflation, measuring it allows us to distinguish between different models of inflation, thus testing fundamental physics at very high energy scales.

However, the level of primordial non-Gaussianity also has an impact on the formation of structure in the Universe – for example, Dalal *et al.* (2008) show that it can result in a scale-dependent galaxy bias (i.e. a bias between the distribution of galaxies and the underlying matter). There have been numerous attempts to test and constrain various kinds of primordial non-Gaussianity, for example by measuring the galaxy power spectrum; however, the most precise constraints today again stem from CMB measurements: Recent results from Planck (Planck Collaboration *et al.* 2014c) including measurements of the CMB bispectrum yield no evidence for the most commonly studied kind of non-Gaussianity (the so-called "local" one), they do, however, find hints of unidentified non-Gaussianity signals.

### 2.2.4 Bayesian methods

Bayesian methods have been widely used in astronomy for many years and are an important foundation for this work, in particular due to the fact that they constitute a key motivation for investigating the probability distributions of physical quantities in the first place. Thus, in this section, we give a brief introduction to Bayesian statistics and mention some example applications in astronomy. Since the amount of astronomical literature making use of Bayesian methods is so vast, this introduction can by no means be exhaustive – for a detailed introduction to Bayesian statistics starting from first principles as well as a more in-depth review of Bayesian methods in astronomy, we refer to Trotta (2008). Liddle (2009) also gives an overview of statistical methods in cosmology, summarizing both Bayesian parameter estimation and model comparison; for a more recent (but brief) overview of Bayesian cosmostatistics with emphasis on recent CMB results, see Leclercq *et al.* (2014).

A central aspect of Bayesian statistics is an understanding of the concept of probability which is fundamentally different from the frequentist one: Namely, probability is considered to be a measure of the "degree of belief" about a proposition. Thus, the fundamental equation of Bayesian statistics, Bayes' Theorem, describes how the probability of a hypothesis (or model, described by a vector of model parameters $\boldsymbol{\theta}$) changes when taking new data $d$ into account:

$$p(\boldsymbol{\theta}|d) = \frac{\mathcal{L}(\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(d)}. \tag{2.17}$$

Here, the left-hand side, i.e. the conditional probability of the model parameters given the data, is usually called the posterior probability of $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta})$ denotes the prior probability of the model parameters. In cosmology, the normalization factor $p(d)$ is called Bayesian evidence for reasons we will explain later. The term we will be most concerned with in this work is the likelihood $\mathcal{L}$ of the model parameters: It is defined as the sample distribution of the data, $\mathcal{L}(\boldsymbol{\theta}) \equiv p(d|\boldsymbol{\theta})$, and is usually written as a function of $\boldsymbol{\theta}$ only, since the data $d$ is considered to be fixed (e.g. a measured value). It should be noted that Eq. (2.17) can easily be derived from the axioms of probability theory and the definition of conditional probability.

If we take the data $d$ to be a measured correlation function $\xi$, applying Bayes' Theorem in order to gain inference on cosmological parameters requires a description of the PDF of $\xi$, since the likelihood $p(\xi|\boldsymbol{\theta})$ enters the equation. Technically, $p(\xi)$ also appears in the denominator of Eq. (2.17), however, it turns out that the Bayesian evidence is not important in parameter estimation, where it only serves as a normalization constant. As we will briefly elaborate later on, it does, however, play a crucial role in Bayesian model comparison.

The first major area of applications in Bayesian statistics we would like to discuss is parameter estimation, i.e. the use of Bayes' Theorem in order to gain inference on model parameters given some data. One important step of the setup is the choice of $\boldsymbol{\theta}$: In general, $\boldsymbol{\theta}$ will be comprised of both physically interesting parameters and uninteresting ones, so-called nuisance parameters (e.g. background or noise). One major advantage of Bayesian methods is the ease of dealing with such nuisance parameters, namely by simply calculating the full posterior distribution of all parameters and then integrating over the ones that are not of interest; this process is called marginalization.

It is important to stress that the output of a Bayesian analysis is always a PDF, whereas in the context of parameter estimation, one might simply be interested in a point estimate (i.e. a "best-fit" value for a parameter including error bars). There is no single "best" choice of how to extract this information from the posterior – the most straightforward thing to do is to give the maximum of the posterior distribution, the maximum a posteriori (MAP), as a point estimate, however, depending on the shape of the posterior and the context, other values might be more suitable. In order to obtain a confidence interval (often called credible intervals in a Bayesian framework), one gives a range of parameter values so that integrating the posterior over this range gives a specific value – again, it is apparent that there are multiple ways of obtaining the interval, and careful consideration should be given to it.

From a technical point of view, sampling the posterior distribution can be quite challenging, in particular when the number of parameters is large, resulting in a high-dimensional problem. In order to explore interesting parts of parameter space, one often has to resort to numerical methods (see e.g. Allison and Dunkley 2014 for a comparison of different sampling techniques), Markov chain Monte Carlo (MCMC) methods being the most common – among other things, MCMC can also be used as a tool for integration, e.g. for marginalizing.

When using Bayes' Theorem for parameter estimation, one obvious challenge is the choice of priors, i.e. the PDF of the model parameters $\boldsymbol{\theta}$ when not taking the new data $d$ into account. In many cases, one is lucky enough to have a specific idea about the parameters and their distribution and can thus pick a shape for the prior distribution (usually called an "informative" prior). When this is not the case, however, a simple flat prior is often used. While this may seem like the most general thing to do in the absence of previous information, a flat prior does have caveats, e.g. its dependence on parametrization, i.e. the choice of $\boldsymbol{\theta}$ – this problem can be solved by using a so-called Jeffreys' prior, which is defined in terms of the Fisher information.

As mentioned before, choosing a likelihood function is, in many cases, problematic: When the

exact shape of the likelihood is unknown, a multivariate Gaussian is often assumed as default, even though this choice may be highly unphysical due to the fundamentally different physical quantities which form the parameter vector $\boldsymbol{\theta}$. In cases where the form of the likelihood function is too complicated for (repeated) computation or even unknown, it may come in handy to use likelihood-free inference, e.g. Approximate Bayesian Computation (ABC) methods: Assuming the process generating the data can be easily simulated, ABC provides a way to probe the posterior distribution without the need to directly compute the likelihood. Example applications are shown by Weyant *et al.* (2013), where ABC is used for cosmological inference from supernovae Ia, or by Cameron and Pettitt (2012) who use ABC in an analysis of morphological transformations of high-redshift galaxies.

Since in precision cosmology, one often has to deal with complex probability distributions, approximations are frequently required. Therefore, Bayesian analyses can profit from tools to validate the (numerically) computed distributions, in other words, error diagnosis methods for posterior distributions. One example of this is the Diagnostics for Insufficiencies of Posterior calculations (DIP) method (Dorn *et al.* 2013a): Basically, it amounts to sampling the model parameters $\boldsymbol{\theta}$ from their prior distribution and using the sample to generate data according to the likelihood, then computing the posterior, and finally applying a statistic constructed specifically for this purpose to check the numerical implementation and approximations made for the computation of the posterior. The DIP approach is introduced and tested in Dorn *et al.* (2013a), an example application can be found in Dorn *et al.* (2013b), where the authors develop an approximation for the posterior of the non-Gaussianity parameter $f_{\mathrm{NL}}$ (as determined from CMB data) and then use DIP to validate it.

Besides parameter estimation, model comparison is another large field of application for the Theorem. As explained in Sect. 2.1.4, the standard model of cosmology still faces numerous challenges, and thus, alternatives, extensions, and refinements are frequently proposed and tested for their compatibility with recent data using Bayesian methods – see for example Kilbinger *et al.* (2010) for a test of several extensions of the standard flat ΛCDM paradigm.

As mentioned before, the denominator on the right-hand side of Bayes' Theorem plays an important role in the context of model comparison: To understand this, it is important to keep in mind that Eq. (2.17) assumes an underlying model $\mathcal{M}$ (which is described by its parameters $\boldsymbol{\theta}$), and in fact, all probabilities in that equation contain an implicit model dependence. When spelling out this dependence, i.e. writing all probabilities as conditional on $\mathcal{M}$, the denominator becomes $p(d|\mathcal{M})$. In order to compare different models $\mathcal{M}_i$, i.e. to try and quantify the probability of a model in the presence of data, we can again use the Theorem:

$$p(\mathcal{M}|d) = \frac{p(d|\mathcal{M}) \cdot p(\mathcal{M})}{p(d)}. \tag{2.18}$$

Thus, we see that the posterior probability of a model given specific data is directly proportional to $p(d|\mathcal{M})$, which is the reason this term is called Bayesian evidence. Furthermore, the ratio of the Bayesian evidences for different models, the so-called Bayes factor, is usually used as an indicator of which model to prefer. An empirical scale originally introduced by Jeffreys (1961) can serve as an aid on how to interpret the numerical values of Bayes factors as strength of evidence. However, it has recently been pointed out by Nesseris and García-Bellido (2013) that a blind use of Jeffreys' scale may lead to biased results.

When considering the relation between Bayesian parameter estimation and model selection, it seems obvious at first glance that the two are separate steps: First, one should decide on which

model to use and then determine parameter estimates in the framework of the "best" model. In many cases, however, model selection might be inconclusive, which raises the necessity to incorporate uncertainties of the underlying model into parameter estimation, i.e. into the PDFs of the parameters under consideration. One solution to this problem is the use of a method called Bayesian model averaging (see Parkinson and Liddle 2013 for a review and example applications in astrophysics).

As a final remark to stress the importance of Bayesian methods in astronomy, it is worth mentioning that, while many publicly available codes for specific astrophysical problems make use of Bayesian statistics "under the hood", a software framework dedicated solely to Bayesian inference (independent of the particular field of research) has recently been developed, namely the Bayesian Inference Engine (BIE) (Weinberg 2013). It implements numerous methods of both parameter estimation and model comparison (including new algorithms for complex data sets and models), in particular Bayesian updating (i.e. the use of multi-component data, e.g. multiple observations, to consecutively update the posterior distribution).

CHAPTER 3

# Constraints on correlation functions

One important ingredient for our approach to obtaining an accurate description of the probability distribution of correlation functions is the fact that their values obey strict mathematical bounds. In this chapter, we will therefore give a brief overview of previous work on algebraic constraints on correlation functions, i.e. show how to compute them and stress their relevance.

As a first remark, one should note that the existence of these constraints is a very fundamental feature of correlation functions that can be derived from first principles, i.e. from pure mathematics – in fact, they originate from the non-negativity of the power spectrum $P(k)$ and thus, the results summarized in this chapter are not limited to astronomical (or even spatial) correlation functions. In this light, it is surprising that until Schneider and Hartlap (2009), there have been no studies in the literature about constrained correlation functions (except for Kurchan 2002, who derives some fundamental constraints on the autocorrelation function). While it is possible to compute the constraints analytically in certain cases (which we will show in Sect. 3.1), it turns out that for two- or higher-dimensional random fields, one has to resort to numerical methods (Sect. 3.2). We conclude with a comparison of analytically and numerically obtained constraints for the one-dimensional case in Sect. 3.3.

## 3.1 Analytical results

In the following, we will summarize the results of Schneider and Hartlap (2009), hereafter SH2009, regarding the analytical computation of constraints on correlation functions. The constraints are best expressed in terms of the correlation coefficients $r_n \equiv \xi(n\ \Delta x)/\xi(0)$, where we made use of the fact that we use a gridded approach, thus denoting $\xi(n\ \Delta x) \equiv \xi_n$, where $\Delta x = L/N$ is the separation between adjacent grid points. As it turns out, the constraints can then be written in the form

$$r_{n\mathrm{l}}(r_1, r_2, \ldots, r_{n-1}) \leq r_n \leq r_{n\mathrm{u}}(r_1, r_2, \ldots, r_{n-1}), \tag{3.1}$$

meaning that the upper and lower boundaries are functions of the $r_i$ with $i < n$.

SH2009 present two ways of calculating the constraints in the case of homogeneous, isotropic random fields. The first one applies the Cauchy-Schwartz inequality and yields the following

constraints:

$$
\begin{aligned}
-1 &\leq r_1 \leq 1, && (3.2)\\
-1+2r_1^2 &\leq r_2 \leq 1, && (3.3)\\
-1+\frac{(r_1+r_{n-1})^2}{1+r_{n-2}} &\leq r_n \leq 1-\frac{(r_1-r_{n-1})^2}{1-r_{n-2}}, \quad n>2. && (3.4)
\end{aligned}
$$

However, while this method provides easy, closed expressions for the upper and lower bounds, stricter constraints can be found – to calculate these, SH2009 use a covariance matrix approach: Since the covariance matrix $C_{ij} = \langle g_i g_j^* \rangle = \xi_{|i-j|}$ has to be positive semi-definite, its eigenvalues must be non-negative, allowing the computation of constraints on $r_i$. The bounds obtained this way differ from the results of the Cauchy-Schwartz method for $n \geq 4$, and SH2009 show that they are, in fact, optimal for a one-dimensional random field, meaning that no stricter bounds exist for a general power spectrum. As we will elaborate in Sect. 3.2, this is no longer true for higher dimensions.

The expressions for constraints are ratios of polynomials – as an example, the upper bound on $r_5$ reads

$$
\begin{aligned}
r_{5\mathrm{u}} = {} & \Big\{-r_1^3+2(-r_2+r_3+r_4+1)r_1^2+\Big(2r_2^2-2(r_3-r_4+1)r_2-r_4^2-2r_4-r_3(2r_4+1)+1\Big)\\
& \times r_1-r_3^3+r_3^2+r_4^2+r_3+2r_2r_3(r_3+r_4-1)-r_2^2(r_3+2r_4-2)-1\Big\}\\
& / \Big\{r_1^2-(2r_2+r_3-1)r_1+r_2^2+r_3-1\Big\}. \qquad (3.5)
\end{aligned}
$$

This illustrates the fact that the polynomials quickly become large and hard to handle (for example, saving the expression for $r_{16\mathrm{l}}$ in plain-text format results in a file of about 12 Megabytes). Nonetheless, it is necessary to use the bounds obtained from the covariance matrix approach and not the much simpler Cauchy-Schwartz results, as can be seen when investigating the strength of the constraints: As a measure of this strength, SH2009 compute the volume fraction of the allowed region in the space spanned by all possible values of the correlation coefficients $r_n$. It can easily be seen from Eq. (2.10) that they are bounded by $\pm 1$, i.e. $-1 \leq r_n \leq 1$, due to the non-negativity of the power spectrum $P(k)$ and thus, if we consider correlation coefficients $r_n$ with $n$ going up to $M$, the whole range of them spans a hypercube of volume $2^M$. In order to compute the volume fraction $V_M$ of the part of the hypercube which is covered by correlation coefficients obeying the constraints, we have to integrate successively, since the bounds on $r_n$ depend on all $r_i$ with $i < n$:

$$
\begin{aligned}
V_M &= \int_{r_{1\mathrm{l}}}^{r_{1\mathrm{u}}} \frac{\mathrm{d}r_1}{2} \int_{r_{2\mathrm{l}}}^{r_{2\mathrm{u}}} \frac{\mathrm{d}r_2}{2} \cdots \int_{r_{M\mathrm{l}}}^{r_{M\mathrm{u}}} \frac{\mathrm{d}r_M}{2}\\
&= \prod_{k=1}^{M} \int_{r_{k\mathrm{l}}}^{r_{k\mathrm{u}}} \frac{\mathrm{d}r_k}{2}. \qquad (3.6)
\end{aligned}
$$

The results of the integration are plotted in Fig. 3.1 as a function of the number $M$ of separations under consideration; the lower panel shows the linear dimension $(V_M)^{1/M}$ of the allowed region. In addition to the corresponding plot in SH2009, we also plot the constraints obtained via the Cauchy-Schwartz method (note that the bump at $M = 16$ is purely a result of the numerical integration and does not have any physical meaning) – it is apparent that the (computationally simpler) bounds from the Cauchy-Schwartz approach are not sufficient, since

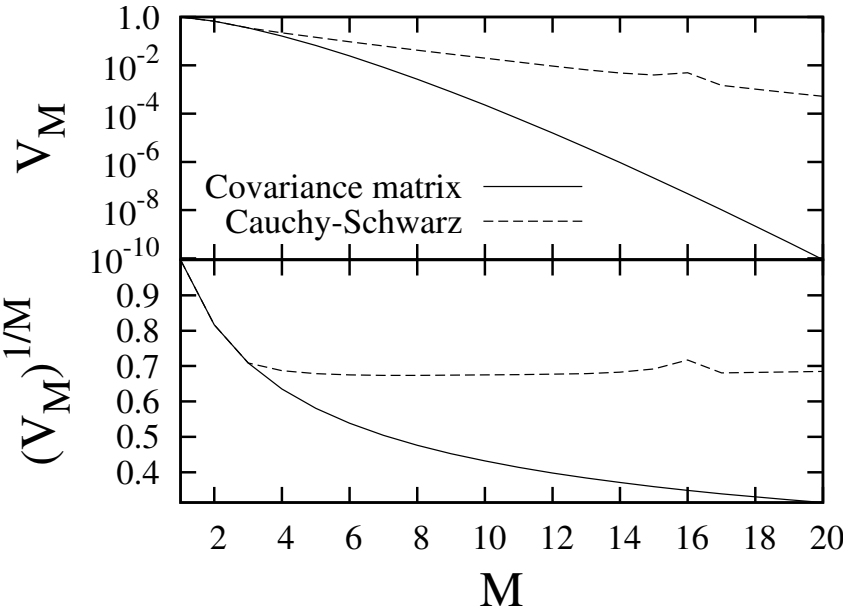


Figure 3.1: Volume fraction of the correlation coefficients obeying the constraints (upper panel) and linear dimension of the allowed region (lower panel); both are plotted as a function of the maximum number of separations.

they are by far weaker than the optimal constraints from the covariance matrix method, in particular for high $M$.

Thus, it is important to find a way of utilizing the optimal constraints $r_{n\mathrm{u}}$ and $r_{n\mathrm{l}}$ without the explicit computation of the ratios of polynomials, which are hard to handle for high $n$, as explained before. However, this can easily be done by computing the bounds on $r_n$ from $r_1, \ldots, r_{n-1}$ directly, applying the matrix methods presented in SH2009, i.e. by calculating the eigenvalues of the covariance matrix and exploiting its non-negativity for each set of values $\{r_1, \ldots, r_{n-1}\}$ separately.

While Fig. 3.1 already illustrates the strength of the constraints, one additional result from SH2009 has to be mentioned in this context: In order to examine a more realistic setting, i.e. a two-dimensional random field, they consider, as an example, a weak lensing survey. Drawing realizations of the shear correlation function from a Gaussian likelihood, they check which fraction of them lie in the allowed region by testing, for each realization, the positive semi-definiteness of the covariance matrix. Without going into details, it is clear that the result has to depend on the size of the survey and the number of bins $n$ under consideration, and SH2009 show that up to about half of the realizations can easily lie in the region explicitly forbidden by the constraints. Apart from further underlining the strength of the constraints and demonstrating the inadequacy of a Gaussian likelihood for $\xi$, this also shows that for practical applications, it is vital to obtain the constraints for higher-dimensional random fields, which is the focus of the following section.

## 3.2 Numerical methods

As mentioned before, the constraints on the correlation coefficients we described so far are only optimal for one-dimensional random fields. In higher dimensions, they are still obeyed; however,

due to the isotropy of the field and the multidimensional integration in Eq. (2.10), tighter constraints hold, which have to be computed numerically. This section mostly summarizes work by Röseler (2013), building up on ideas from SH2009.

For the case of a two-dimensional random field, the definition of the correlation function, Eq. (2.10), can be rewritten as

$$\xi(x) = \int \frac{\mathrm{d}k}{2\pi}\; kP(k)\; J_0(kx) \equiv \int \mathrm{d}k\; \hat{P}(k)\; J_0(kx), \tag{3.7}$$

where $\hat{P}(k) > 0$. Following SH2009, we apply a quadrature formula for the integral, yielding

$$\xi(x) = \sum_{j=1}^{K} w_j \hat{P}(k_j)\; J_0(k_j x) \equiv \sum_{j=1}^{K} W_j\; J_0(k_j x), \tag{3.8}$$

where the $w_j$ are positive weights given by the quadrature formula, and we defined $W_j \geq 0$. Thus, the correlation coefficient can be expressed as

$$r(x) \equiv \xi(x)/\xi(0) = \sum_{j=1}^{K} V_j\; J_0(k_j x), \tag{3.9}$$

where the coefficients

$$V_j = \left[\sum_{i=1}^{K} W_i\right]^{-1} W_j \tag{3.10}$$

fulfill $0 \leq V_j \leq 1$ and $\sum V_j = 1$. For fields of different dimensionality than two, it is possible to find similar expressions for $r(x)$ – for example, in three dimensions, the Bessel function $J_0$ is replaced by the spherical Bessel function $j_0$. Since Eq. (3.9) was obtained using a quadrature formula for the integral in Eq. (3.7), it is an approximation – however, it becomes arbitrarily accurate as $K \to \infty$.

If we now consider correlation coefficients measured for $N$ different separations $x_i$, each point $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)$ in this $N$-dimensional space can be written as a weighted sum along the curve $\boldsymbol{c}(\lambda) = (J_0(\lambda x_1), \ldots, J_0(\lambda x_N))$, where we used a continuous variable $\lambda$ with $0 \leq \lambda < \infty$ instead of discrete wave numbers $k_j$:

$$\boldsymbol{r} = \sum_{j=1}^{K} V_j\; \boldsymbol{c}(\lambda_j). \tag{3.11}$$

Since $0 \leq V_j \leq 1$ and $\sum V_j = 1$, each point $\boldsymbol{r}$ has to lie within the convex envelope of the curve $\boldsymbol{c}(\lambda)$, which corresponds to the constraints on the correlation coefficients – for example, constructing the convex envelope of the curve $\boldsymbol{c}(\lambda)$ for two points $(r_1, r_2)$ in the one-dimensional case reproduces the analytically known bounds $r_{2\mathrm{u,l}}(r_1)$.

Thus, finding the constraints reduces to describing the convex envelope of the curve $\boldsymbol{c}(\lambda)$. Unfortunately, there does not seem to be a general analytical solution for this problem, and one has to resort to numerical methods: For example, the qhull algorithm (Barber *et al.* 1996, publicly available at `http://www.qhull.org`) provides an efficient implementation for computing, among other things, convex hulls. It is, however, limited to inputs of dimensionality lower than 9, meaning that it is only applicable for a maximum number of separations of $N = 8$.

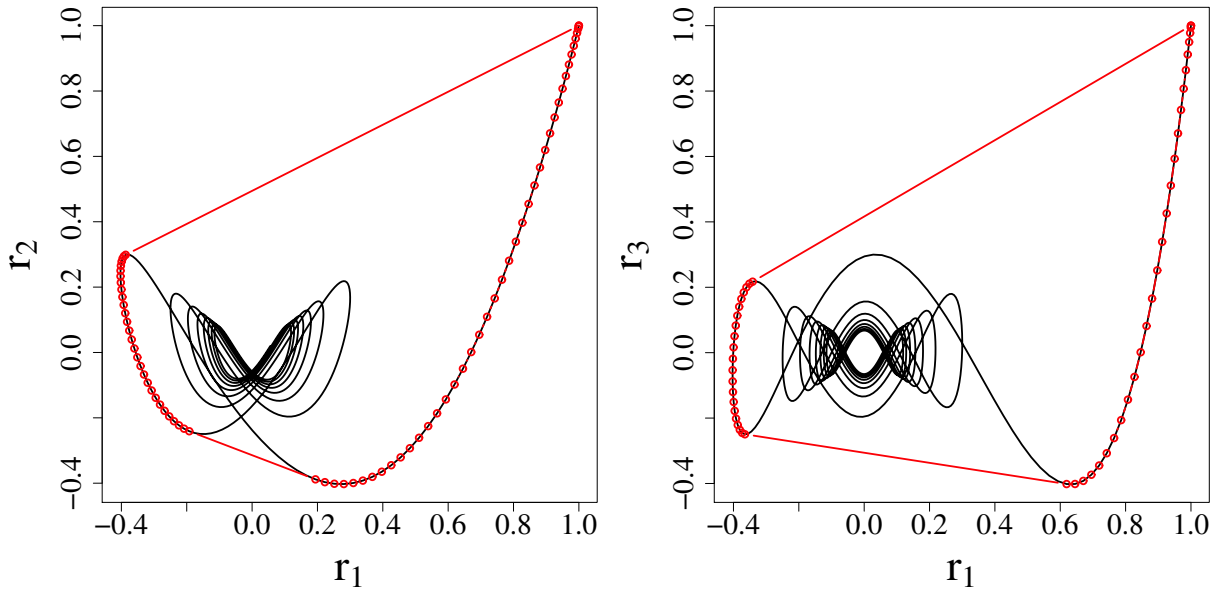


Figure 3.2: Two examples of the curve $\boldsymbol{c}(\lambda)$ for a two-dimensional random field: In the $r_1 - r_2$-plane (left panel), $\boldsymbol{c}(\lambda) = (J_0(\lambda), J_0(2\lambda))$, and in the $r_1 - r_3$-plane (right panel), $\boldsymbol{c}(\lambda) = (J_0(\lambda), J_0(3\lambda))$, respectively. The red circles and the line connecting them show the convex hulls determined by qhull.

As examples for the determination of the constraints, Fig. 3.2 shows the $r_1 - r_2$-plane (left panel) and the $r_1 - r_3$-plane (right panel); in both cases, the curve $\boldsymbol{c}(\lambda)$ is plotted in black up to $\lambda = 50$. In each panel, the red circles show points on the convex envelope of the curve as determined by qhull, the interconnecting red line is the convex hull. For a given $r_1$, the upper and lower bounds on $r_2$ are given as intersection with the red hull plotted in the left panel of Fig. 3.2 – this method can of course be generalized to higher dimensions (e.g. the determination of $r_{5\mathrm{u,l}}$ from $r_1, \ldots, r_4$, where the convex hull is a hypersurface in a five-dimensional space). Following this procedure, Röseler (2013) developed a code for computing the constraints for two- and three-dimensional fields, which we will use in our analysis of correlation functions measured from the Millennium Simulation in Sect. 6.

## 3.3 Comparison of the constraints for one-dimensional fields

In the following, we will test the numerical method of obtaining the constraints – to do so, we compare the numerical results to the analytically computed bounds. As described in Sect. 3.1, an analytical calculation of the constraints is only possible for one-dimensional Gaussian random fields, so we confine our comparison to that case.

There are several ways of approaching this: Most straightforward is to compare the constraints from the two methods directly, i.e. to compute the upper and lower bounds $r_{n\mathrm{u}}$ and $r_{n\mathrm{l}}$ both analytically and numerically and check how much they differ. An alternative approach involves the coordinate transformation $r_n \to y_n$, which is a central ingredient of the quasi-Gaussian approximation: As briefly mentioned in Sect. 1 (more details can be found in Sect. 4.3, where we give a thorough motivation and introduction to the quasi-Gaussian approach), we use the constraints in order to Gaussianize the correlation function, namely by

computing

$$y_n = \operatorname{atanh} \frac{2r_n - r_{n\mathrm{u}} - r_{n\mathrm{l}}}{r_{n\mathrm{u}} - r_{n\mathrm{l}}}. \tag{3.12}$$

Since this transformation is where the constraints are mostly used in this work, it can and should be applied as a means to compare the analytically and numerically obtained bounds, namely by using the different sets of constraints in the transformation and comparing the resulting $y_n$.

As previously described, the constraints on $r_n$ are functions of the correlation coefficients with lower lags, and thus, we need input values for $r_1, \ldots, r_{n-1}$ in order to compute and compare the different $r_{n\mathrm{l}}$ and $r_{n\mathrm{u}}$. Again, two possibilities arise: We can, on the one hand, use realizations of correlation coefficients obtained from numerical simulations (described later in Sect. 4.1) – while this approach is closest to "real-life" applications, it requires assumptions about the underlying random field and, in particular, its power spectrum. Thus, a more general approach is to draw the correlation coefficients randomly, i.e. from a uniform distribution over the allowed range, $r_n \in ]r_{n\mathrm{l}}, r_{n\mathrm{u}}[$. Due to the nature of the constraints, this is an iterative procedure, meaning that one has to draw $r_1 \in ]r_{1\mathrm{l}}, r_{1\mathrm{u}}[$, compute $r_{2\mathrm{l,u}}$ from this $r_1$, then draw $r_2 \in ]r_{2\mathrm{l}}, r_{2\mathrm{u}}[$ in order to determine $r_{3\mathrm{l,u}}$, and so on.

A direct comparison of analytically and numerically obtained $r_{n\mathrm{u}}$ and $r_{n\mathrm{l}}$ is shown in Fig. 3.3. For each bound $r_{n\mathrm{u,l}}$, the required input values of the correlation coefficients $r_i$ with $i < n$ are drawn uniformly as previously described. In order to perform a statistically significant check, this procedure is repeated 500 times, meaning that we generate 500 realizations of the input correlation coefficients and compute the upper and lower bounds both numerically and analytically for each realization. The values plotted in the figure are obtained by averaging over the 500 realizations – it can be clearly seen that the numerical and analytical bounds are in good agreement.

In order to quantify this result, we check the difference between the analytical and numerical bounds. In addition, we also investigate how much impact the sampling of the convex hull

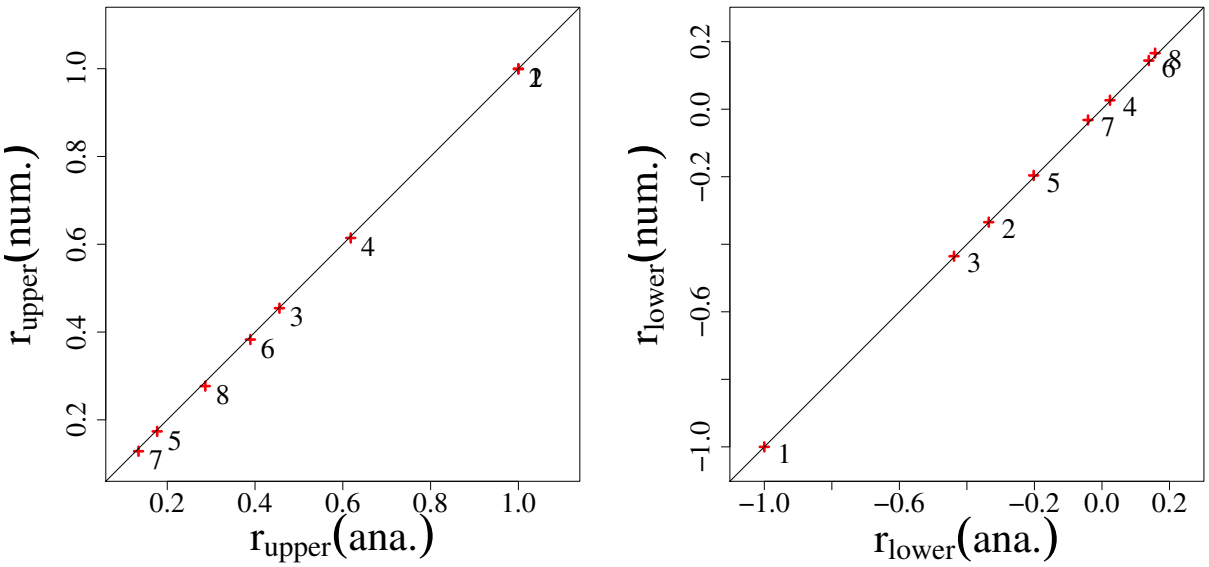


Figure 3.3: Direct comparison of the analytically and numerically obtained bounds, averaged over 500 realizations. Each cross in the left panel denotes a pair of upper bounds $(r_{n\mathrm{u}}^{\mathrm{ana}}, r_{n\mathrm{u}}^{\mathrm{ana}})$ and is labeled with the value of $n$, the right panel shows the same plot for the lower bounds (see text for details).

of the curve $\boldsymbol{c}(\lambda)$ has on the accuracy of the numerical bounds. As shown by Röseler (2013), it is sufficient to sample the curve $\boldsymbol{c}(\lambda)$ for values of $0 \leq \lambda \leq 2\pi$; for the direct comparison shown in Fig. 3.3, the sampling rate in the code developed by Röseler (2013) was set to 100 steps – in the following, we vary this number. The results can be seen in Fig. 3.4: Here, the differences $r_{\mathrm{u,l}}^{\mathrm{ana}} - r_{\mathrm{u,l}}^{\mathrm{num}}$ are plotted as functions of $n$, where the convex hull is sampled using 100 (red crosses), 200 (blue circles), and 300 (green triangles) steps. The upper three sets of points, i.e. the ones with positive slopes, show the difference $r_{\mathrm{u}}^{\mathrm{ana}} - r_{\mathrm{u}}^{\mathrm{num}}$ between the analytical and numerical upper bounds, whereas the ones with negative slopes depict the deviation between the lower bounds, i.e. $r_{\mathrm{l}}^{\mathrm{ana}} - r_{\mathrm{l}}^{\mathrm{num}}$.

Three conclusions can be drawn from Fig. 3.4: First, the deviation of the numerically obtained bounds from the analytical shows a trend to grow with $n$ – this is to be expected, since the sampling of the convex hull becomes more challenging with increasing dimensionality. Second, the impact of this sampling has a strong impact on the accuracy of the numerical calculation of the bounds; namely, the difference between the numerical and the analytical results decreases by about a factor of three when doubling the number of steps used for the convex hull sampling. In fact, this sampling is the limiting factor for the accuracy of the numerical bounds, as can be seen from the third observation: In the case of the upper bounds, the numerical results are systematically smaller than the exact analytical values, whereas for the lower bounds, the numerical values are too high. This effect is an expected consequence of the non-continuous approximation for the smooth hull – due to convexity, the splines interconnecting the points

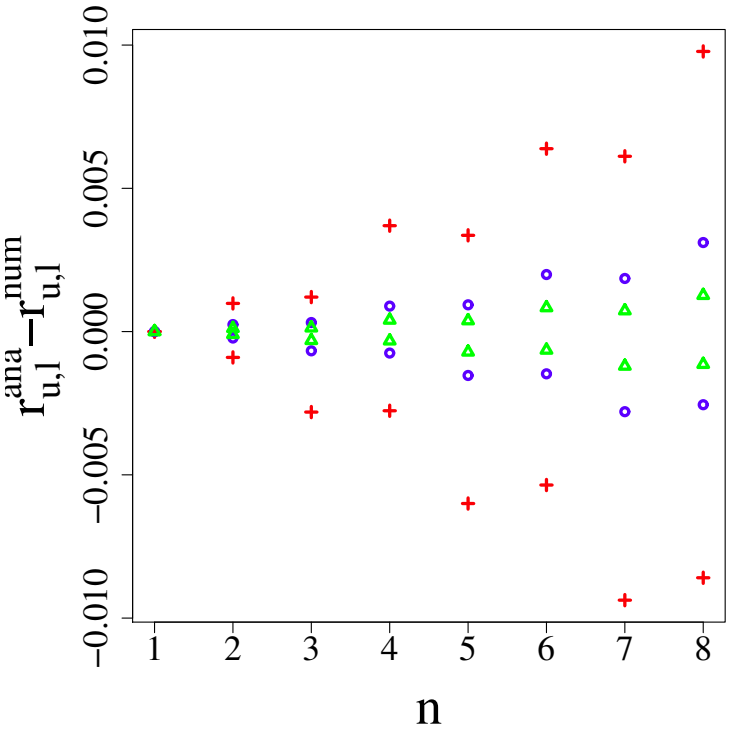


Figure 3.4: The difference between the analytically and numerically obtained bounds, averaged over 500 realizations. The upper three sets of points correspond to the difference $r_{\mathrm{u}}^{\mathrm{ana}} - r_{\mathrm{u}}^{\mathrm{num}}$, whereas the ones with negative slopes show $r_{\mathrm{l}}^{\mathrm{ana}} - r_{\mathrm{l}}^{\mathrm{num}}$. Furthermore, the different symbols denote the number of steps used to sample the convex hull of the curve $\boldsymbol{c}(\lambda)$ for values of $0 \leq \lambda \leq 2\pi$, namely 100 (red crosses), 200 (blue circles), and 300 (green triangles) steps.

used to sample the hull always have to be located inside the hull. To clarify this, the illustrative plot in Fig. 3.5 shows the exemplary contours of the allowed region in $r_i - r_j$-space: Here, the (arbitrarily chosen) black, smooth curve depicts the exact region, whereas the red lines show the non-continuous approximation. For illustrative purposes, we use a very low number of points to sample the hull in order to show its impact – it can be clearly seen that the numerical approximation overestimates the lower bound and underestimates the upper one.

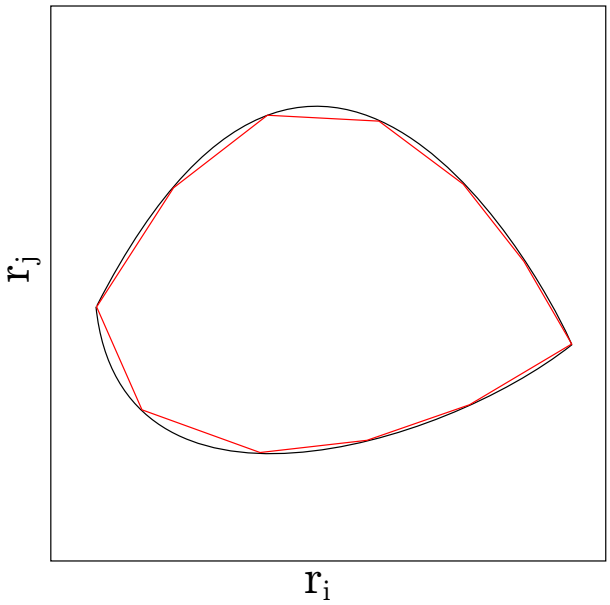


Figure 3.5: Illustration of the effect of the convex hull sampling. The black curve shows the exact shape of the allowed region in the $r_i - r_j$-plane, and the red lines are a non-continuous approximation, which overestimates the lower bound and underestimates the upper one. Note that the shape of the region is chosen arbitrarily, since it serves a purely illustrative purpose.

In summary, the accuracy of the numerical constraints can be increased by improving the sampling of the hull. While a larger number of steps could presumably improve the results even further, using more than 300 steps for the sampling becomes impractical due to the computational costs – however, as the following tests will further demonstrate, using 300 steps seems sufficiently accurate.

As mentioned before, another important check for the accuracy of the numerical methods used to compute the bounds is to apply the transformation $r_n \to y_n$ and to compare the resulting $y_n$. In the following, we adopt correlation coefficients from simulations instead of uniformly drawn ones as input for the computation of the bounds. Namely, we use 500 realizations of the correlation function on a one-dimensional Gaussian field of length $L$ with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$; for details on the simulations used in this work, we refer to Sect. 4.1.

For each simulated realization of the correlation coefficients, we compute the bounds $r_{n\mathrm{u}}$ and $r_{n\mathrm{l}}$ for each $n$ both numerically and analytically, and use them to transform $r_n$ to $y_n$ as defined in Eq. (3.12). To visualize the resulting values for $y_n$, we again average over the 500 realizations and plot the numerical and analytical values against each other – as can be seen in Fig. 3.6, they agree very well.

As before, we also plot the difference between the analytical and numerical values, and

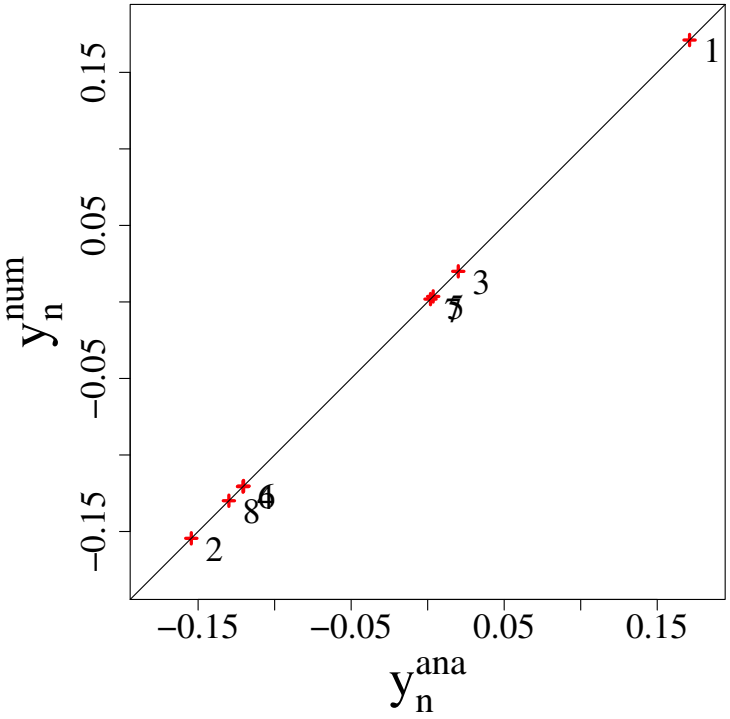


Figure 3.6: Direct comparison of the $y_n$, computed using the analytically and numerically obtained bounds, averaged over 500 realizations. The input values for the computation of the bounds used in the transformation $r_n \to y_n$ stem from 500 simulated realizations of the correlation function on a one-dimensional Gaussian field of length $L$ with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$. Again, the labels next to the crosses denote the value of $n$.

investigate the impact of the sampling of the convex hull – Fig. 3.7 shows $y_n^{\mathrm{ana}} - y_n^{\mathrm{num}}$ as a function of $n$, where 100 (left panel), 200 (central panel), and 300 (right panel) steps were used in the hull sampling. As before, the values plotted in the figure are the average over the 500 realizations; the error bars denote the standard deviations. The accuracy of the numerical

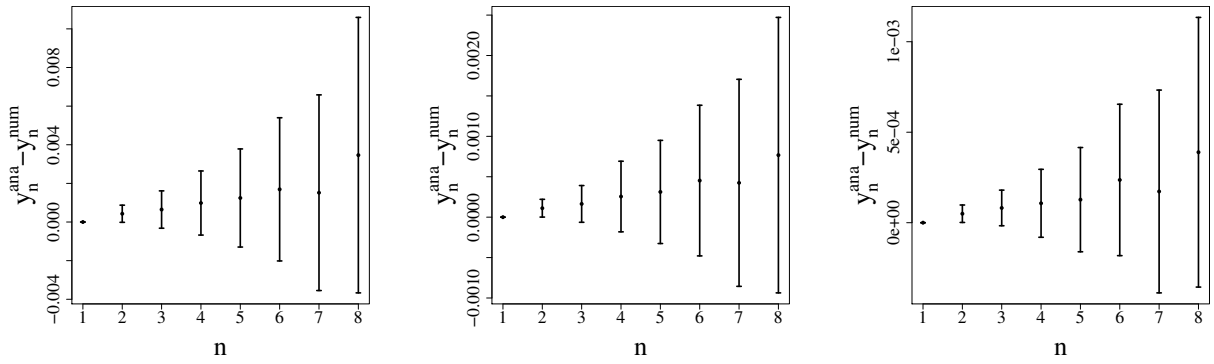


Figure 3.7: The difference between the $y_n$ computed using the analytically and the numerically obtained bounds. As described in the text and the caption of Fig. 3.6, the plotted values are again averages over 500 realizations; the error bars show the standard deviations. The three panels differ in the number of steps used to sample the convex hull of the curve $\boldsymbol{c}(\lambda)$ for $0 \leq \lambda \leq 2\pi$, namely 100 (left), 200 (center), and 300 (right).

approximation again shows a strong dependence on the number of steps used to sample the convex hull. Nonetheless, the difference between the values of $y_n$ computed using the analytical and the numerical bounds becomes very small in the rightmost panel of Fig. 3.7 – thus, we conclude that the problem of the non-continuous approximation of the curve $\boldsymbol{c}(\lambda)$ and its convex hull can be tackled, and using 300 steps in the sampling yields sufficiently accurate bounds.

CHAPTER 4

# A new approximation for the probability distribution of correlation functions

In the following, we will show how to use the constraints on correlation functions described in the previous chapter in order to obtain a new approximation for the PDF of correlation functions of one-dimensional Gaussian random fields – most results in this chapter have been published in Wilking and Schneider (2013). Before introducing the new approximation, which we call quasi-Gaussian, in Sect. 4.3, it is necessary to briefly explain the numerical simulations needed to obtain and test it as well as go into the analytical work on $p(\xi)$ by Keitel and Schneider (2011). We conclude with some tests regarding the accuracy of the quasi-Gaussian approximation and also investigate two alternative approaches.

## 4.1 Simulations

In this work, numerical simulations in order to generate realizations of the correlation functions of random fields with given power spectra play a crucial role. The purpose of these simulations is manifold: One the one hand, they can be used to verify that the constraints on correlation functions are actually obeyed, but additionally, the possibility to generate many realizations of $\xi$ allows us to probe the distribution $p(\xi)$ directly. Thus, it is possible to test various descriptions of this distribution, in particular study the Gaussianity of $\xi$, and also check the analytical results for $p(\xi)$.

Since details of the simulations used in this work can be found in Wilking and Schneider (2013), we will limit ourselves to a brief description: Instead of generating a random field and computing the correlation function from the field using an estimator, we apply a new method. Namely, after analytically computing the PDF of the power spectrum, it is possible to directly draw realizations of the power spectrum and obtain $\xi$ by Fourier transforming. Additionally, in order to further improve our simulations, we apply quasi-random (or "sub-random") sampling instead of uniform sampling, which guarantees that we probe the whole space of power spectra even in the case of small sample sizes. An introduction to quasi-randomness can be found in Press *et al.* (2007) – one way to generate quasi-random samples is by using a so-called Sobol' sequence (see Joe and Kuo 2008), we use an implementation available at the authors' website (`http://web.maths.unsw.edu.au/~fkuo/sobol/`).

## 4.2 Analytical work

In this section, we will summarize some results from Keitel and Schneider (2011) regarding the analytical calculation of $p(\xi)$ which are necessary ingredients for our quasi-Gaussian approach. A central mathematical quantity in this context is the characteristic function, usually denoted by $\psi(s)$, which is defined as the Fourier transform of the PDF of a random variable. As such, the corresponding characteristic function completely determines a probability distribution and has some useful properties: For example, the moments of a distribution can be expressed in terms of the derivative of its characteristic function (see Kendall and Stuart 1977 for details). Since we will need to compute characteristic functions in the derivation of the mean and covariance of $\xi$, we postpone mathematical details and a rigorous definition of $\psi(s)$ to Sect. 4.4.1.

From the characteristic function, Keitel and Schneider (2011) compute $p(\xi)$ via Fourier transform, i.e. by integration using the theorem of residues. However, this turns out to be challenging both numerically and analytically, and they obtain only the uni- and bivariate distributions. Thus, these analytical results are of limited use for a likelihood analysis, where one would deal with correlation functions measured at more lags than two – they do, however, play a crucial role in computing the quasi-Gaussian PDF, which requires $p(\xi_0)$ to be known, as we will show in Sect. 4.3. In the univariate case, the exact form of $p(\xi)$ as computed by Keitel and Schneider (2011) reads

$$\begin{aligned} p(\xi) &= \sum_{n=1}^{\infty} \{H(\xi)H(C_n) - H(-\xi)H(-C_n)\} \\ &\times \exp\left(-\frac{\xi}{2C_n}\right) \frac{1}{2C_n} \prod_{m\neq n}^{\infty} \frac{1}{1-\frac{C_m}{C_n}}, \end{aligned} \tag{4.1}$$

where $H(\xi)$ denotes the Heaviside step function and the $C_n$ are given by $C_n = \sigma_n^2 \cos(k_n x)$, i.e. as functions of the wave number $k_n$, the lag parameter $x$, and the $\sigma_n$ of the underlying random field. Thus, for the zero-lag correlation function $\xi_0$, $C_n = \sigma_n^2$ holds. As a final remark, it needs to be stressed that the analytical results by Keitel and Schneider (2011) have been shown to agree well with simulations.

## 4.3 The quasi-Gaussian approach

In order to obtain an accurate description of the full multivariate PDF of $\boldsymbol{\xi}$, the quasi-Gaussian approach follows up on an idea originally suggested by Schneider and Hartlap (2009), namely to transform $\boldsymbol{\xi}$ to a new variable $\boldsymbol{y}$ for which the Gaussian approximation holds, and then obtain $p(\boldsymbol{\xi})$ by transforming the Gaussian in $y$-space back to $\xi$-space. If the constraints are the main reason for the non-Gaussianity of $p(\boldsymbol{\xi})$ (which seems to be a valid assumption, since, as Schneider and Hartlap (2009) show, the shape of the distribution of the correlation coefficients is similar to the shape of the region allowed by the upper and lower bounds, even if it is located well inside the bounds), then a viable transformation to "Gaussianize" the $r_n$ is

$$r_n \to y_n = \operatorname{atanh} \frac{2r_n - r_{n\mathrm{u}} - r_{n\mathrm{l}}}{r_{n\mathrm{u}} - r_{n\mathrm{l}}}. \tag{4.2}$$

Here, the function in the argument of the inverse hyperbolic tangent uses the constraints to map the allowed range of $r_n$ to the interval $(-1, +1)$, which is then mapped to the entire real axis $(-\infty, +\infty)$. The choice of atanh by Schneider and Hartlap (2009) was an "educated guess" rather than based on theoretical arguments. It is notable that despite its simple analytical form, the transformation $r \to y$ is highly non-linear due to the nature of the constraints – as explained in Sect. 3, the upper and lower bounds on $r_n$ depend on all $r_i$ with $i < n$.

While the transformation in Eq. (4.2) yields a quantity $y$ which, by construction, is unbounded, this fact does not mean that it is Gaussian distributed. To compare the quality of the Gaussian approximation in $\xi$- and $y$-space, Fig. 4.1 shows two examples of bivariate distributions. Here, the black contours are the iso-probability contours obtained from 400 000 simulated realizations, and the red ones show the best-fitting Gaussians, i.e. normal distributions with means and covariance matrices computed from the simulated samples. It is clear

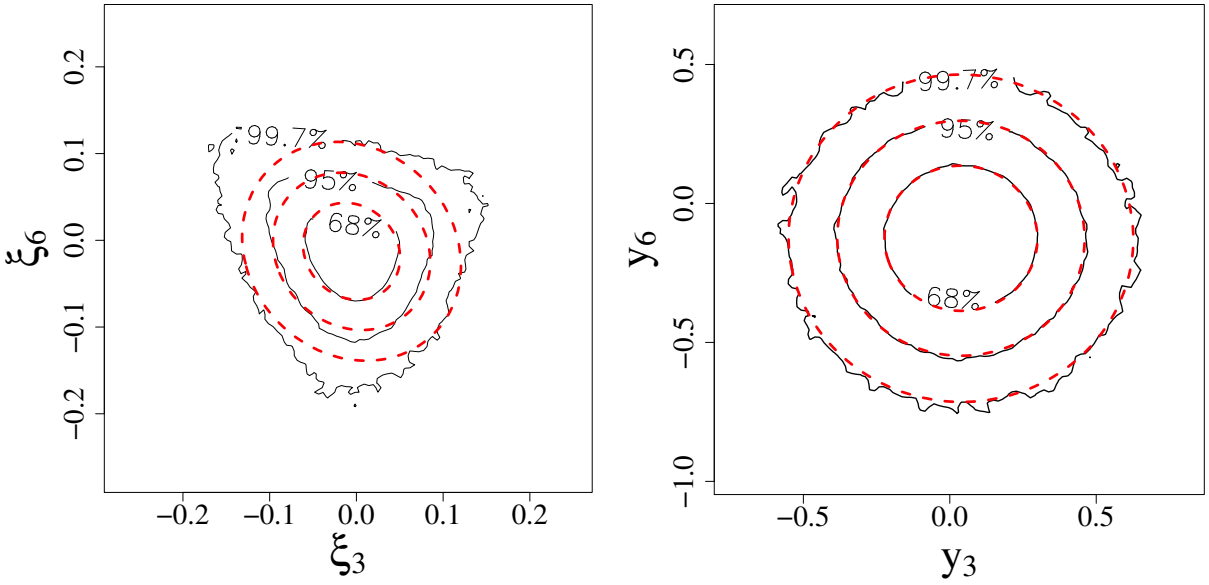


Figure 4.1: Iso-probability contours of $p(\xi_3, \xi_6)$ (left) and $p(y_3, y_6)$ (right) for 400 000 realizations of a field with $N = 32$ grid points and a Gaussian power spectrum, where the length $L$ of the field and the width $k_0$ of the power spectrum are related by $Lk_0 = 80$. The dashed red contours show the best-fitting Gaussian approximations in both cases.

"by eye" that the Gaussian approximation in $y$-space is very accurate (in particular, far more accurate than the Gaussian in $\xi$-space), but nonetheless, it is important to test this assumption in a more rigorous way. Results from these tests are presented in Sect. 4.3.1; for the following, we simply assume that the Gaussian approximation in $y$-space holds and show how to obtain the quasi-Gaussian PDF for $\boldsymbol{\xi}$.

As an intermediate step, we first need to transform the Gaussian distribution for $\boldsymbol{y}$ to $r$-space. To do so, we make use of the conservation of probability, which, in the univariate case, can be written as $p_r(r)\ \mathrm{d}r = p_y(y)\ \mathrm{d}y$. Thus, in general,

$$p_r\left(r_1, \ldots, r_n\right) = p_y\left(y_1, \ldots, y_n\right) \cdot \left|\det\left(J^{r \to y}\right)\right|, \tag{4.3}$$

where $J^{r \to y}$ denotes the Jacobian of the transformation, i.e. $J_{ij}^{r \to y} = \partial y_i / \partial r_j$. The partial derivatives required for the determinant of the Jacobian can be computed from the constraints

derived in Sect. 3.

As a final next step, the distribution $p_r$ has to be transformed to $\xi$-space. The determinant of the Jacobian for the transformation $\xi \to r$ is simple, since $r_i = \xi_i/\xi_0$; however, we have to take into account that in $\xi$-space, an additional variable, namely $\xi_0$, exists. Thus, we have to write down the transformation as follows:

$$\begin{aligned} p(\xi_0, \xi_1, \ldots, \xi_n) \prod_{i=0}^{n} \mathrm{d}\xi_i &= p'(\xi_1, \ldots, \xi_n \mid \xi_0)\; p(\xi_0) \prod_{i=1}^{n} \mathrm{d}\xi_i \\ &= p_r(r_1, \ldots, r_n \mid \xi_0)\; p(\xi_0) \prod_{i=1}^{n} \mathrm{d}r_i \\ &= p_y(y_1, \ldots, y_n \mid \xi_0)\; p(\xi_0) \prod_{i=1}^{n} \mathrm{d}y_i. \end{aligned} \tag{4.4}$$

Hence, we need the conditional probability on $\xi_0$, which raises the problem of a potential $\xi_0$-dependence of the mean $\langle \boldsymbol{y} \rangle$ and the covariance matrix $C_y$ of the Gaussian in $y$-space. As shown in Wilking and Schneider (2013), the covariance matrix is "quite independent" of $\xi_0$, and neglecting this dependence is a reasonably good approximation. The mean, however, does show a non-negligible $\xi_0$-dependence. In principle, an analytical computation of $\langle \boldsymbol{y} \rangle$ and $C_y$ should be possible, and our attempts in this respect will be discussed in detail in Sect. 4.4. However, since this computation yields no practically usable results, we determine both $\langle \boldsymbol{y} \rangle$ and $C_y$ from simulations, namely as sample mean and covariance from realizations with a $\xi_0$ close to the given value.

Equipped with all these ingredients (i.e. the Gaussian in $y$-space including its mean and covariance matrix, as well as the determinant of the Jacobians), we can finally compute the quasi-Gaussian $p(\boldsymbol{\xi})$. As an example, Fig. 4.2 shows the bivariate distribution $p(\xi_1, \xi_2)$ obtained from simulations (black contours) as well as the quasi-Gaussian approximation (red contours), which clearly agree very well.

### 4.3.1 Quality of the quasi-Gaussian approximation

While it is obvious from plots of uni- or bivariate distributions (such as Fig. 4.2) that the quasi-Gaussian approximation describes the true PDF of the simulated correlation functions very accurately, it is important to quantify these results. To do so, we performed several tests, described in detail in Wilking and Schneider (2013): One well-established method of comparing probability distributions is the so-called Kullback-Leibler (KL) divergence, which gives a measure of distance between PDFs. In the case of the univariate distribution $p(\xi_1)$, for example, the KL divergence of the true PDF obtained from simulations and the quasi-Gaussian one is smaller than the one of the true and the Gaussian PDF by more than one order of magnitude.

An additional, straightforward test is to directly quantify the level of Gaussianity of the transformed variables $y$ and compare it to the one of the original quantity $\xi$. One advantage of this approach is that multivariate tests can easily be performed, which is essential, since the fact that the univariate PDFs $p(y_i)$ are "quite Gaussian" does not imply that the full multivariate distribution $p(\boldsymbol{y})$ is well described by a multivariate Gaussian distribution.

While numerous tests for multivariate Gaussianity exist (see e.g. Henze 2002 for a review), we confine our analysis to moment-based tests. To do so, it is necessary to generalize skewness

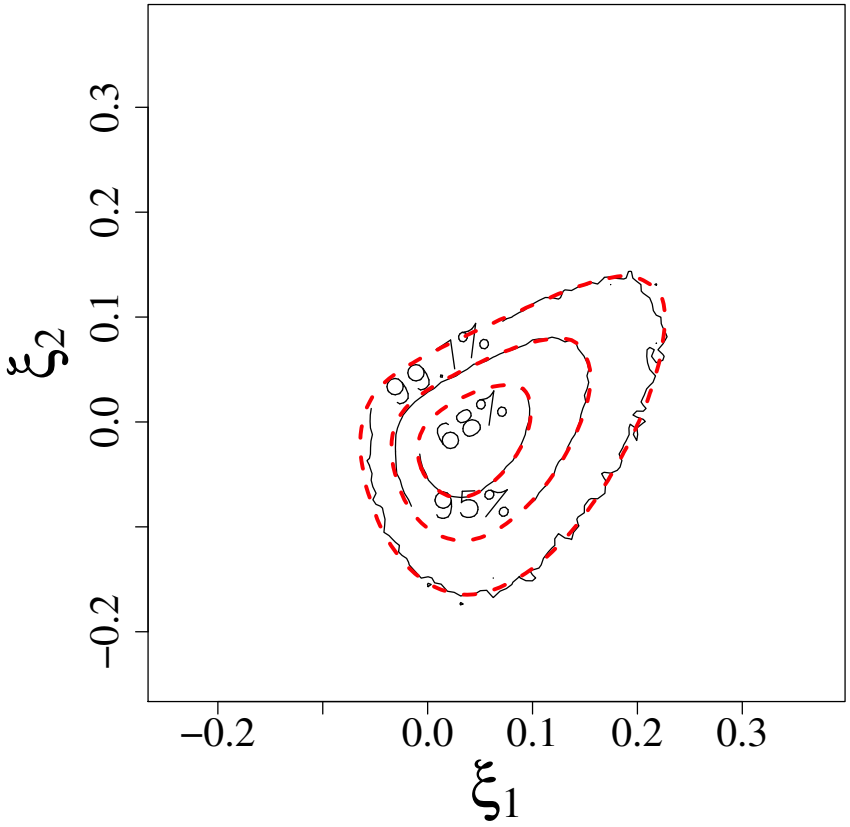


Figure 4.2: $p(\xi_1, \xi_2)$ for a field with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$. The dashed red contours show the transformed Gaussian approximation from $y$-space.

and kurtosis to higher dimensions – we use the well-established definitions by Mardia (1970, 1974): Considering a sample of $d$-dimensional vectors $x_i$, a measure of skewness for a $d$-variate distribution can be written as

$$\gamma_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}} C^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}) \right\}^3 , \tag{4.5}$$

where $n$ denotes the sample size, and $\boldsymbol{\mu}$ and $C$ are the mean and covariance matrix of the sample. The corresponding kurtosis measure is

$$\kappa_d = \frac{1}{n} \sum_{i=1}^{n} \left\{ (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}} C^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\}^2 - d(d+2), \tag{4.6}$$

where we subtracted the last term to make sure that the kurtosis of a Gaussian sample is zero.

Fig. 4.3 shows the results of the multivariate test. For a simulation run of a field with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$ (using only 5000 realizations to speed up calculations), the skewness and kurtosis of the $n$-variate distributions, i.e. $p(\xi_0, \ldots, \xi_{n-1})$ (black circles) and $p(y_1, \ldots, y_n)$ (red, upward triangles) are plotted as a function of $n$. For comparison, the blue squares show the moments of the $n$-variate distributions marginalized from a 15-variate Gaussian. It is apparent that also in the multivariate case, the assumption of

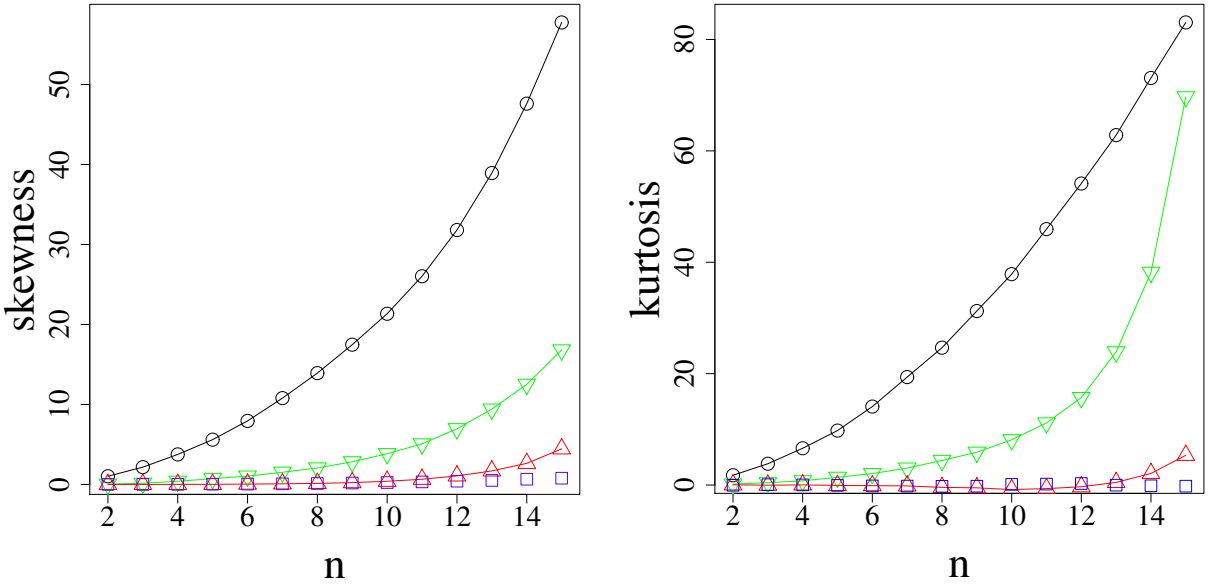


Figure 4.3: Mardia's skewness (left-hand panel) and kurtosis (right-hand panel) of $n$-variate $\{\xi\}$- (black circles) and $\{y\}$-samples (red, upward triangles) for a Gaussian power spectrum with $Lk_0 = 80$. The green, downward triangles show the skewness / kurtosis of $\{y\}$-samples obtained under the assumption of a Gaussian $\{\xi\}$-sample, and the blue squares show the skewness / kurtosis of Gaussian samples. See text for more details.

Gaussianity is by far better justified for the transformed variables $y$ than for $\xi$, although the approximation is not perfect.

To avert comparing PDFs of different random variables (namely $\xi$ and $y$), we perform an additional check: We draw a 15-variate Gaussian sample in $\xi$-space and transform it to $y$-space (using only the realizations which lie inside the constraints); the corresponding values of skewness and kurtosis are shown as green, downward triangles. Clearly, they are by far higher than those obtained for the "actual" $y$-samples, further justifying our approach.

## 4.4 Analytical calculation of mean and covariance matrix

As mentioned in Sect. 4.3, the analytical calculation of the $\xi_0$-dependence of the mean and the covariance matrix does not produce practically usable results. Nonetheless, these results are interesting from a theoretical point of view and we will thus present them in the following. The computation of the quasi-Gaussian approximation requires the mean $\boldsymbol{y}$ and the covariance matrix $C_y$. However, we will first show calculations in $\xi$-space before addressing the problem of how to transform the results to $y$-space.

### 4.4.1 Calculations in $\xi$-space

In order to compute the $\xi_0$-dependence of the mean $\langle\xi_1\rangle$ (where the index is purely a numbering and does not denote the lag), we express the mean as the first moment of the probability distribution:

$$\langle\xi_1\rangle\,(\xi_0) = \int \mathrm{d}\xi_1\ \xi_1\ p(\xi_1|\xi_0) \tag{4.7}$$

with the conditional probability

$$p(\xi_1|\xi_0) = \frac{p(\xi_0, \xi_1)}{p(\xi_0)}. \tag{4.8}$$

Following Keitel and Schneider 2011, hereafter KS2011, we can compute the left-hand side of Eq. (4.7) by making use of the fact that the moments of a probability distribution are related to the derivatives of its characteristic function, which is defined as the Fourier transform of the PDF ($\Phi \leftrightarrow p$ in short-hand notation, see Kendall and Stuart 1977 for details on characteristic functions). Thus, we need to compute the characteristic function of $p(\xi_1|\xi_0)$, which is defined by

$$\Phi(s_1;\xi_0) = \int \mathrm{d}\xi_1 \ \mathrm{e}^{\mathrm{i}s_1\xi_1} \ p(\xi_1|\xi_0), \tag{4.9}$$

$$p(\xi_1|\xi_0) = \int \frac{\mathrm{d}s_1}{2\pi} \ \mathrm{e}^{-\mathrm{i}s_1\xi_1} \ \Phi(s_1;\xi_0). \tag{4.10}$$

Making use of the characteristic function $\Psi(s_0, s_1)$ (where $\Psi(s_0, s_1) \leftrightarrow p(\xi_0, \xi_1)$) computed in KS2011, we can also write

$$\begin{aligned} p(\xi_1|\xi_0) &= \frac{1}{p(\xi_0)} \int \frac{\mathrm{d}s_0}{2\pi} \int \frac{\mathrm{d}s_1}{2\pi} \ \mathrm{e}^{-\mathrm{i}s_0\xi_0} \ \mathrm{e}^{-\mathrm{i}s_1\xi_1} \ \Psi(s_0, s_1) \\ &= \int \frac{\mathrm{d}s_1}{2\pi} \ \mathrm{e}^{-\mathrm{i}s_1\xi_1} \ \frac{1}{p(\xi_0)} \int \frac{\mathrm{d}s_0}{2\pi} \mathrm{e}^{-\mathrm{i}s_0\xi_0} \ \Psi(s_0, s_1). \end{aligned} \tag{4.11}$$

Comparison with Eq. (4.10) yields

$$\Phi(s_1;\xi_0) = \frac{1}{p(\xi_0)} \int \frac{\mathrm{d}s_0}{2\pi} \ \mathrm{e}^{-\mathrm{i}s_0\xi_0} \ \Psi(s_0, s_1). \tag{4.12}$$

As stated before, the $k$th moment $M_k$ of a PDF is related to the $k$th derivative of the corresponding characteristic function $\psi(s)$, namely by

$$M_k = \mathrm{i}^{-k} \left. \frac{\mathrm{d}^k \psi(s)}{\mathrm{d}s^k} \right|_{s=0}. \tag{4.13}$$

Thus, we can calculate the mean, i.e. the left-hand side of Eq. (4.7), from the characteristic function in Eq. (4.12):

$$\langle \xi_1 \rangle(\xi_0) = \frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_1} \Phi(s_1;\xi_0) \Big|_{s_1=0} \tag{4.14}$$

$$= \frac{1}{p(\xi_0)} \int \frac{\mathrm{d}s_0}{2\pi} \ \mathrm{e}^{-\mathrm{i}s_0\xi_0} \ \frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_1} \Psi(s_0, s_1) \Big|_{s_1=0}. \tag{4.15}$$

Using the result from KS2011 for the bivariate characteristic function,

$$\Psi(s_0, s_1) = \prod_{n=1}^{\infty} \frac{1}{1 - 2\mathrm{i}s_0 C_{n0} - 2\mathrm{i}s_1 C_{n1}}, \tag{4.16}$$

where $C_{nm} = \sigma_n^2 \cos(k_n x_m)$, we can calculate the derivative as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_1}\Psi(s_0,s_1)\Big|_{s_1=0} &= \sum_{n=1}^{\infty}\frac{2C_{n1}}{(1-2\mathrm{i}s_0C_{n0})^2}\prod_{k\neq n}\frac{1}{1-2\mathrm{i}s_0C_{k0}} \\ &= \sum_{n=1}^{\infty}\frac{2C_{n1}}{1-2\mathrm{i}s_0C_{n0}}\prod_{k=1}^{\infty}\frac{1}{1-2\mathrm{i}s_0C_{k0}} \\ &= \Psi(s_0)\sum_{n=1}^{\infty}\underbrace{\frac{2C_{n1}}{1-2\mathrm{i}s_0C_{n0}}}_{Y_n(s_0)}, \end{aligned} \tag{4.17}$$

where we inserted the univariate characteristic function computed in KS2011,

$$\Psi(s_0) = \prod_{n=1}^{\infty}\frac{1}{1-2\mathrm{i}s_0C_{n0}}. \tag{4.18}$$

To calculate the integral in Eq. (4.15), we use a Taylor expansion of $Y_n(s_0)$ from Eq. (4.17):

$$Y_n(s_0) \approx \sum_{k=0}^{\infty} 2^{k+1}\ (\mathrm{i}s_0)^k\ C_{n0}^k C_{n1}. \tag{4.19}$$

We insert the derivative into Eq. (4.15) and thus obtain

$$\begin{aligned}\langle\xi_1\rangle(\xi_0) \approx \frac{1}{p(\xi_0)}\int\frac{\mathrm{d}s_0}{2\pi}\ \mathrm{e}^{-\mathrm{i}s_0\xi_0}\ \Psi(s_0) \\ \sum_{n=1}^{\infty}\sum_{k=0}^{\infty} 2^{k+1}\ (\mathrm{i}s_0)^k\ C_{n0}^k C_{n1}. \end{aligned} \tag{4.20}$$

According to the definition of $\Psi(s_0) \leftrightarrow p(\xi_0)$,

$$\frac{\mathrm{d}^k p(\xi_0)}{\mathrm{d}\xi_0^k} = \int\frac{\mathrm{d}s_0}{2\pi}(-\mathrm{i}s_0)^k\ \mathrm{e}^{-\mathrm{i}s_0\xi_0}\ \Psi(s_0), \tag{4.21}$$

and thus, after changing the order of summation and integration, Eq. (4.20) can finally be written as

$$\langle\xi_1\rangle(\xi_0) = \sum_{k=0}^{\text{order}}\sum_{n=1}^{\text{modes}} 2^{k+1}\ C_{n0}^k C_{n1}\ (-1)^k\ \frac{\mathrm{d}^k p(\xi_0)}{\mathrm{d}\xi_0^k}\ \frac{1}{p(\xi_0)}. \tag{4.22}$$

Inserting the known result for $p(\xi_0)$ and calculating its derivatives allows us to compare the analytical result to simulations. The results can be seen in Fig. 4.4; here, the black points with error bars show the mean of $\xi_n$ for different lags $n$ as determined from simulations (100 000 realizations, Gaussian power spectrum with $Lk_0 = 100$), and the colored symbols show the analytical results to different order (see figure caption for details). It seems that, although the Taylor series in Eq. (4.22) does not converge, a truncation at order 10 yields sufficient accuracy, barring some numerical issues for very low $\xi_0$-values. We will briefly comment on these after showing results of the analytical calculation of the covariance matrix.

The $\xi_0$-dependence of the covariance matrix $C_\xi$ can be computed in a similar way to the

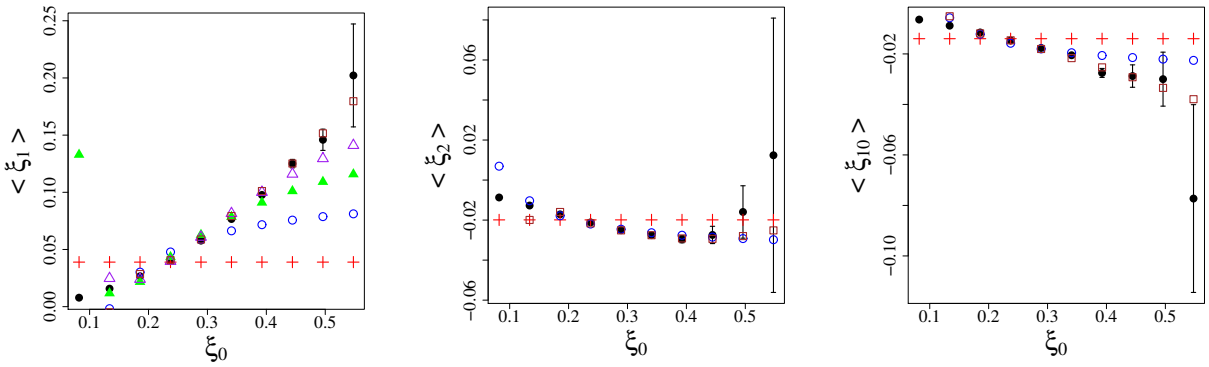


Figure 4.4: The mean of $\xi_n$ for different $n$ as function of $\xi_0$, determined from simulations (black points with error bars) and analytically to zeroth (red crosses), first (blue circles), second (green filled triangles; left panel only), third (purple empty triangles; left panel only), and tenth (brown squares) order.

mean. We start from the general definition of covariance,

$$\begin{aligned}
\mathrm{cov}\left(\xi_1,\xi_2\right)\left(\xi_0\right) &\equiv \left\langle \left(\xi_1 - \langle\xi_1\rangle\right)\ \left(\xi_2 - \langle\xi_2\rangle\right)\right\rangle_{\xi_0} \\
&= \int \mathrm{d}\xi_1\ \mathrm{d}\xi_2\ \left(\xi_1 - \langle\xi_1\rangle\right)\ \left(\xi_2 - \langle\xi_2\rangle\right)\ p(\xi_1,\xi_2|\xi_0) \\
&= \underbrace{\int \mathrm{d}\xi_1\ \mathrm{d}\xi_2\ \xi_1\xi_2\ p(\xi_1,\xi_2|\xi_0)}_{\equiv A} \\
&+ \langle\xi_1\rangle\left(\xi_0\right)\ \langle\xi_2\rangle\left(\xi_0\right) \underbrace{\int \mathrm{d}\xi_1\ \mathrm{d}\xi_2\ p(\xi_1,\xi_2|\xi_0)}_{=1} \\
&- \langle\xi_1\rangle\left(\xi_0\right) \underbrace{\int \mathrm{d}\xi_1\ \mathrm{d}\xi_2\ \xi_2\ p(\xi_1,\xi_2|\xi_0)}_{=\langle\xi_2\rangle(\xi_0)} \\
&- \langle\xi_2\rangle\left(\xi_0\right) \underbrace{\int \mathrm{d}\xi_1\ \mathrm{d}\xi_2\ \xi_1\ p(\xi_1,\xi_2|\xi_0)}_{=\langle\xi_1\rangle(\xi_0)}.
\end{aligned}$$

The integral $A$ can again be expressed in terms of the characteristic function $\Phi(s_1,s_2;\xi_0) \leftrightarrow p(\xi_1,\xi_2|\xi_0)$:

$$A = \frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_1}\frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_2}\ \Phi(s_1,s_2;\xi_0)\Big|_{s_1=s_2=0} \tag{4.23}$$

Similar to the previous calculations,

$$\Phi\left(s_1,s_2;\xi_0\right) = \frac{1}{p\left(\xi_0\right)}\int\frac{\mathrm{d}s_0}{2\pi}\ \mathrm{e}^{-\mathrm{i}s_0\xi_0}\ \Psi\left(s_0,s_1,s_2\right) \tag{4.24}$$

with the trivariate characteristic function

$$\Psi\left(s_0,s_1,s_2\right) = \prod_{n=1}^{\infty}\frac{1}{1-2\mathrm{i}s_0C_{n0}-2\mathrm{i}s_1C_{n1}-2\mathrm{i}s_2C_{n2}}. \tag{4.25}$$

Calculating the second derivative of (4.25) yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_1}\frac{\mathrm{d}}{\mathrm{i}\mathrm{d}s_2} \quad \Psi \quad (s_0,s_1,s_2)\Big|_{s_1=s_2=0} \\ = \quad \Psi\left(s_0\right)\left[\sum_{n=1}^{\infty}\underbrace{\frac{4C_{n1}C_{n2}}{(1-2\mathrm{i}s_0C_{n0})^2}}_{Z_n(s_0)}\right. \\ \left. + \quad \sum_{n,k=1}^{\infty}\underbrace{\frac{4C_{n1}C_{k2}}{(1-2\mathrm{i}s_0C_{n0})(1-2\mathrm{i}s_0C_{k0})}}_{Z_{k,n}(s_0)}\right].\end{aligned} \tag{4.26}$$

The Taylor expansions of $Z_n(s_0)$ and $Z_{k,n}(s_0)$ read

$$Z_n(s_0) \approx \sum_{k=0}^{\infty} 2^{k+2}\ (k+1)\,(\mathrm{i}s_0)^k\ C_{n0}^k C_{n1}C_{n2}, \tag{4.27}$$

$$Z_{k,n}(s_0) \approx \sum_{l=0}^{\infty} 2^{l+2}\ (\mathrm{i}s_0)^l\ C_{k2}C_{n1}\sum_{p=0}^{l} C_{k0}^p\ C_{n0}^{l-p}. \tag{4.28}$$

Using it as well as the expansion (4.19), we finally obtain

$$\begin{aligned}\mathrm{cov}\left(\xi_1,\xi_2\right)(\xi_0) \quad = \quad \sum_{k=0}^{\mathrm{order}}\ \frac{1}{p(\xi_0)}\frac{\mathrm{d}^k p\left(\xi_0\right)}{\mathrm{d}\xi_0^k}\left[\sum_{n}^{\mathrm{modes}}(-1)^k\ 2^{k+2}\ (k+1)\right. \\ \times \quad C_{n0}^k C_{n1}C_{n2} + \sum_{n,m}^{\mathrm{modes}}(-1)^k\ 2^{k+2}\ C_{m2}\ C_{n1} \\ \left.\times \quad \sum_{p=0}^{k} C_{m0}^p\ C_{n0}^{k-p}\right] - \langle\xi_1\rangle(\xi_0)\ \langle\xi_2\rangle(\xi_0).\end{aligned} \tag{4.29}$$

We show a comparison of the results (for different elements of the covariance matrix) from simulations and the analytical formula in Fig. 4.5. Again, the black dots are obtained from simulations and the colored symbols represent the results from Eq. (4.29), where the last term (i.e. the one containing the mean values $\langle\xi_n\rangle$) was calculated up to tenth order, thus providing sufficient accuracy, as previously shown. As before, there are some numerical problems for very small values of $\xi_0$. Additionally, the analytical results do not agree with the simulations for small lags, as can be seen from the left-most panel (the same holds for other covariance matrix elements involving small lags). However, for the higher-lag examples (i.e. the two right-hand panels), a truncation of the Taylor series at tenth order seems to be accurate enough.

The reasons for both the problems occurring for low-lag covariance matrix elements and the overall issues for small $\xi_0$-values are unclear – while the latter are not very severe and probably stem from insufficient numerical accuracy, an exact computation of all covariance matrix elements is crucial. Since we tested the various Taylor expansions used in the computation separately and they seem accurate enough, it remains unclear which part of the approxima-

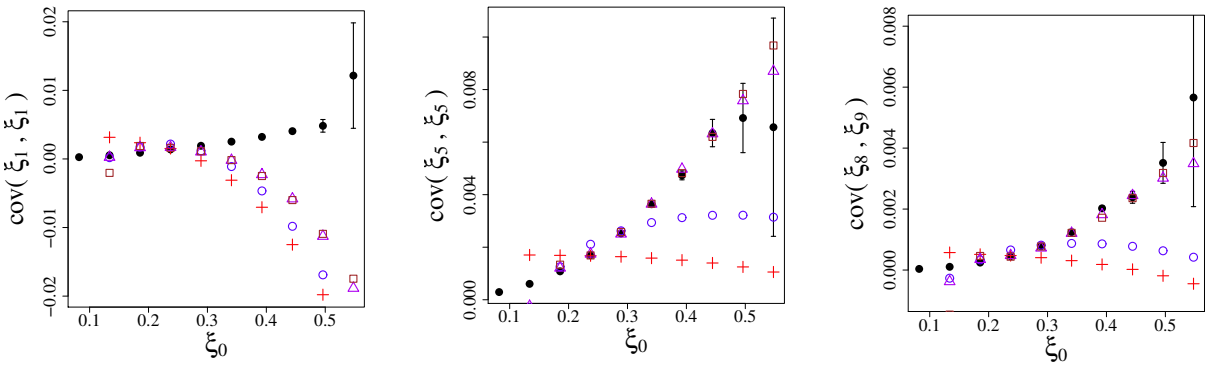


Figure 4.5: Different elements of the covariance matrix $C(\{\xi_n\})$, determined from simulations (black points with error bars) and analytically to zeroth (red crosses), first (blue circles), fifth (purple triangles), and tenth (brown squares) order.

tion breaks down. As we shall see in the following, however, the transformation of the results presented in this section to $y$-space poses an even greater challenge.

### 4.4.2 Transformation of mean and covariance matrix to $y$-space

In the previous section, we showed how to calculate the ($\xi_0$-dependent) mean and covariance matrix in $\xi$-space and addressed some occurring problems. The computation of the quasi-Gaussian PDF, however, requires the mean $\boldsymbol{y}$ and the covariance matrix $C_y$ in $y$-space, which cannot be obtained from those in $\xi$-space analytically due to the complex nature of the transformation $\xi \to y$.

As a first approach, we test a linear approximation for the transformation. Under this assumption, the mean in $y$-space can simply be obtained by transforming the mean in $\xi$, i.e. $\langle \boldsymbol{y} \rangle = \boldsymbol{y}(\langle \boldsymbol{\xi} \rangle)$. Since covariance is itself a linear function, the covariance matrix in $y$-space then reads

$$\mathrm{cov}(y_n, y_m) \approx \sum_{i=0}^{n} \sum_{j=0}^{m} \frac{\partial y_n}{\partial \xi_i} \frac{\partial y_m}{\partial \xi_j} \, \mathrm{cov}(\xi_i, \xi_j). \tag{4.30}$$

In order to test the linear approximation, we calculate the approximated mean and covariance matrix in $y$-space from $\langle \boldsymbol{\xi} \rangle$ and $C_\xi$ (obtained from a simulated sample of correlation functions) and compare them to the true mean and covariance matrix, i.e. the ones obtained by individually transforming each realization $\boldsymbol{\xi}$ to $y$-space and computing a sample mean and covariance matrix. While the results for the mean and the variance are still quite acceptable, the linear approximation completely breaks down for the off-diagonal elements of the covariance matrix, which is not surprising due to the fact that the transformation $\xi \to y$ is highly non-linear.

Thus, instead of settling for a linear approximation, we have to choose a more computationally expensive approach. Namely, we calculate the first and second moments (in $\xi$) of the quasi-Gaussian distribution as functions of the mean and (inverse) covariance matrix in $y$-space and equate the result to the analytical results, i.e. we solve a set of equations of the form

$$\int \mathrm{d}\xi \; \xi_i \; p\left(\boldsymbol{\xi}; \langle \boldsymbol{y} \rangle, C_y^{-1}\right) = \langle \xi_i \rangle_{\mathrm{ana}} \tag{4.31}$$

$$\int \mathrm{d}\xi \; \xi_j \; \xi_k \; p\left(\boldsymbol{\xi}; \langle \boldsymbol{y} \rangle, C_y^{-1}\right) = \langle \xi_j \; \xi_k \rangle_{\mathrm{ana}}, \tag{4.32}$$

where we did not write down the $\xi_0$-dependence explicitly for the sake of readability.

It is worth noting that this is a complicated procedure which requires powerful numerical tools: On the one hand, the set of equations can only be solved numerically; in the following, we use a multi-dimensional root-finding algorithm provided within the GNU Scientific Library (GSL) (Galassi *et al.* 2009). Additionally, the integration on the equations' left-hand sides can only be performed numerically – to do so, we make use of a Monte-Carlo code from Press *et al.* (2007).

Due to the complexity of this approach, we first test it in a simpler setup, thus avoiding some subtleties and numerical challenges in the calculation of the quasi-Gaussian PDF. Instead, we assume an $N$-variate Gaussian distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, C)$ with known mean and covariance matrix and try to recover those with the method described above. In this case, the set of equations we need to solve looks as follows:

$$\int \mathrm{d}^N x \; x_i \; \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, C) \quad = \quad a_i \tag{4.33}$$

$$\int \mathrm{d}^N x \; x_i x_j \; \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, C) \quad = \quad b_{ij}, \tag{4.34}$$

where the right-hand sides denote the known, fiducial values for $\langle x_i \rangle$, $\langle x_i x_j \rangle$. Thus, for an $N$-variate distribution, the system consists of $N + \frac{1}{2}N(N+1)$ equations – for testing purposes, we use $N = 8$, resulting in a set of 44 equations (8 for mean, and 36 for the covariance matrix elements).

As fiducial values, we choose a mean of zero and a covariance matrix equal to diag(0.5), thus setting the $a_i = \langle x_i \rangle_{\mathrm{fid}} \equiv 0$ and the $b_{ij} = \langle x_i x_j \rangle_{\mathrm{fid}}$ (recall that $C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$). Furthermore, the numerical solver requires starting values, which we set randomly and "not too far" from the fiducial values. Using 500 000 Monte-Carlo steps for integration, the solver converges after $\sim$ 10 iterations. The results for the elements of the mean are quite close to the fiducial values of zero, they differ by about 0.05 at most. Fig. 4.6 visualizes the results for the covariance matrix: The fiducial values (left panel), the output values recovered by the numerical solver (central panel), as well as their difference (right panel) are plotted. It can be seen that the numerical procedure does indeed recover the fiducial mean and covariance matrix of the Gaussian distribution quite accurately, providing a proof of concept for the methods used – it remains to be tested whether the precision is high enough in the case of the quasi-Gaussian

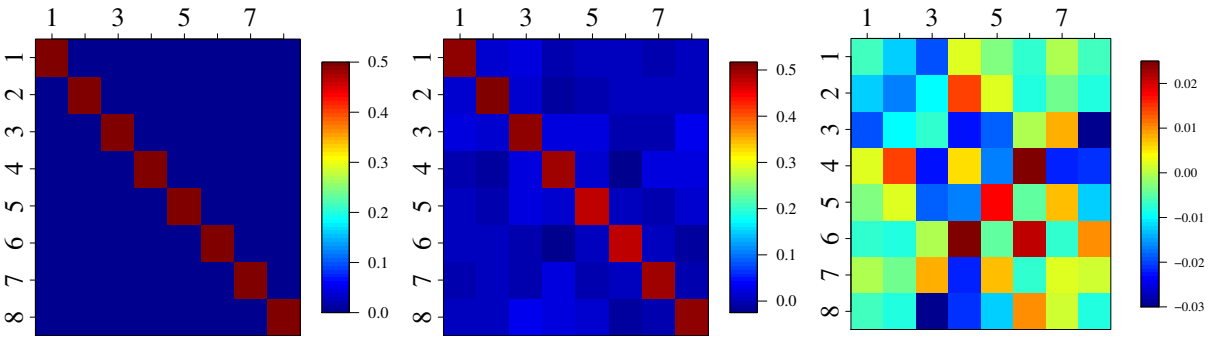


Figure 4.6: Results of the test setup used to recover the covariance matrix of an 8-variate Gaussian distribution. The first panel shows the fiducial covariance matrix, whereas the values recovered by the numerical methods described in the text are visualized in the second panel. The third panel depicts the difference between the fiducial and the recovered covariance matrices.

distribution.

However, when implementing the quasi-Gaussian PDF instead of a Gaussian, we are unable to obtain correct results for the mean and covariance matrix. One reason for this might be the more complicated shape of the probability contours, which poses a challenge for the numerical integration – to circumvent this problem, we perform the integration in $y$-space (where the Gaussian approximation holds), and thus rewrite, for example, the left-hand side of Eq. (4.31) as

$$\int \mathrm{d}^N\xi \; \xi_i \; p(\boldsymbol{\xi}) = \int \mathrm{d}^{N-1}y \; \mathrm{d}\xi_0 \; \xi_i(\boldsymbol{y}) \; p(\xi_0) \; p(\boldsymbol{y}|\xi_0). \tag{4.35}$$

An additional complication are the computational costs, so we have to limit ourselves to the case $N = 4$ (i.e. $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_3)^{\mathrm{T}}$), resulting in 9 equations (namely for $\langle\xi_1\rangle$, $\langle\xi_2\rangle$, $\langle\xi_3\rangle$; $\langle\xi_1^2\rangle$, $\langle\xi_2^2\rangle$, $\langle\xi_3^2\rangle$; $\langle\xi_1\ \xi_2\rangle$, $\langle\xi_1\ \xi_3\rangle$, $\langle\xi_2\ \xi_3\rangle$) and the same number of unknowns ($\langle y_1\rangle$, $\langle y_2\rangle$, $\langle y_3\rangle$,...). Nonetheless, the method does not give usable results, which may be due to a faulty implementation – in any case, it seems likely that, even after solving those problems, we would have to accept drawbacks in accuracy in order to keep computational costs reasonably low. Thus, it is expected that any possible gain in accuracy (which was one motivation for the analytical calculation of the mean and covariance matrix in the first place) would be averted by the required heavy use of purely numerical methods.

In summary, this procedure does not seem practical, leading us to abandon any further attempts in this direction. Thus, as described in Sect. 4.3, we refrain from using our analytical results for the mean and covariance matrix and simply determine them (as well as their $\xi_0$-dependence) from simulations, which we have shown to be sufficiently accurate.

## 4.5 Alternative approaches

In this section, we briefly investigate two potential alternative ways of finding a new approximation for the likelihood of correlation functions, namely a copula approach and a method involving Box-Cox transformations. Both methods have previously been applied in astronomy, we will show, however, that they do not yield satisfactory results in the case of correlation functions.

### 4.5.1 A copula approach

As the name suggests, a copula function provides a way of coupling univariate probability distributions in order to obtain a multivariate PDF. In recent years, copula methods have gained notoriety due to the use of Gaussian copulae at Wall Street in order to estimate the PDF of losses, and have even been connected to the recent financial crisis (see for example MacKenzie and Spears 2014).

Since the exact univariate PDF of $\xi$ is known (Keitel and Schneider 2011), using a copula approach to compute the correlation function likelihood seems to be an obvious step. According to the definition of the copula, the joint PDF of $n$ random variables $\xi_i$ can be calculated from the univariate distributions $p_i(\xi_i)$ as

$$p(\xi_1, \xi_2, \ldots, \xi_n) = c(u_1, u_2, \ldots, u_n) \cdot \prod_{i=1}^{n} p_i(\xi_i), \tag{4.36}$$

where the copula density function $c$ depends on $u_i = P_i(\xi_i)$, i.e. on the cumulative distribution

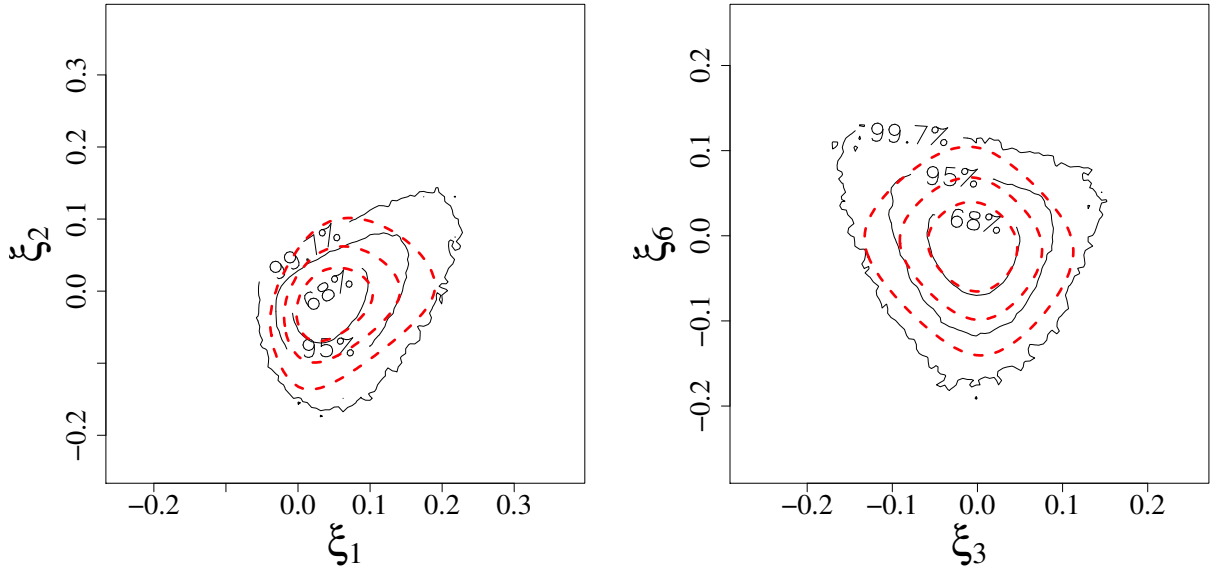


Figure 4.7: $p(\xi_1, \xi_2)$ (left) and $p(\xi_3, \xi_6)$ (right) for a field with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$. The dashed red contours show the approximation obtained from a Gaussian copula.

functions (CDFs) of $\xi_i$. In the simplest case, the copula function is assumed to be a Gaussian.

Copulae have previously been used in cosmology, e.g. for the PDF of the density field of LSS (Scherrer *et al.* 2010), or the weak lensing convergence power spectrum (see Sato *et al.* 2011 and its companion paper Sato *et al.* 2010). In the case of a Gaussian copula, the bivariate joint PDF can be calculated (see e.g. Sato *et al.* 2011) as

$$\begin{aligned} p\left(\xi_1, \xi_2\right) &= \frac{1}{\sqrt{2\pi \ \det\left(C\right)}} \ \exp\left\{-\frac{1}{2}\left(\boldsymbol{q}-\boldsymbol{\mu}\right)^{\mathrm{T}} C^{-1}\left(\boldsymbol{q}-\boldsymbol{\mu}\right)\right\} \\ &\times \prod_{i=1}^{2}\left(\frac{1}{\sqrt{2\pi}\sigma_i} \ \exp\left\{-\frac{\left(q_i-\mu_i\right)^2}{2\sigma_i^2}\right\}\right)^{-1} \cdot p\left(\xi_i\right), \end{aligned} \tag{4.37}$$

where $q_i = \Phi^{-1}_{\mu_i,\sigma_i}\left(P\left(\xi_i\right)\right)$. Here, $\boldsymbol{\mu}$ and $C$ denote the mean and covariance matrix of the copula function, $\Phi_{\mu,\sigma}$ is the Gaussian CDF. Note that contrary to our usual notation, here the indices of $\xi$ are purely a numbering, and Eq. (4.37) can in fact be applied for arbitrary lags.

To calculate the copula likelihood, we implement the analytical univariate formulae for $p_i(\xi_i)$ and $P_i(\xi_i)$ derived by Keitel and Schneider (2011); the mean and covariance matrix of the Gaussian copula are calculated directly from the simulated $\{\boldsymbol{\xi}\}$-sample. Fig. 4.7 shows the bivariate PDFs from the simulation (black contours) as well as the copula likelihood (red dashed contours) for two different combinations of lags. It is apparent that the copula likelihood does not describe the true PDF very well. In particular, it does not even seem to be a more accurate description than the simple multivariate Gaussian used in the left-hand panel of Fig. 4.1, leading us to the conclusion that our quasi-Gaussian approximation should be favored also over the copula likelihood. Of course, the accuracy of the latter might improve if a more realistic coupling than the Gaussian one was found.

### 4.5.2 Box-Cox transformations

As previously described, a central idea of the quasi-Gaussian approach is to transform the correlation function $\xi$ to a new variable $y$, which is (almost perfectly) Gaussian distributed. This suggests testing the performance of another, well-established method of transforming random variables in order to Gaussianize them, namely Box-Cox transformations. They are a form of power transforms originally introduced by Box and Cox (1964) and have been used in astronomical applications, see e.g. Joachimi *et al.* (2011) for results on Gaussianizing the one-point distributions of the weak gravitational lensing convergence.

For a sample of correlation functions $\{\xi_i\}$ at a certain lag, the form of the Box-Cox transformation is

$$\bar{\xi}_i(\lambda_i, a_i) = \begin{cases} \left[(\xi_i + a_i)^{\lambda_i} - 1\right]/\lambda_i & \lambda_i \neq 0 \\ \ln(\xi_i + a_i) & \lambda_i = 0 \end{cases} , \tag{4.38}$$

with free transformation parameters $\lambda_i$ and $a_i$, where the case $\lambda_i = 0$ corresponds to a log-normal transformation. For the following statements, we determine the optimal Box-Cox parameters using the procedure of maximizing the log-likelihood for $\lambda_i$ and $a_i$ explained in Joachimi *et al.* (2011) and references therein.

Note that since we cannot assume the transformation parameters to be identical for each lag $i$, we need to determine the full sets $\{\lambda_i\}$ and $\{a_i\}$. There are two different ways of addressing this: Since we are, in the end, interested in the multivariate likelihood of the correlation function, the most straightforward approach is to optimize the sets of Box-Cox parameters $\{\lambda_i\}$ and $\{a_i\}$ in such a way that the full $(n+1)$-variate distribution $p(\bar{\xi}_0, \bar{\xi}_1, \ldots, \bar{\xi}_n)$ of the transformed variables $\bar{\xi}_i$ is close to a multivariate Gaussian. Alternatively, one can treat all univariate distributions $p(\xi_i)$ separately, i.e. determine the optimal Box-Cox parameters in such a way that the univariate PDFs $p(\bar{\xi}_0), \ldots, p(\bar{\xi}_n)$ of the transformed variables are univariate Gaussians.

The first approach, i.e. trying to Gaussianize the full $(n+1)$-variate distribution, turns out to be unsuccessful: The multivariate moments (as defined in Sect. 4.3.1) of the transformed quantities $\bar{\xi}$ are hardly any different from those of the original correlation functions $\xi$. Additionally, there is barely any improvement in the Gaussianity of the univariate distributions $p(\bar{\xi}_i)$ compared to $p(\xi_i)$. In contrast to that, the transformation from $\xi$ to $y$ used in our calculation of the quasi-Gaussian likelihood resulted in an improvement in skewness and kurtosis by about an order of magnitude, as described in Sect. 4.3.1.

The second approach, i.e. treating all univariate distributions independently by trying to Gaussianize them separately, of course leads to lower univariate skewness and kurtosis in the transformed quantities $\bar{\xi}_i$; however, the multivariate moments are again almost unchanged compared to the untransformed correlation functions, indicating that this approach does not lead to a better description of the multivariate correlation function likelihood.

Thus, in summary, the quasi-Gaussian approach seems far more accurate than using Box-Cox transformations. This is not too surprising, since the latter obviously cannot properly take into account correlations between the different random variables (i.e. the $\xi_i$), whereas in contrast to that, the transformation $\xi \to y$ is specifically tailored for correlation functions and thus mathematically better motivated than a general Gaussianizing method such as power transforms.

CHAPTER 5

# Performance of the quasi-Gaussian approximation in a likelihood analysis

In the previous chapters, we described the fundamental constraints on correlation functions and showed how to use them in order to obtain the quasi-Gaussian approximation for the probability distribution of $\xi$. Furthermore, we discussed some subtleties in the computation of the quasi-Gaussian PDF and performed checks to show that it agrees well with the PDF obtained from simulations. In this section, we will construct an application for this new description for the probability distribution of correlation functions, i.e. we will test its performance as a likelihood function in a Bayesian parameter estimation analysis.

## 5.1 Setup

As explained in Sect. 2.2.4, the central equation for Bayesian parameter estimation is Bayes' Theorem, which states how to compute the posterior probability of a vector of model parameters $\boldsymbol{\theta}$ given some data, in our case a correlation function $\xi$. In this setup, the theorem reads

$$p(\boldsymbol{\theta}|\xi) = \frac{\mathcal{L}(\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\xi)}. \tag{5.1}$$

In the following, we will study the impact of using the quasi-Gaussian PDF as likelihood function $\mathcal{L}(\boldsymbol{\theta}) \equiv p(\xi|\boldsymbol{\theta})$.

For the sake of simplicity, we use artificial data instead of correlation functions actually measured on the sky in our analysis – applying the simulation methods described in Sect. 4.1, we generate 400 000 realizations of the correlation function of a Gaussian random field with $N = 64$ grid points and a Gaussian power spectrum with $L\ k_0 = 100$. To complete the setup of our analysis, we have to define a set of parameters about which we wish to obtain inference from this data. We choose the parameters of the power spectrum (i.e. its amplitude $A$ and its width $k_0$), so $\boldsymbol{\theta} = (A, k_0)$. To facilitate this choice of parameters, we parametrize the power spectrum as $P(k) = A \cdot 100/\left(Lk_0\right)\ \exp\{-(k/k_0)^2\}$, where we choose $A = 1\,\mathrm{Mpc}$ and $k_0 = 1\,\mathrm{Mpc}^{-1}$ as fiducial values (corresponding to $L\ k_0 = 100$ and a field length $L = 100\,\mathrm{Mpc}$). Since we aim to focus our analysis on the role of the likelihood function, we use a flat prior for $\boldsymbol{\theta}$ – thus, and due to the fact that the denominator in Eq. (5.1) (the Bayesian evidence) acts solely as a

normalization in the context of parameter estimation (again, see Sect. 2.2.4), the shape of the posterior $p(\boldsymbol{\theta}|\xi)$ is determined entirely by the likelihood.

In order to assess the effect of the quasi-Gaussian likelihood on the shape of the posterior, in a first analysis, we use the classical Gaussian approximation for the likelihood, which is the usual approach in the literature, as explained in Sect. 1. The Gaussian likelihood reads

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &\equiv p(\xi|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C_\xi}} \\ &\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\xi}\left(\boldsymbol{\theta}\right) - \boldsymbol{\xi}_{\mathrm{fid}}\right)^{\mathrm{T}} \cdot C_\xi^{-1} \cdot \left(\boldsymbol{\xi}\left(\boldsymbol{\theta}\right) - \boldsymbol{\xi}_{\mathrm{fid}}\right)\right\}, \end{aligned} \tag{5.2}$$

where $\boldsymbol{\xi} \equiv (\xi_0, \ldots, \xi_{n-1})^{\mathrm{T}}$, and $C_\xi$ denotes the covariance matrix computed from the $\{\boldsymbol{\xi}\}$-sample; note that we use only $n = 32$ $\boldsymbol{\xi}$-components, since the last 32 components yield no additional information due to periodicity. In an analysis of actual data, the measured value of the correlation function would have to be inserted as mean of the Gaussian distribution – since in our toy-model analysis, we use our simulated sample of correlation functions as data, we insert $\boldsymbol{\xi}_{\mathrm{fid}} \equiv \boldsymbol{\xi}\left(\boldsymbol{\theta}_{\mathrm{fid}}\right)$ instead, which is practically identical to the sample mean of the generated realizations.

## 5.2 Components of the quasi-Gaussian likelihood

Since the quasi-Gaussian PDF has a more complicated structure than the Gaussian likelihood given in Eq. (5.2), its computation in the context of our Bayesian analysis requires some attention, and we will thus study its different components in more detail. For our analysis, we transform the simulated realizations of the correlation function to $y$-space and calculate the likelihood as

$$\begin{aligned} p(\xi|\boldsymbol{\theta}) &= \frac{1}{(2\pi)^{(n-1)/2}\sqrt{\det C_y}} \\ &\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}\left(\boldsymbol{\theta}\right) - \langle \boldsymbol{y}\left(\xi_0\right)\rangle\right)^{\mathrm{T}} \cdot C_y^{-1} \cdot \left(\boldsymbol{y}\left(\boldsymbol{\theta}\right) - \langle \boldsymbol{y}\left(\xi_0\right)\rangle\right)\right\} \\ &\times p(\xi_0) \cdot \left|\det\left(J^{\xi\to y}\right)\right|. \end{aligned} \tag{5.3}$$

Here, instead of inserting the fiducial value $\boldsymbol{y}_{\mathrm{fid}}$ as "measured value", we incorporate the $\xi_0$-dependence of the mean $\langle \boldsymbol{y}\rangle$ in the same way as described in Sect. 4., i.e. by calculating the average only over those realizations with a $\xi_0$-value close to the "current" value of $\xi_0$, which is the one determined by the fixed value of $\boldsymbol{\theta}$, i.e. $\xi_0(\boldsymbol{\theta})$. In contrast to that, $C_y$ denotes the covariance matrix of the full $y$-sample, meaning that we neglect its $\xi_0$-dependence, although we have shown in Sect. 4.3.1 that incorporating this dependence increases the accuracy of the quasi-Gaussian approximation. The reason for this is that for some values of $\xi_0$, the number of realizations with a $\xi_0$-value close to it is so small that the sample covariance matrix becomes singular. However, since the toy-model analysis presented in this section is about a proof of concept rather than about maximizing the accuracy, this is a minor caveat – of course, when applying our method in an analysis of real data, the $\xi_0$-dependence of $C_y$ should be taken into account. It should also be mentioned that apart from the $\xi_0$-dependence, we also neglect any possible dependence of the covariance matrix on the model parameters $\boldsymbol{\theta}$, since this is not

expected to have a strong influence on parameter estimation (for example in the case of BAO studies, Labatie *et al.* 2012b show that even if $C$ does have a slight dependence on the model parameters, incorporating it in a Bayesian analysis only has a marginal effect on cosmological parameter constraints).

The $\boldsymbol{\theta}$-dependence of the last two terms in Eq. (5.3) also merits some discussion, in particular, the role of the fiducial model parameters $\boldsymbol{\theta}_{\mathrm{fid}}$ has to be specified: While the Gaussian likelihood discussed previously is of course centered around the fiducial values by construction, since $\boldsymbol{\xi}\left(\boldsymbol{\theta}_{\mathrm{fid}}\right)$ is inserted as mean of the Gaussian distribution, this cannot be done in the case of the quasi-Gaussian likelihood due to its more complicated mathematical form.

The $p(\xi_0)$-term in Eq. (5.3) can be treated in a straightforward way: Namely, we fix the shape of the PDF $p(\xi_0)$ by determining it from the fiducial power spectrum parameters $\boldsymbol{\theta}_{\mathrm{fid}}$ and then evaluate it at the current value $\xi_0(\boldsymbol{\theta})$, thus computing this term as $p_{\boldsymbol{\theta}_{\mathrm{fid}}}(\xi_0(\boldsymbol{\theta}))$. To be more explicit, the shape of the analytically known univariate PDF $p(\xi)$, given in Eq. (4.1), is determined by the values of $C_n$ and $C_m$. As described in Sect. 4.2, in the case of $p(\xi_0)$, the $C_n$ are simply equal to the variance $\sigma_n^2$ of the underlying random field in Fourier space, which is in turn related to the power spectrum via $P(k_n) = \sigma_n^2\, L$, where $L$ is the length of the random field. Thus, we compute $C_n$ as well as $C_m$ using the fiducial power spectrum (given by $\boldsymbol{\theta}_{\mathrm{fid}}$), and then evaluate the resulting PDF $p(\xi_0)$ at $\xi_0(\boldsymbol{\theta})$.

The last term, i.e. the determinant of the transformation matrix, however, has to be evaluated for the fiducial value, yielding $|\det(J^{\xi\to y}(\boldsymbol{\theta}_{\mathrm{fid}}))|$. Thus, this term has no $\boldsymbol{\theta}$-dependence at all and plays no role in parameter estimation; similarly to the Bayesian evidence, it does not have to be computed, but can rather be understood as part of the normalization of the posterior. This can be explained in a more pragmatic way: Assume that a specific value of $\boldsymbol{\xi}$ has been measured and one wants to use it for inference on the underlying power spectrum parameters, incorporating the quasi-Gaussian likelihood. Then one would transform the measurement to $y$-space and rather use the resulting $\boldsymbol{y}$-vector as data in the Bayesian analysis than $\boldsymbol{\xi}$, and thus, the $|\det J|$-term would not even show up when writing down the likelihood. Here, we nonetheless included it in order to follow the train of thought from Sect. 4 and for better comparison with the Gaussian likelihood in $\xi$-space.

Note that the previous argument cannot be applied to the $p(\xi_0)$-term: Calculating it as $p(\xi_0(\boldsymbol{\theta}_{\mathrm{fid}})|\boldsymbol{\theta}_{\mathrm{fid}})$ (i.e. as independent of $\boldsymbol{\theta}$, making it redundant for parameter estimation), yields biased results, as we shall see in the following. Furthermore, incorporating the non-negligible $\xi_0$-dependence of the mean $\langle\boldsymbol{y}\rangle$ immediately requires the introduction of the conditional probability in Eq. (4.4), thus automatically introducing the $p(\xi_0)$-term.

To further study the role of the different terms of the quasi-Gaussian likelihood, we plot the two central components separately in Fig. 5.1. For the left panel, we drop not only the $|\det(J^{\xi\to y})|$-term from Eq. (5.3) (which has no impact on the parameter estimation, as previously explained), but also the $p(\xi_0)$-term – thus, this panel shows the impact of the Gaussian component of the quasi-Gaussian likelihood on the estimation of the parameters $A$ and $k_0$. While the kinks in the contours at low values of $A$ are simply due to the steep shape of the contours and the low resolution of the graph (we divided the displayed region of the $A-k_0$-plane into $50 \times 50$ pixels to produce the plot), it is notable that the contours of the likelihood show a strong degeneracy in the $A$ of the power spectrum.

This justifies our previous argument on how to evaluate the $p(\xi_0)$-term: Since the left panel of Fig. 5.1 shows what the posterior would look like if we evaluated the term as $p(\xi_0(\boldsymbol{\theta}_{\mathrm{fid}})|\boldsymbol{\theta}_{\mathrm{fid}})$, i.e. independent of $\boldsymbol{\theta}$, the visible degeneracy in $A$ clearly shows that by doing so, we do not use the full information contained in the measured correlation function (in particular in $\xi_0$; recall

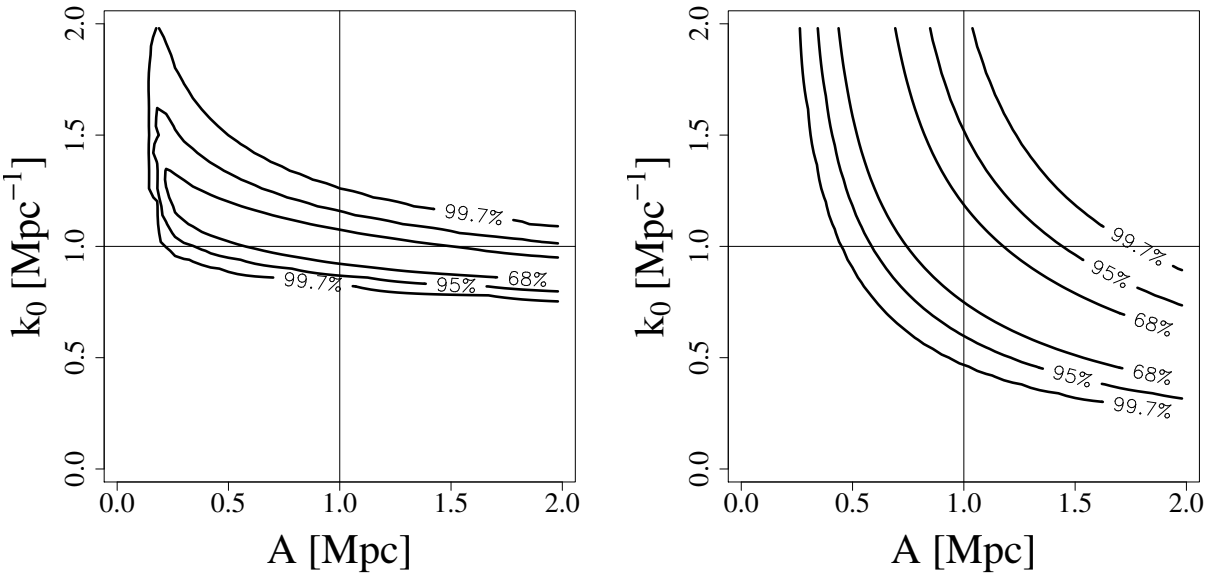


Figure 5.1: The components of the quasi-Gaussian likelihood. For these plots, we neglect all terms of the quasi-Gaussian likelihood except for one (namely the Gaussian in $y$-space for the left panel and the $p(\xi_0)$-term for the right panel), thus illustrating the impact of the separate terms on the parameter estimation of $\boldsymbol{\theta} = (A, k_0)$. The horizontal and vertical lines are the fiducial values $(1.0, 1.0)$.

that the transformation from $\boldsymbol{\xi}$ to $\boldsymbol{y}$ involves only ratios $\xi_i/\xi_0$). Thus, in order to get stronger constraints on the value of $A$, we need to incorporate the $p(\xi_0)$-term, which is plotted separately in the right panel of Fig. 5.1. Clearly, this term of the quasi-Gaussian likelihood does have an $A$-dependence, which can be understood when considering the conditional probability $p(\xi_0|A)$: It can be written in the form

$$p(\xi_0|A) = \frac{1}{A}\, f\left(\frac{\xi_0}{A}\right), \tag{5.4}$$

as can be seen when integrating both sides over $\xi_0$, yielding 1 due to normalization. However, the function $f$ is constant, since $\xi_0$ is proportional to $A$ (recall that is is computed as a Fourier integral over $P(k)$). Hence, the conditional probability $p(\xi_0|A)$ is proportional to $A^{-1}$, meaning that in the framework of our Bayesian analysis, large values of $\xi_0$ favor small values of $A$.

In summary, we can only use the quasi-Gaussian likelihood to extract the full information contained in the data if we treat the different terms and their dependence on the fiducial values $\boldsymbol{\theta}_{\mathrm{fid}}$ as previously described.

## 5.3 Results

Combining both the Gaussian and $p(\xi_0)$-term of the quasi-Gaussian likelihood (and assuming a flat prior for $\boldsymbol{\theta}$) finally allows us to compute the posterior distribution and to compare it to the one obtained when using the classical Gaussian likelihood. The posteriors resulting from the two analyses can be seen in Fig. 5.2 – here, the left-hand panel shows the result for the case of a Gaussian likelihood as given in Eq. (5.2), and the right-hand one is the result of the quasi-Gaussian analysis, where we put together the different components of the quasi-Gaussian likelihood previously plotted. Already for this simple toy model, the impact of the more accurate likelihood on the posterior is visible. The difference may be larger for a

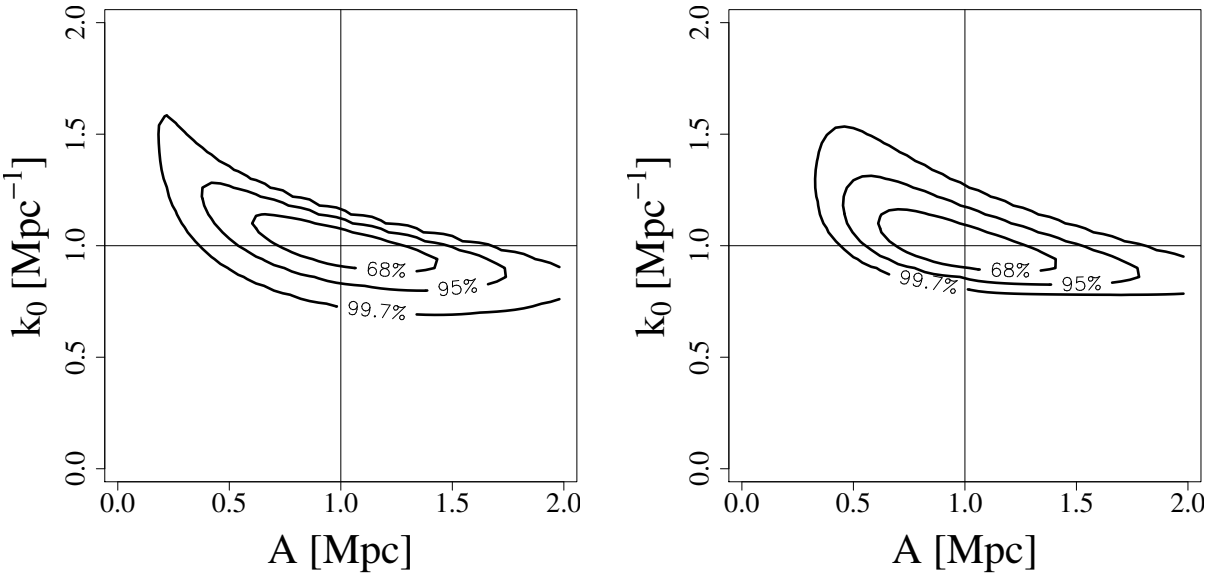


Figure 5.2: Posterior probability for the power spectrum parameters $(A, k_0)$ using the Gaussian (left panel) and the quasi-Gaussian (right panel) likelihood and flat priors.

different choice of power spectrum, where the deviation of the likelihood from a Gaussian is more pronounced. Nonetheless, it is evident that the change in the shape of the contours is noticeable enough to have an impact on cosmological parameter estimation.

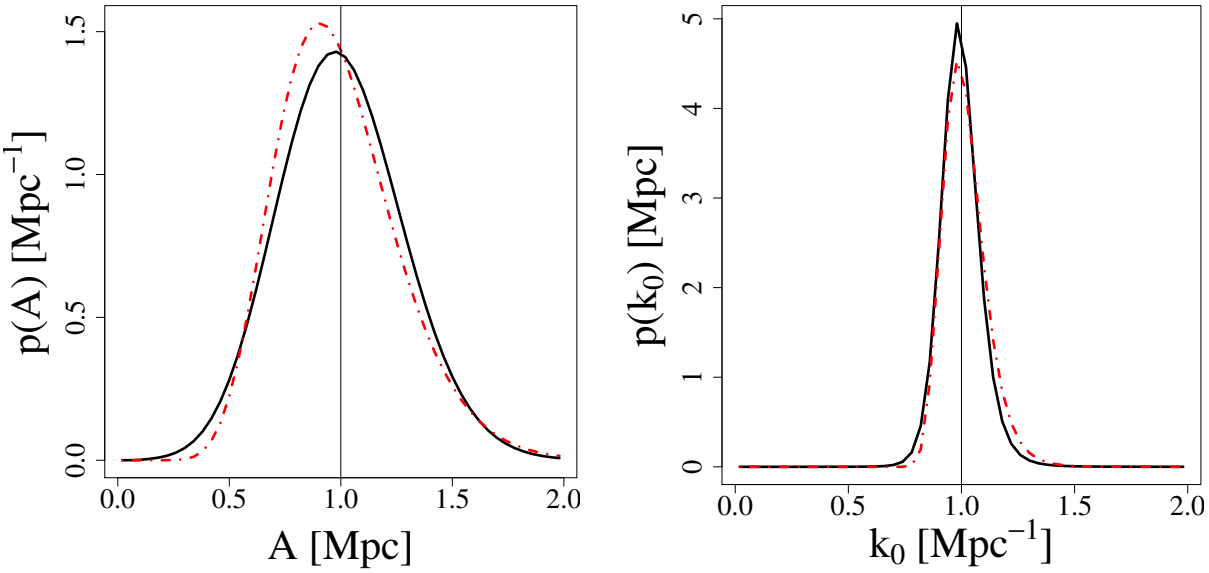


Figure 5.3: Marginalized posterior probabilities for the power spectrum parameters $A$ and $k_0$. The solid black curves are the results obtained from the Gaussian likelihood, whereas the red dot-dashed curves show the results from the quasi-Gaussian analysis.

Fig. 5.3 shows the marginalized posteriors for $A$ and $k_0$, again for the Gaussian (black solid curve) and the quasi-Gaussian (red dot-dashed curve) case. As for the full posterior, there is a notable difference. While it may seem alarming that the marginalized posteriors in the quasi-

Gaussian case are not centered around the fiducial value, this is in fact not worrisome: First, as explained in Sect. 2.2.4, there are different ways of obtaining parameter estimates in Bayesian statistics, and in the case of a skewed posterior (as the quasi-Gaussian one), the maximum a posteriori (MAP) estimate, i.e. the maximum of the posterior distribution, is not necessarily the most reasonable one. Furthermore, in our case, it should again be stressed that the Gaussian likelihood is of course constructed to be centered around the fiducial value, since $\boldsymbol{\theta}_{\mathrm{fid}}$ is explicitly put in as mean of the distribution, whereas the quasi-Gaussian likelihood is mainly constructed to obey the boundaries of the correlation functions. These are mathematically fundamental constraints, and thus, although none of the two methods are guaranteed to be bias-free, the quasi-Gaussian one should be favored.

CHAPTER 6

# Constrained correlation functions from the Millennium Simulation

So far, we have examined statistical properties of correlation functions in a very general, mathematical framework. In this chapter, we will investigate our results in a more astrophysical context by applying them to cosmological correlation functions. Since we are not only interested in testing the relevance of the constraints on measured correlation functions, but also in their statistical properties, we will compute correlation functions from numerical simulations: The size of modern $N$-body simulations enables us to easily measure multiple realizations of $\xi$, thus allowing for an approximate determination of the underlying probability distribution.

## 6.1 $N$-body simulations

As described in Sect. 2.1.2, an analytical description of cosmological structure formation is not feasible as soon as density fluctuations become non-linear. Due to the increasing availability of high-performance computing resources, numerical simulations have become a common and very powerful tool to cope with this issue, since they allow for an analysis of cosmological models also in the non-linear regime. While these simulations have become more and more sophisticated over the last years (see Kuhlen *et al.* 2012 for a review of state-of-the-art methods), for our purposes, the impact of, for example, baryon physics is not as vital, and thus we restrict our discussion of $N$-body simulations in this section to the most basic principles, namely to DM-only simulations. Nonetheless, it is worth mentioning that by combining hydrodynamical and DM simulations, recent large-scale simulation projects such as Illustris (Vogelsberger *et al.* 2014) are even able to reproduce the observed galaxy populations, i.e. their distributions and internal properties.

The most fundamental idea of simulations of structure formation is to trace representative DM particles, since a simulation on the level of individual elementary particles is not possible on cosmological scales. Thus, one examines the behavior of $N$ macroscopic particles of mass $M$. As simulation volume, one typically chooses a comoving cube of side length $L$, where $L$ should be "large enough" for the cube to be representative, i.e. larger than the largest structures one wishes to simulate. Additionally, periodic boundary conditions are imposed.

Simulations typically start at very high redshift, where the initial conditions are usually fixed in such a way that the matter follows a Gaussian distribution with a power spectrum

corresponding to the theoretical (linear) power spectrum of the underlying cosmological model. In order to compute the evolution of the density field, the equations of motion of the particles have to be solved. In principle, this requires the computation of the force acting on each particle according to Newton's law; however, two problems arise: First, the force law would allow for collisions between particles, which is unphysical in this setup – it would simply be a consequence of the choice of macroscopic DM particles. Thus, one modifies the force law in such a way that it deviates from the Newtonian $1/r^2$ behavior on small scales, i.e. on scales below the so-called softening length $\epsilon$. The choice of $\epsilon$ is important, since it automatically determines a limit on the spatial resolution of the simulation. Typically, one chooses a softening length of the order of the mean separation between particles (which in turn depends on the particle mass $M$).

In practice, another challenge occurs: Usually, the number of particles is too large to compute the force on individual particles due to the large number of terms. Thus, this straightforward approach (the so-called particle-particle or PP method) is replaced by a more sophisticated one, namely the particle-mesh (PM) method, where, instead of calculating the force on a particle exactly, one computes an approximate force field: To do so, the particle distribution is Fourier transformed, and the gravitational acceleration is computed via Poisson's equation which has a simple form in Fourier space. Since an effective numerical implementation using fast Fourier transform (FFT) methods requires the introduction of a grid, one again has to pay attention to the effect on the simulation's spatial resolution – usually, the grid cell size is chosen in agreement with the softening length.

In order to optimize accuracy, it is also possible to split up the force law into different regimes by applying the PP method for short-range interactions and the PM method for long ranges, resulting in a so-called $P^3M$ (particle-particle particle-mesh) approach. All of the aforementioned simulation methods can be modified in numerous ways to optimize accuracy and performance – for example, tree code can be used, which essentially means that one divides the simulation volume into subcells and adjusts the grid resolution in each subcell.

Since running an $N$-body simulation requires a fixed choice of cosmology, changing the values of the underlying cosmological parameters in principle involves rerunning the simulation. This can easily become too expensive from a computational point of view, in particular if one wants to study the influence of the values of cosmological parameters on structure formation. One method that deals with this problem is simulation remapping: Angulo and White (2010) show that by rescaling length, mass, and velocity units as well as the time axis and the amplitudes of fluctuation modes, it is possible to adjust the output of a simulation to a (slightly) different set of cosmological parameters – for example, they rescale the output of a simulation with an underlying WMAP1 cosmology to the WMAP3 values of the parameters. This approach is taken one step further by Mead and Peacock (2014), who present a remapping which immediately acts on the halo catalog produced by the simulation instead of the particle distribution, thus facilitating the generation of, for example, mock galaxy catalogs for different sets of cosmological parameters.

## 6.2 The Millennium Simulation

The $N$-body simulation we use in the following analyses is the Millennium Run (Springel *et al.* 2005). Although it is no longer the largest cosmological simulation, the Millennium Simulation is still widely used in astrophysics and has spawned a great number of ground-breaking results – to date, more than 500 papers using Millennium-related data have been published. The

Table 6.1: Cosmological parameters and simulation parameters used for the Millennium Simulation.

| $\Omega_\mathrm{m}$ | $\Omega_\mathrm{b}$ | $\Omega_\Lambda$ | $h$ | $n_S$ | $\sigma_8$ | $L/\,(h^{-1}\,\mathrm{Mpc})$ | $N_\mathrm{p}$ | $M_\mathrm{P}/\,(h^{-1}\,\mathrm{M}_\odot)$ | $\epsilon/\,(h^{-1}\,\mathrm{kpc})$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 0.045 | 0.75 | 0.73 | 1 | 0.9 | 500 | $2160^3$ | $8.61 \times 10^8$ | 5 |

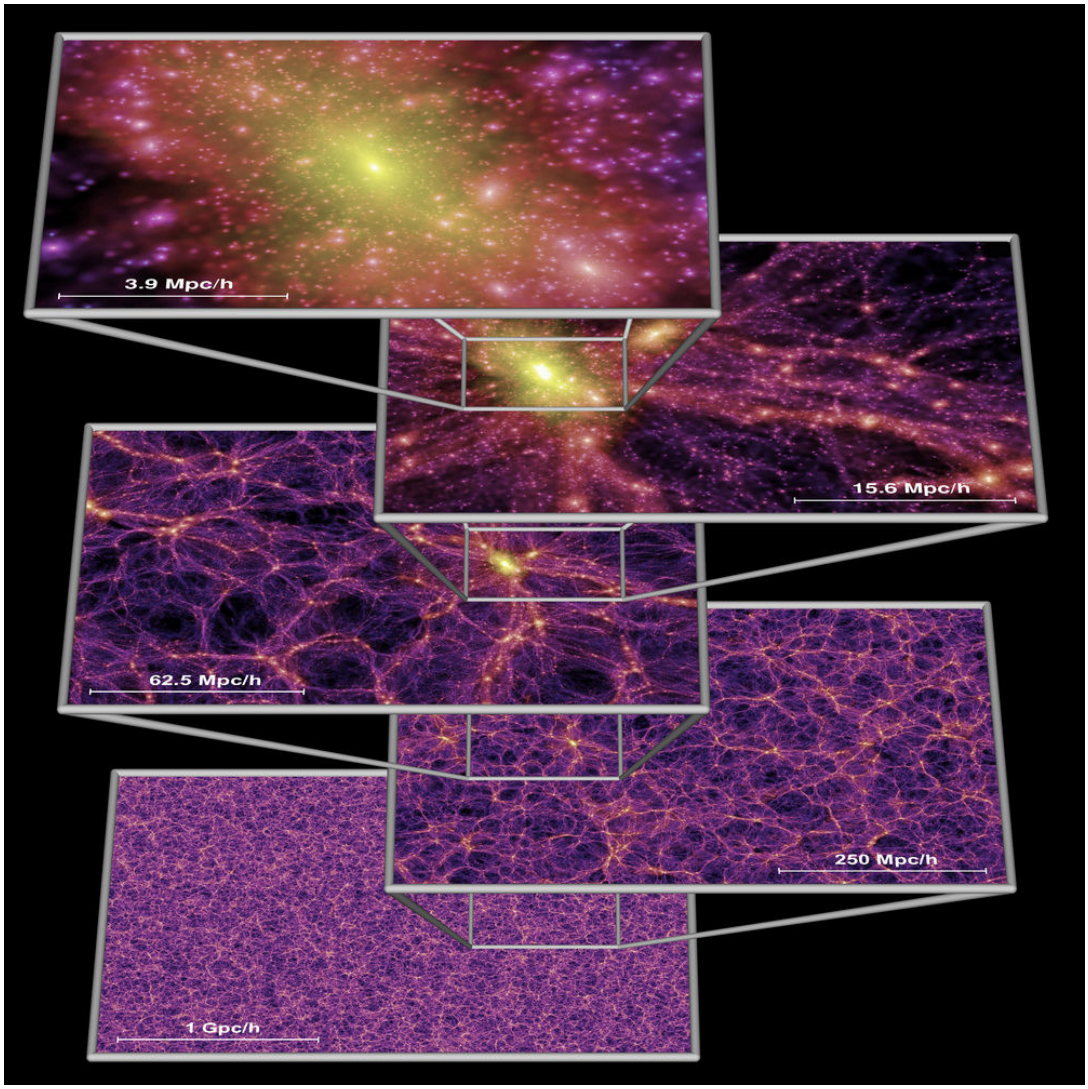


Figure 6.1: The DM density obtained from the Millennium Simulation at different scales, where each image displays the density projected in a slab of thickness 15 $h^{-1}$ Mpc (from Springel *et al.* 2005).

simulation makes use of a TreePM code, namely GADGET-2 (Springel 2005), to follow $\approx 10^{10}$ particles in a 500 $h^{-1}$ Mpc box from redshift $z = 127$ until the present. The parameters of the underlying cosmology as well as the simulation parameters are listed in Table 6.1. As an illustrative example of the output of the simulation, Fig. 6.1 shows the DM density at different

scales.

One advantage of using the Millennium Simulation is that it is easily accessible: A vast amount of simulation output is stored in a Structured Query Language (SQL) database, including for example properties of halos and subhalos, halo merger trees, and assembly histories of galaxies (see Lemson and the Virgo Consortium 2006 for a brief description of the database). The data can be accessed via a web interface – Fig. 6.2 shows an exemplary SQL query which produces a list of halos with a given mass threshold. Both the Millennium Simulation databases used in this work and the web application providing online access to them were constructed as part of the activities of the German Astrophysical Virtual Observatory (GAVO).

```
select haloID,np,m_Crit200,halfmassRadius,x,y,z
  from MPAHaloTrees..MR
 where snapnum=63
   and m_Crit200 > 100
   and x between 0 and 500
   and y between 0 and 500
   and z between 0 and 500
```

Figure 6.2: Example of an SQL query to access the output of the Millennium Run: This query produces a list of halos (including their number of particles, $M_{200}^{\mathrm{crit}}$, half-mass radius, and position) with a given mass (here: $M_{200}^{\mathrm{crit}} > 100 \times 10^{10}\ h^{-1}\,\mathrm{M}_{\odot}$) at a given redshift (specified by the snapshot number, where snapnum= 63 corresponds to $z = 0$) within a certain part of the simulation volume ($x$,$y$,$z$), in this case, the full box of size $500^3\ \left(h^{-1}\,\mathrm{Mpc}\right)^3$.

## 6.3 Computing correlation functions

In the following, we compute the correlation function of DM halos in the Millennium Simulation. Since we are not interested in redshift evolution, we only make use of the halo catalog from the $z = 0$ simulation snapshot, from which we then select typical galaxy-mass halos by choosing a mass cut $M_{200}^{\mathrm{crit}} > 10^{12}\ h^{-1}\,\mathrm{M}_{\odot}$, yielding a total number of 440 000 halos. In order to perform a statistical analysis, we require different realizations of the correlation function – thus, as a first attempt, we divide the full simulation cube into 1000 subcubes of size $50^3\ \left(h^{-1}\,\mathrm{Mpc}\right)^3$ and measure $\boldsymbol{\xi}$ in each of the subcubes.

The calculation of the correlation function can be implemented as described in Sect. 2.2.2: In addition to the halo catalog from the simulation, we need a random catalog, so for each subcube, we draw halo positions uniformly. We then determine the number of halo pairs for given pair separations in both the data and the random catalog as well as the cross-correlation. From the count rates $DD(s)$, $RR(s)$, $DR(s)$ at different pair separations $s$, we compute $\boldsymbol{\xi}$ using one of the estimators listed in Table 2.2. As explained in Sect. 3, in order to calculate the constraints, we need to measure the correlation function at equidistant lags, i.e. $\xi_n \equiv \xi(n \cdot \Delta s)$, where the maximum number of lags is $n = 8$ due to limitations of the numerical computation of the constraints.

The size of the random catalog merits some discussion – ideally, it should be infinitely large, i.e. $N_{\mathrm{rand}} \to \infty$. However, the computation of the pair separations is the most time-consuming step in the calculation of $\boldsymbol{\xi}$ and thus, the number of halos in the random catalogs for each of the 1000 subcubes is subject to practical limitations. We study the impact of the random catalog

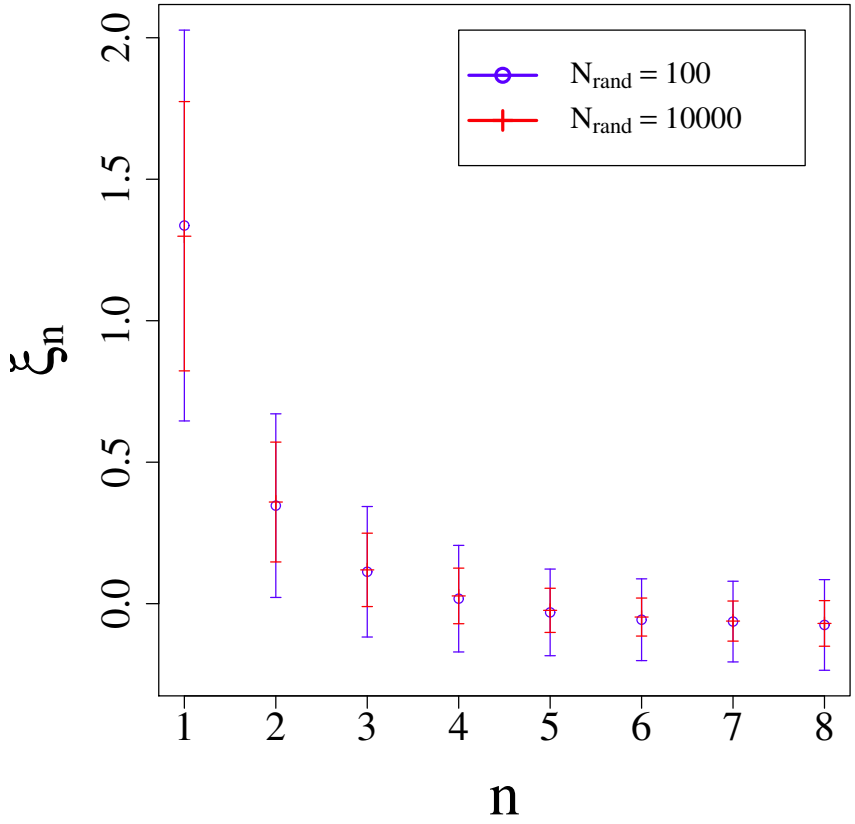


Figure 6.3: The correlation function from 1000 subcubes of the Millennium Simulation, computed using the LS estimator. The points and error bars show the correlation function $\xi_n \equiv \xi(n \cdot \Delta s)$ for $\Delta s = 5\ h^{-1}\,\mathrm{Mpc}$ (see text for details) averaged over the 1000 subcubes of size $50^3\ \left(h^{-1}\,\mathrm{Mpc}\right)^3$, as well as the standard deviation. For the blue circles, the random catalog for each subcube contained 100 halos, as opposed to 10000 halos for the red crosses.

size in Fig. 6.3: Here, we show the correlation function for an exemplary choice of lags – we will go into details on the choice of lags in the analysis of the constraints. For now, we choose $\Delta s = 5\ h^{-1}\,\mathrm{Mpc}$, i.e. we measure $\xi_1, \xi_2, \ldots, \xi_8$ at lags of $5, 10, \ldots 40\ h^{-1}\,\mathrm{Mpc}$. In practice, we need to allow for a range of pair separations in order to obtain sufficiently large numbers of pairs – thus, we adapt a bin size of width $1\ h^{-1}\,\mathrm{Mpc}$, so for example in the computation of $\xi_1$, we use all pairs with separations ranging from 4.5 to 5.5 $h^{-1}\,\mathrm{Mpc}$. For the auto-correlation $\xi_0$, i.e. the correlation function at zero lag (which we do not plot in the figure, but which is required for the calculation of the constraints), we count all pairs with very small separations, e.g. $s \leq 1\ h^{-1}\,\mathrm{Mpc}$ – again, we refer to the next section for a discussion on the measurement of $\xi_0$. The points and error bars show the mean and standard deviation over the 1000 subcubes of size $50^3\ \left(h^{-1}\,\mathrm{Mpc}\right)^3$; as estimator, we use the LS estimator, since it is the most commonly used one – in parts due to the fact that it is less sensible to the size of the random catalog than others (see Kerscher *et al.* 2000). For the blue circles, a small random catalog ($N_{\mathrm{rand}} = 100$ halos for each subcube) was used, whereas the choice of $N_{\mathrm{rand}} = 10000$ for the red crosses results in noticeably smaller standard deviations over the 1000 realizations at the cost of a longer computation time.

Thus, we aim to find a trade-off between those two values. First, one should note that,

although the mean correlation functions for the two random catalog sizes plotted in Fig. 6.3 do not seem to differ very much by eye, choosing the catalog size as small as $N_{\mathrm{rand}} = 100$ is a quite extreme case, since a large fraction of the realizations yield a diverging auto-correlation $\xi_0$ due to count rates in $RR$ or $DR$ being zero, at least when measuring it as previously described. However, when increasing the random catalog to 1000 halos per subcube, the mean correlation function for non-zero lag only shows a deviation of $\sim 1\,\%$ compared to the mean $\xi$ for $N_{\mathrm{rand}} = 10000$ (and even here, about a tenth of the realizations show a diverging $\xi_0$). At $N_{\mathrm{rand}} = 5000$, no such problems occur, and also the error bars as plotted in the figure become indistinguishable from those at $N_{\mathrm{rand}} = 10000$ – thus, a random catalog size of 5000 is a reasonable trade-off between accuracy and computational expenses.

An additional observation from Fig. 6.3 is that $\xi$ becomes negative for higher lags, i.e. around $20-25\ h^{-1}\,\mathrm{Mpc}$, which is unphysical – the reason for this is the so-called integral constraint (see e.g. Peacock 1999). This effect arises when measuring correlation functions in finite volumes, where the global mean density is unknown and is usually approximated by the mean observed density. In our case, it occurs in the normalization of count rates: As explained in Sect. 2.2.2, we normalize $DD$ using the total number of halo pairs in the subcube. However, the average halo number density in each subcube $\bar{n}_{\mathrm{sub}}$ is not necessarily a good approximation for the underlying global density, giving rise to integral constraints.

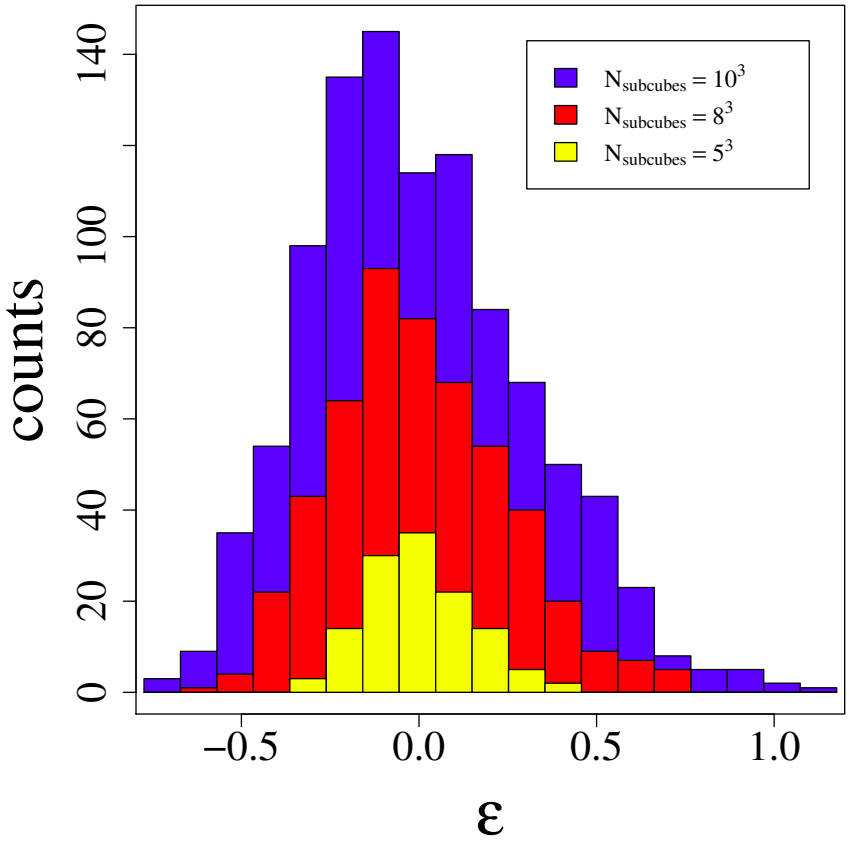


Figure 6.4: Histograms of the distributions of overdensities in the subcubes of the simulation volume, where $\epsilon$ denotes the number overdensity of halos in the subcube relative to the mean halo number density in the whole simulation box, see Eq. (6.1). The colors indicate the number of subcubes the simulation box was divided into.

One straightforward attempt to remedy this is to use the density $\bar{n}_{\mathrm{box}}$ from the whole simulation box to normalize the count rates – while this approach does indeed increase the values of $\xi$, it also greatly drives up the variance over the subcubes, thus making any further analysis hard. Alternatively, we investigate other ways of renormalizing the count rates: One way to address this is to estimate the overdensity $\epsilon$ in each subcube by comparing the mean number density in the subcube to the one from the whole simulation volume,

$$\bar{n}_{\mathrm{sub}} = \bar{n}_{\mathrm{box}} \, (1 + \epsilon) \, . \tag{6.1}$$

Assuming $\epsilon$ is small, we can Taylor expand the count rate $DD$ to first order, thus allowing us to correct the measured correlation function $\xi_{\mathrm{meas}}$ and to obtain the true one, $\xi_{\mathrm{true}}$. Applying the Peebles and Hauser (PH) estimator for simplicity, this procedure yields

$$\xi_{\mathrm{true}} = \xi_{\mathrm{meas}} \ (1 - 2\epsilon) - 2\epsilon. \tag{6.2}$$

However, this type of correction also results in a large variance of the correlation functions measured over the subcubes and is thus not practical.

Hence, as a final resort, we decrease number of subcubes in our analysis, and use the density $\bar{n}_{\mathrm{sub}}$ in the individual subcubes for the normalization of the count rates. While this is a trade-off in the sense that we obtain fewer realizations of $\xi$, it lessens the impact of the integral constraint by making the individual subcubes more representative, at the same time yielding a sufficiently small variance to allow for a precise analysis. In order to decide on the number of subcubes, we examine the distributions of the overdensity $\epsilon$, which we plot as histograms in Fig. 6.4. Here, we slice the simulation volume into different numbers of subcubes and compute the overdensity as defined in Eq. (6.1) in each subcube. It is apparent that the distribution $p(\epsilon)$ is very broad for the value $N_{\mathrm{subcubes}} = 10^3$ used so far – as expected, it becomes quite narrow for the case of $5^3$ subcubes, indicating that the integral constraint does not pose a problem in this case. The resulting correlation functions for the case of $8^3$ and $5^3$ subcubes (and otherwise same parameters as before, i.e. same lags and random catalog size) are shown in Fig. 6.5. Evidently, slicing the simulation volume into $5^3$ subcubes yields reasonable results, i.e. a non-negative correlation functions with a sufficiently small variance.

As a final ingredient in the measurement of correlation functions, we will briefly discuss the choice of estimator. To do so, instead of measuring $\xi_n$ at a few different lags $n$, it is advisable to compute $\xi(s)$ for all lags in each subcube – since this is obviously not doable in practice, we compute it at a high number of lags, meaning that we divide the range of pair separations into adjacent bins of width 0.2 $h^{-1}$ Mpc. As previously discussed, we slice the simulation volume into $5^3$ subcubes and use a fixed size of random catalog for each subcube, namely $N_{\mathrm{rand}} = 5000$. In the determination of $\xi(s)$ from the count rates, we try different estimators; the results are shown in Fig. 6.6. The lines of different color denote the five estimators from Table 2.2, where we again average over the 125 subcubes, and the gray shaded region depicts the standard deviation in the case of the most commonly used LS estimator. For the sake of clarity, in the left panel, we only plot scales from 8 to 40 $h^{-1}$ Mpc, whereas the right panel shows the correlation function for very low lags, i.e. $4 - 8$ $h^{-1}$ Mpc.

It is apparent that the numerous estimators yield very similar results, in particular compared to the standard deviation of the 125 realizations – we will briefly discuss the influence of the choice of estimator on the constraints on correlation functions at the end of the following section.

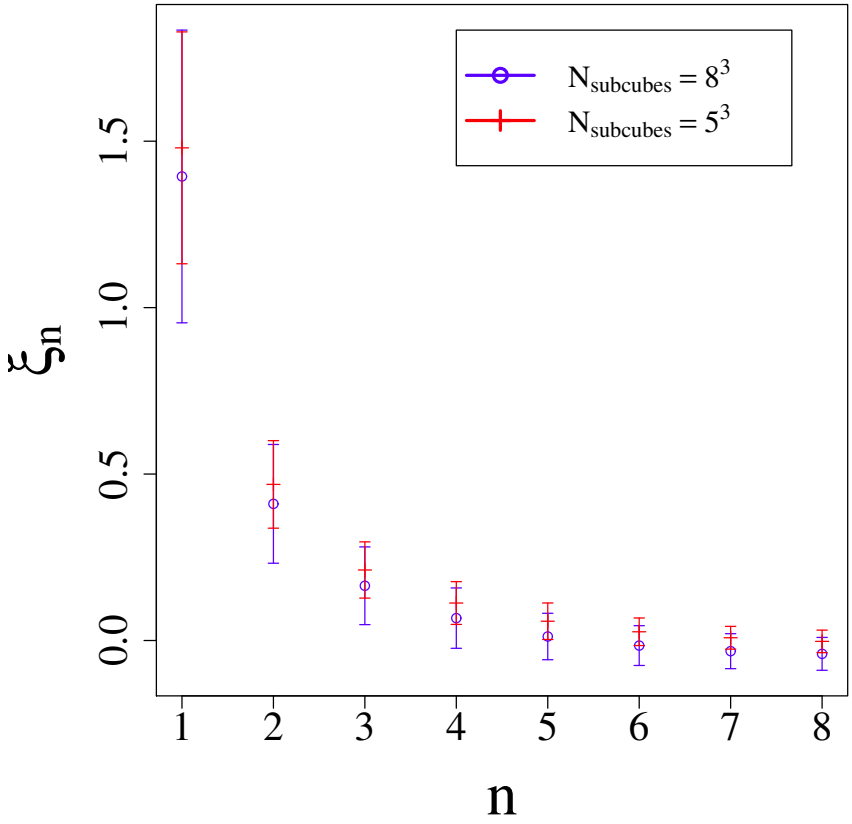


Figure 6.5: The correlation function from the Millennium Simulation, computed using the LS estimator, a random catalog size of 5000, and a lag distance of $\Delta s = 5\ h^{-1}\,\mathrm{Mpc}$. Again, the points and error bars show the mean and standard deviation computed over the subcubes of the simulation, where the simulation box was sliced into $8^3$ subcubes for the blue data points, as opposed to $5^3$ for the red ones.

## 6.4 Testing the constraints

In this section, we will investigate whether the correlation functions computed from the halo catalog of the Millennium Simulation obey the constraints introduced in Sect. 3. Since no analytical expressions for the constraints are available in the case of random fields in three dimensions, we have to compute them numerically – as explained in Sect. 3.2, this requires the calculation of the convex hull of the curve $\boldsymbol{c}(\lambda)$. While we argue in Sect. 3.3 that using 300 points to sample the hull yields sufficiently good agreement between the numerical and the exact analytical values of the constraints, we have to restrict ourselves to 270 points in the case of a 3D random field due to the computational costs – although the convex hull only has to be computed once and can then be used to determine the constraints for all sets of correlation coefficients, sampling the hull for a 3D random field with the given accuracy poses memory problems for the qhull algorithm, which is out of scope for us to resolve. However, this does not pose a problem: When comparing the accuracy of the numerical constraints as plotted in Fig. 3.4, it is apparent that the improvement in accuracy when going from 200 to 300 steps is far smaller than the one from 100 to 200, and thus, we expect the use of 270 points to be precise enough.

To test the constraints, we again measure the correlation function at eight lags with separa-

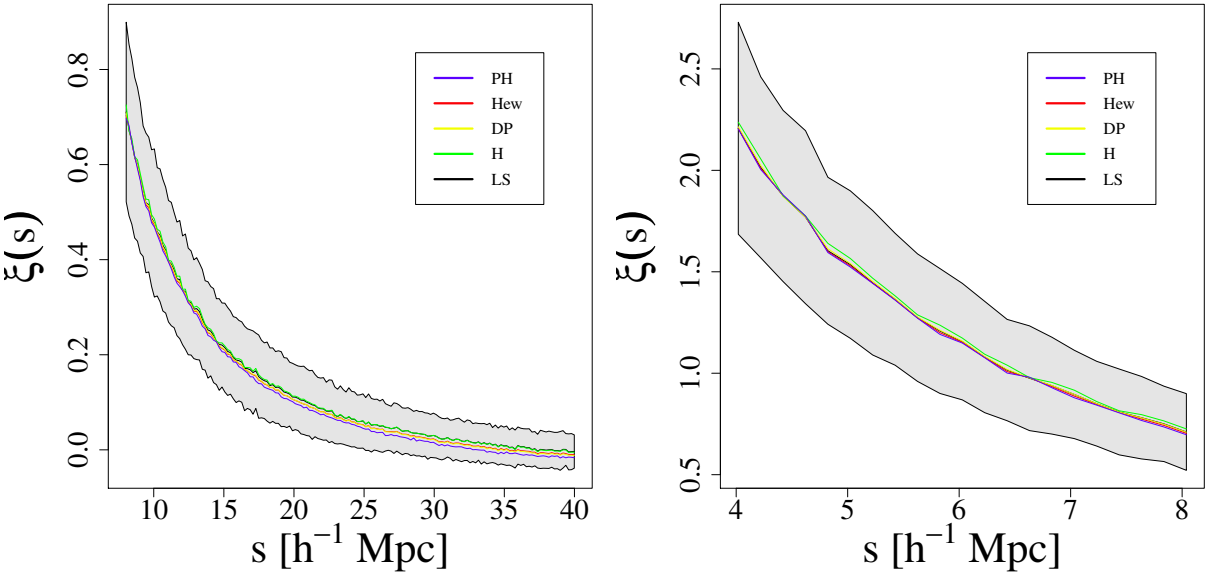


Figure 6.6: The correlation function measured for "all" lags (see text for details) as a function of the pair separation $s$, with a random catalog size of $N_{\rm rand} = 5000$. The lines denote the mean $\xi(s)$ measured in the 125 subcubes using the different estimators from Table 2.2, and the gray shaded region shows the standard deviation. In the left panel, the $s$-range from 8 to 40 $h^{-1}$ Mpc is plotted, and the right panel shows the results for very low lags, from 4 to 8 $h^{-1}$ Mpc.

tions of $\Delta s = 5\ h^{-1}$ Mpc and a bin width of 1 $h^{-1}$ Mpc (thus, for example $\xi_2$ is determined using all pairs with separations from 9.5 to 10.5 $h^{-1}$ Mpc). As described in the previous section, we divide the simulation volume into 125 subcubes and use a random catalog of size $N_{\rm rand} = 5000$ in each subcube. Since the constraints are best expressed in terms of the correlation coefficients, we compute $r_n \equiv \xi_n/\xi_0$ for each realization, as well as the upper and lower bounds $r_{n\rm u}$ and $r_{n\rm l}$. It turns out that the width of the $\xi_0$-bin has a strong huge influence, in particular on the width of the distributions of the correlation functions $r_n$. For illustrative purposes, we first choose a relatively broad bin, i.e. we measure $\xi_0$ for all pair separations from 0 to 2 $h^{-1}$ Mpc – this choice is primarily motivated by the fact that increasing the spread of the correlation coefficients over the 125 realizations allows us to test how close to the edges to the allowed region the $r_n$ move. Towards end of this section, we will further study the impact of the width of the $\xi_0$-bin.

One question that arises is how to visualize the constraints – the simplest approach to this are scatter-plots with dots for the individual realizations; an example in the $r_1 - r_2$-plane is shown in Fig. 6.7. Here, the red dots show the different realizations of $r_1$ and $r_2$, computed for the subcubes using the LS estimator; additionally, we plot iso-density contours containing 68, 95 and 99.7 % of the realizations. For the left panel, we sliced the simulation volume into 1000 subcubes, as opposed to 125 for the left panel. While we argued in the previous section that the choice of 1000 subcubes yields unphysical correlation functions due to the integral constraint, we still include the left-hand plot in order to illustrate how the higher number of subcubes greatly increases the spread of the correlation functions, which can clearly also be observed in $r$-space. In both panels, the upper and lower blue lines are the constraints, i.e. $r_{2\rm u,l}(r_1)$ – note that while it would be possible to construct them analytically from the corresponding Bessel functions in the case of the $r_1 - r_2$-plane, for simplicity, we use our numerical methods to compute the bounds for each realization of $r_1$ shown in the figure and plot them as connected

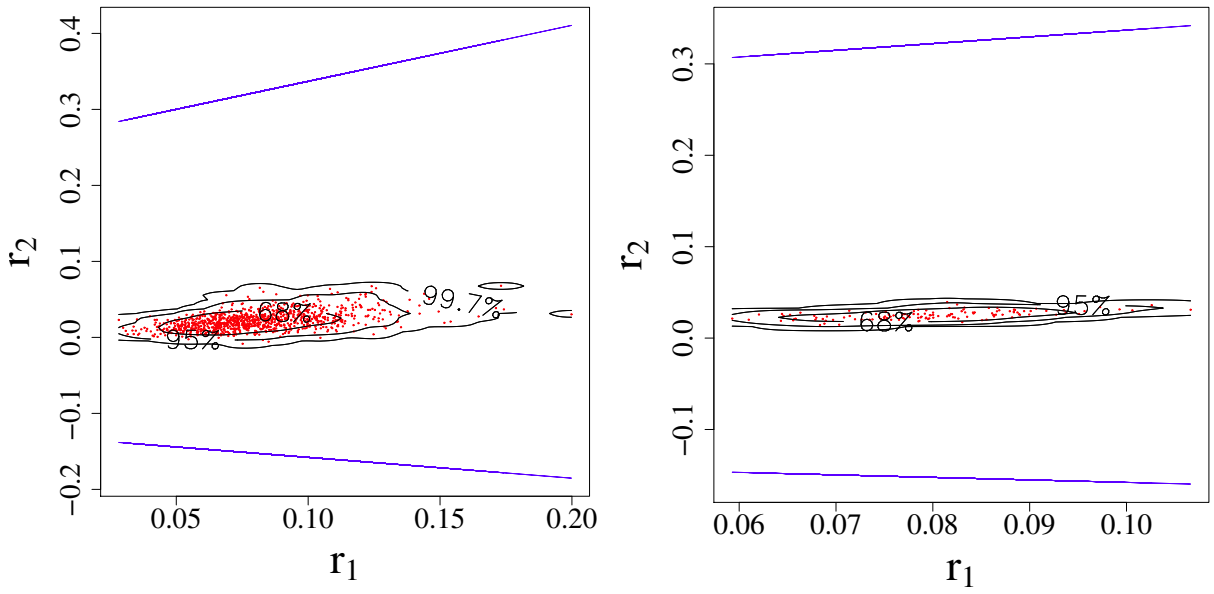


Figure 6.7: The correlation coefficients $r_1$ and $r_2$ measured from the halo catalogs in the subcubes of the Millennium Simulation using the LS estimator, where we slice the simulation volume into 1000 subcubes for the left-hand panel and 125 for the right-hand one, and the random catalog for each subcube contains 5000 halos. In both cases, we measure $\xi$ at lags of separation $\Delta s = 5\ h^{-1}\,\mathrm{Mpc}$, and use all halo pairs with pair separations of 0 to 2 $h^{-1}\,\mathrm{Mpc}$ to compute $\xi_0$. The red dots show the 1000 (125) realizations, and the black lines are iso-density contours containing the given percentages of the realizations. The upper and lower constraints $r_{2\mathrm{u,l}}(r_1)$, computed individually for each realization of $r_1$, are shown as a blue lines.

lines. It can be clearly seen that all realizations lie well inside the constraints.

As an additional way of depicting the constraints, we apply a part of the quasi-Gaussian transformation in order to map the allowed range of the correlation coefficients to $(-1, +1)$, namely by transforming the correlation coefficients $r_n$ to

$$x_n = \frac{2r_n - r_{n\mathrm{u}} - r_{n\mathrm{l}}}{r_{n\mathrm{u}} - r_{n\mathrm{l}}}. \tag{6.3}$$

Scatter-plots of the 125 realizations of $x_1$ and $x_8$ (using the same lags, random catalog size, and estimator as previously described) are shown in Fig. 6.8. While the left panel again shows the relatively narrow distribution for the lowest lag, the transformed correlation coefficients at highest lag plotted in the right panel extend closer to the edge of the allowed range, but still lie clearly inside.

Since scatter-plots are a slightly inept way of visualizing the constraints, we make use of a powerful and well-established tool to visualize probability distributions, namely box-and-whisker plots: In this kind of diagram, samples from a distribution are displayed as boxes whose upper and lower borders show the first and third quartiles of the sample, i.e. the values that split off the upper and lower 25 % of the data. Additionally, one usually plots the sample median (i.e. the second quartile) as a line inside the box, as well as two whiskers, where in the most widely used type of box-and-whisker plot, the ticks at the end of the whiskers denote the minimum and maximum of the data. In our case, we instead use them to display the upper and lower constraints: Since $r_{n\mathrm{u,l}}$ are functions of all $r_i$ with $i < n$, we show the mean $r_{n\mathrm{u}}$ and $r_{n\mathrm{l}}$

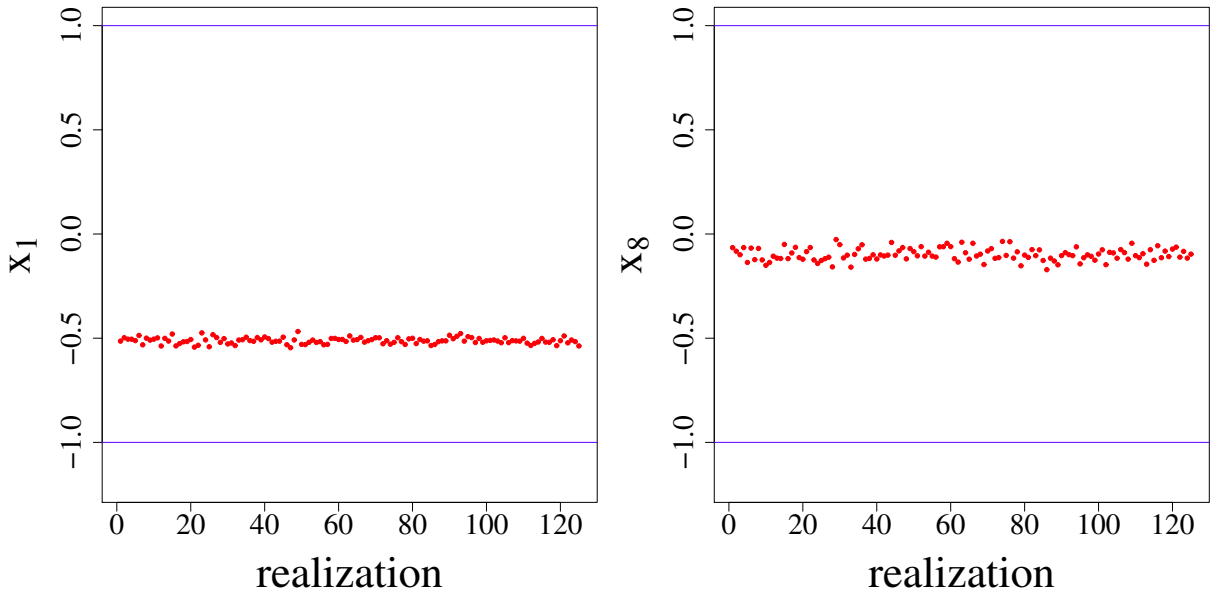


Figure 6.8: The transformed correlation coefficients at lowest and highest lag, as defined in Eq. (6.3). Again, each dot corresponds to one realization, i.e. the value measured in one of the 125 subcubes of the Millennium Simulation using the same lags, random catalog size, and estimator as before (see text and previous figure captions for details).

over all realizations for plots in $r$-space. For the transformed values $x_n$, the bounds are simply $\pm 1$, so there is no need to average over the realizations.

Fig. 6.9 shows box-and-whisker plots of $r_n$ and $x_n$ at all eight lags $n$, where we use the same lags and random catalog size as before, as well as the LS estimator. Again, it can be seen that the constraints are clearly obeyed and although the distributions becoming broader for increasing lag, the boxes showing the upper and lower quartiles only occupy a small fraction of the allowed region. The distributions are not necessarily centered within the allowed region, which is not surprising, since their exact shape and position also depend on the underlying power spectrum.

Furthermore, we compare the impact of the choice of estimator in Fig. 6.10. Again, we show the transformed correlation coefficients at lowest and highest lag, i.e. $x_1$ and $x_8$, but we make use of the numerous estimators described in Sect. 2.2.2 and listed in Table 2.2. As was to be expected from Fig. 6.6, the different estimators yield quite similar results; one thing that is worth noting is the slightly higher variance of the Peebles and Hauser estimator for the higher lag. In particular, none of the estimators shift the distribution within the allowed region.

As mentioned in the beginning of this section, the main influence on the variance of the distributions in $\xi$- and correspondingly in $r$- and $x$-space seems to be the width of the $\xi_0$-bin – in Fig. 6.11 we investigate this observation and also study the impact of the choice of the separation between the lags at which we measure $\xi$. In the two panels of the figure, we again show box-and-whisker plots of the transformed correlation coefficients at highest and lowest lag, i.e. $x_1$ and $x_8$, and we vary the separation $\Delta s$ of the lags as well as the bin widths of the pair separations used to measure $\xi_0$ and the correlation functions at non-zero lag, $\xi_1 \dots \xi_8$. In the case of the four left-most distributions in each panel, we use a lag separation of $\Delta s = 5\,h^{-1}\,\mathrm{Mpc}$, where we adapt a bin width for $\xi_1 \dots \xi_8$ of $1\,h^{-1}\,\mathrm{Mpc}$ for the first and second distribution, and a bin width of $2\,h^{-1}\,\mathrm{Mpc}$ for the third and forth one. In both cases, we separately use a narrow

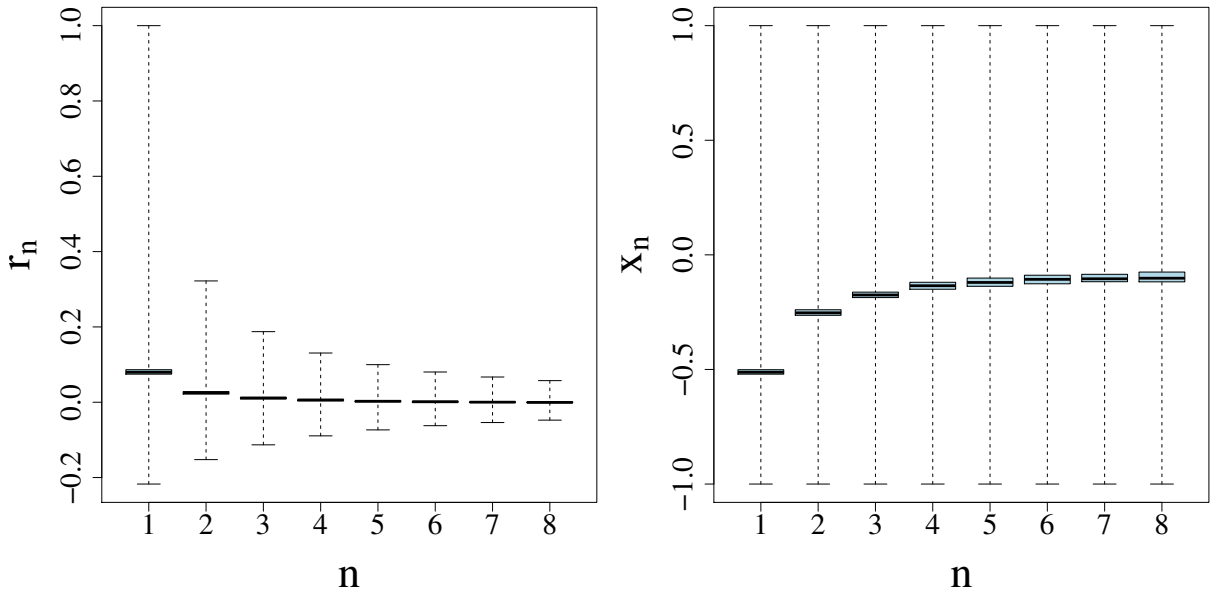


Figure 6.9: Box-and-whisker plots for $r_n$ and $x_n$, where each data point shows the upper and lower quartile (edges of the box), median (line inside the box), and mean upper and lower boundaries (whiskers) for the 125 realizations measured from the simulation subcubes at the same lags as before, using the LS estimator.

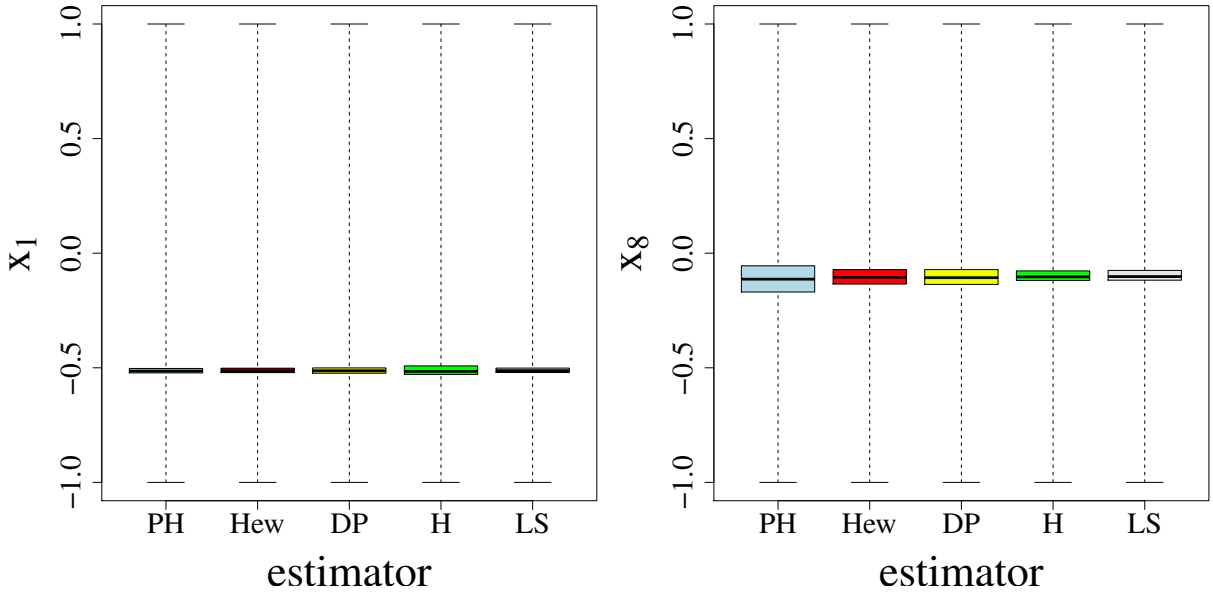


Figure 6.10: Box-and-whisker plots (see previous figure caption) of the transformed correlation coefficients at highest and lowest lag, using the five estimators listed in Table 2.2.

and a broad bin width for the measurement of $\xi_0$ (also 1 and 2 $h^{-1}$ Mpc). The figure illustrates that the width of the distributions of $x_n$ is mainly determined by the $\xi_0$-bin size, whereas the width of the bins for $\xi_n$ at lags $n > 0$ barely has any influence. Even more noteworthy is the fact that the width of the $\xi_0$-bin also shifts the $x_1$-distributions significantly – in particular, this shift is larger than compared to a case where we measure $\xi_n$ at different lags altogether, as

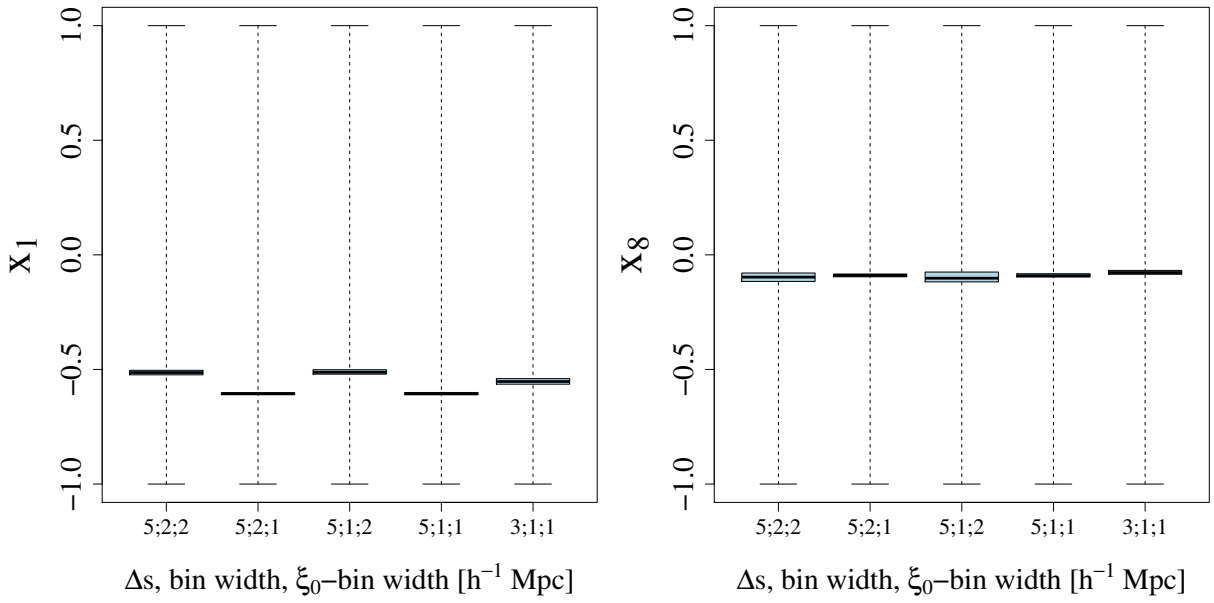


Figure 6.11: Box-and-whisker plots of the transformed correlation coefficients at highest and lowest lag for varying lag separation and bin width. The triple labeling each distribution gives the lag separation $\Delta s$, the bin width of pair separations at non-zero lag (i.e. for $\xi_1 \dots \xi_8$), and the width of the $\xi_0$-bin. For example for the second case shown in each panel, we measure $\xi_0$ from halo pairs with separations from 0 to 1 $h^{-1}$ Mpc, $\xi_1$ from those with separations from 4 to 6 $h^{-1}$ Mpc, $\xi_2$ from 9 to 11 $h^{-1}$ Mpc, and so on.

illustrated for a lag separation of $\Delta s = 3\ h^{-1}$ Mpc in the fifth distribution shown in the figure. In this context, it is important to stress that the problem of how to measure $\xi_0$ in practice is well-known, since in most applications, it is difficult to measure $\xi$ at very low lags. As we have shown, however, this poses a particularly hard challenge when analyzing measured correlation functions in a quasi-Gaussian framework, since here, the exact determination of $\xi_0$ is vital – the auto-correlation function enters everywhere, since one would always transform $\xi$ to $y$ (or at least to $r$) for an analysis involving the constraints.

In summary, all correlation functions we measured from the Millennium Simulation are quite far away from the edge of the allowed region, at least for the several choices of parameters presented in this section. In particular, the fact that every single realization of correlation functions obeys the constraints is a clear sign for the validity of the numerous choices made in our analysis – for example, using very small random catalog sizes does indeed yield single realizations outside the allowed region. On a related note, allowing for halo pairs to extend too closely to the edge of the subcubes (for examples by using a very high number of subcubes or by choosing lags such that the $\xi_8$-bin extends all the way to the border of the subcubes) again does produce realizations of the correlation function that lie outside the allowed region, even for sufficiently large random catalogs. Finally, we have shown the importance of a precise determination of $\xi_0$, which is crucial when computing the bounds on correlation functions – this can pose a challenge when analyzing correlation functions obtained from actual data in a quasi-Gaussian framework.

## 6.5 Quality of the Gaussian approximation in $\xi$ and $y$-space

In this final section, we study some properties of the PDF of the correlation functions measured from the Millennium Simulation – in particular, we transform $\xi$ to $y$ as explained in Sect. 4.3 and test the Gaussianity of the distributions in $y$ and $\xi$. This way, it is possible to investigate to which degree the transformation $\xi \to y$ does indeed Gaussianize the correlation function, which is a central ingredient for the quasi-Gaussian approach presented in this work.

While it would be preferable to directly assess the quality of the quasi-Gaussian approximation, i.e. to check how well the quasi-Gaussian PDF agrees with $p(\xi)$ as obtained from the Millennium Simulation, computing the quasi-Gaussian PDF requires measuring the underlying power spectrum, which is out of scope for this work. However, as explained in Sect. 5, in a real-life application, one would in any case transform the measured correlation function to $y$-space in order to perform a Bayesian analysis, and thus, the Gaussianity of $p(y)$ is pivotal. Nonetheless, knowledge about the underlying power spectrum would still be required in order to make use of the analytically known $p(\xi_0)$.

As described in Sect. 4.3.1, many tests for Gaussianity exist, and we focus on the calculation of moments – in particular, we compute the skewness and kurtosis, which are defined in such a way that they are zero for a Gaussian distribution. The definitions for the multivariate case are given in Eq. (4.5) and (4.6); here, we also test the Gaussianity of the univariate PDFs. In the univariate case, the skewness $\gamma$ of a distribution $p(x)$ reads

$$\gamma = \left\langle \frac{(x-\mu)^3}{\sigma^3} \right\rangle \equiv \frac{m_3}{m_2^{3/2}}, \tag{6.4}$$

where $m_i = \langle (x-\mu)^i \rangle$ denotes the central $i$th-order moment. Thus, $\gamma$ is essentially the (renormalized) third-order moment, and the kurtosis

$$\kappa = \left\langle \frac{(x-\mu)^4}{\sigma^4} \right\rangle - 3 \equiv \frac{m_4}{m_2^2} - 3 \tag{6.5}$$

is closely related to the fourth-order moment.

To test the impact of the quasi-Gaussian transformation on Gaussianity, we transform the 125 realizations of the correlation function (measured for eight lags of distance $\Delta s = 5\ h^{-1}\,\mathrm{Mpc}$ with bins of width $1\ h^{-1}\,\mathrm{Mpc}$ for $\xi_0 \dots \xi_8$) to $y$ and compute skewness and kurtosis of the distributions in $\xi$- and $y$-space. Similar to our tests in Sect. 4.3.1, we also draw Gaussian samples with same mean and covariance matrix as our samples $\{y\}$, both for comparison and to account for small sample sizes. The results for the univariate distributions are plotted in Fig. 6.12 – here, we show the skewness and kurtosis of the distributions $p(\xi_0), \dots, p(\xi_8)$ and $p(y_1), \dots, p(y_8)$ as well as corresponding Gaussian samples. Evidently, the distributions in $y$ are far more Gaussian than those in $\xi$ (with exception of $p(\xi_0)$) and in particular show a kurtosis comparable to the Gaussian samples. It is worth mentioning that due to the small sample size, the moments of Gaussian samples fluctuate quite a bit – when drawing multiple samples of size 125, it is even very possible to obtain a Gaussian sample with higher skewness or kurtosis than the $y$-samples. Note that since their purpose is only a rough comparison, we do not use the same Gaussian samples for the two panels of the figure.

As stated in Sect. 4.3.1, the Gaussianity of the univariate distribution does not imply Gaussianity of the multivariate PDFs – hence, we also compute the moments of the $n$-variate distri-

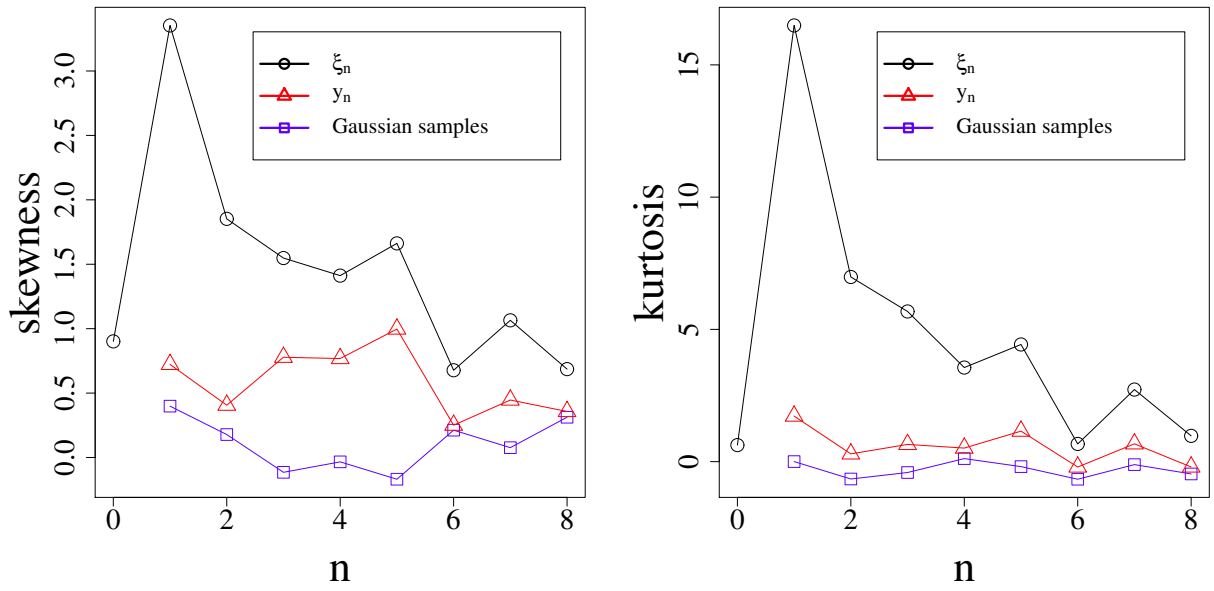


Figure 6.12: Test for the univariate Gaussianity of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation, using a lag separation of $\Delta s = 5\ h^{-1}\,\mathrm{Mpc}$ and bin widths of $1\ h^{-1}\,\mathrm{Mpc}$ for all $\xi_n$, including $\xi_0$. The data points show the univariate skewness and kurtosis of the distributions $p(\xi_n)$, $p(y_n)$, and of corresponding Gaussian samples (see text for details).

butions $p(\xi_0, \ldots, \xi_{n-1})$, $p(y_1, \ldots, y_n)$ and of corresponding multivariate Gaussian samples and plot them as functions of $n$; the results are shown in Fig. 6.13. Here the difference between the level of Gaussianity in $\xi$- and $y$-space becomes even larger, reaching about one order of magnitude in $\gamma$ and $\kappa$.

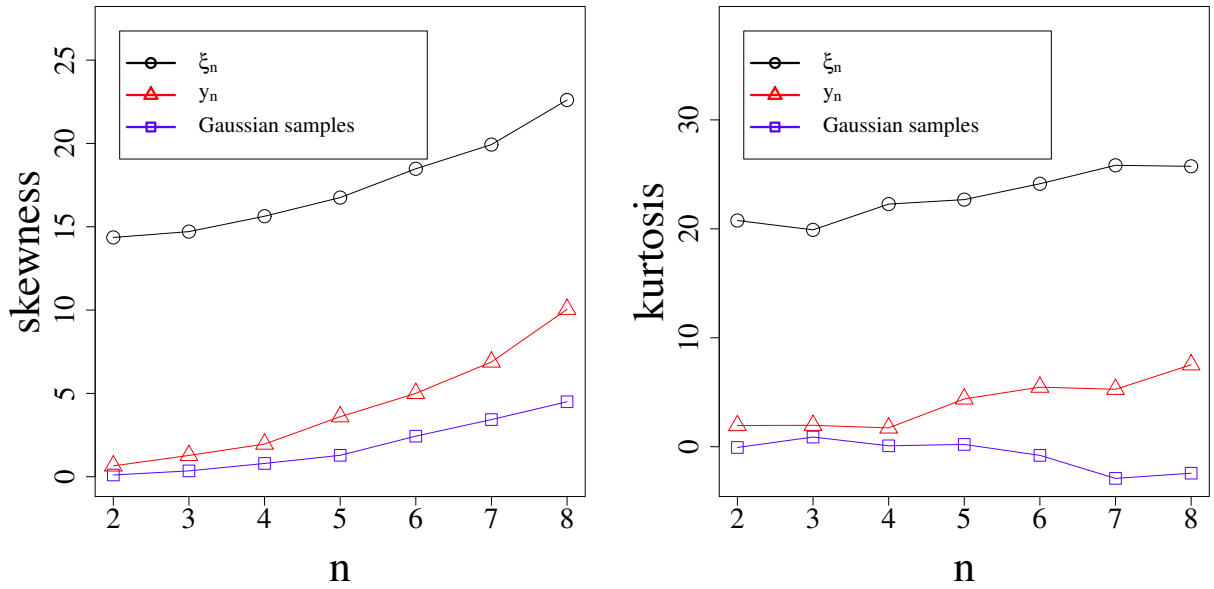


Figure 6.13: Multivariate Mardia's skewness and kurtosis of the $n$-variate distributions of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation and of corresponding Gaussian samples, using the same parameters as before (see previous figure caption and text for details).

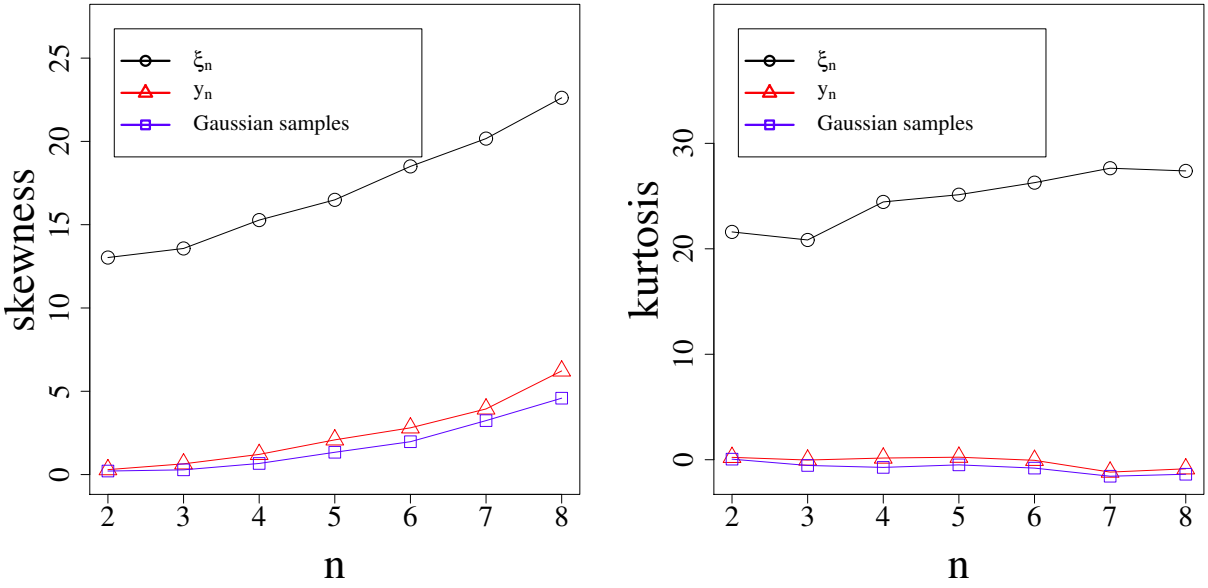


Figure 6.14: Multivariate skewness and kurtosis of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation and of corresponding Gaussian samples. In comparison to the previous figure, we adapt a broader $\xi_0$-bin, i.e. we measure the auto-correlation function from all halo pairs with pair separations from 0 to 2 $h^{-1}$ Mpc.

As we showed in the previous section, the width of the $\xi_0$-bin, i.e. the range of pair separations used to measure the auto-correlation function, has an impact on the distributions of the correlation coefficients and thus on those of the $y_n$ – hence, we vary the $\xi_0$-bin width and again study the multivariate moments of the corresponding distributions. Fig. 6.14 shows a similar plot to Fig. 6.13, however, we use a $\xi_0$-bin width of 2 instead of 1 $h^{-1}$ Mpc. As it turns out, this yields distributions in $y$-space which are almost perfectly Gaussian – their moments are hardly distinguishable from those of the corresponding Gaussian samples with same sample size. In fact, when performing the same test multiple times while drawing new Gaussian samples each time, they often yield a skewness and kurtosis slightly higher than the ones of the $y$-samples. As before, it seems that the width of the $\xi_0$-bin has a far higher impact on the results than the bin widths for $\xi_1 \dots \xi_8$ – using bins of 2 $h^{-1}$ Mpc for the higher-lag correlation functions barely influences the outcome.

As a final test, we also vary the lag separation, i.e. we measure $\xi$ for $\Delta s = 3\ h^{-1}$ Mpc, again adapting bin widths of 1 $h^{-1}$ Mpc both for $\xi_0$ and $\xi_1 \dots \xi_8$. The results for the skewness and kurtosis are shown in Fig. 6.15, plotted analogously to the previous figures – it is apparent that changing the lags does not alter our conclusion, namely that the distributions in $y$-space are far more Gaussian than those in $\xi$ and do indeed have skewness and kurtosis comparable to those of Gaussian samples of the same size. In summary, the tests shown in this section indicate the validity of the quasi-Gaussian approach independent of the specific parameters used to measure the correlation function.

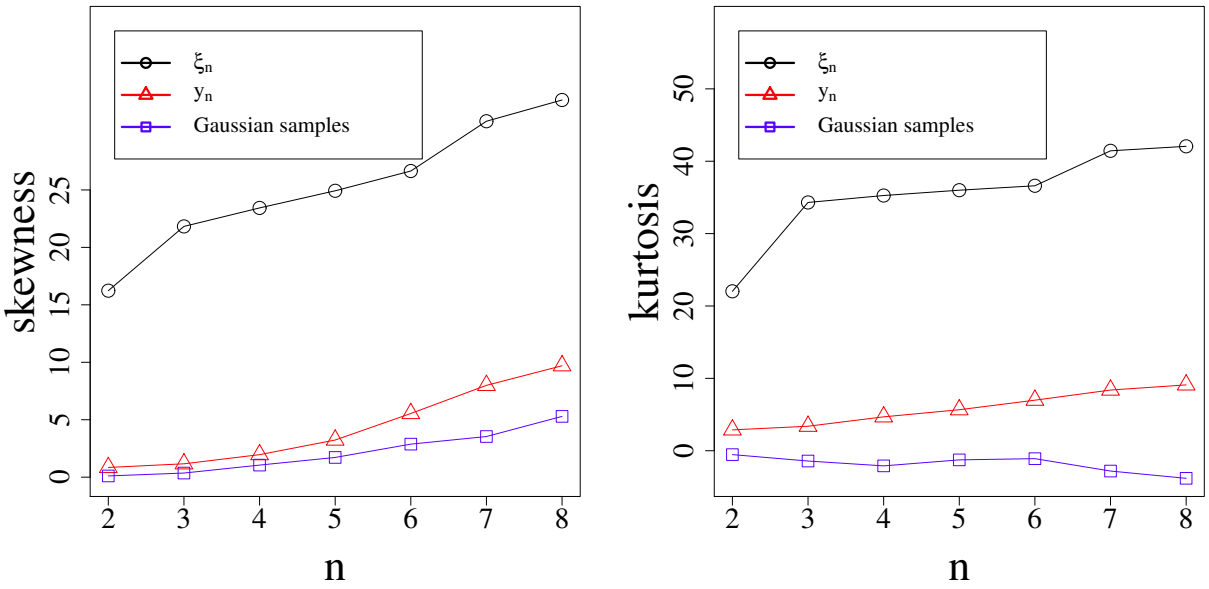


Figure 6.15: Multivariate skewness and kurtosis of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation using different lags than before, and of corresponding Gaussian samples. Here, we measure $\xi$ of lag separations $\Delta s = 3\ h^{-1}\,\mathrm{Mpc}$ with bin widths of $1\ h^{-1}\,\mathrm{Mpc}$ for $\xi_0 \ldots \xi_8$.

CHAPTER 7

# Conclusion

## 7.1 Summary

In this final chapter, we summarize the results presented in this work and give a brief outlook on potential next steps.

When analyzing measured correlation functions in the framework of Bayesian statistics, their likelihood is usually assumed to be Gaussian – however, this approximation is not necessarily accurate due to the existence of fundamental constraints on the allowed values of correlation functions, which can be derived from purely mathematical properties.

We presented a derivation of those constraints and showed how to compute them, which is possible analytically only for a one-dimensional random field. For fields of higher dimensionality, which constitute the main areas of application in astrophysics, one has to resort to numerical methods. We showed how the constraints can be calculated by constructing a convex hull in a high-dimensional space, and tested the accuracy of this method by comparing the results to the analytical constraints for the one-dimensional case.

The constraints can then be used to transform the correlation function $\xi$ to an unbounded quantity $y$, where the Gaussian approximation for the likelihood is expected to hold to a higher degree of accuracy. We showed that this is indeed the case also for multivariate distributions, and then demonstrated how to make use of this observation to construct a new approximation for the likelihood $\mathcal{L}(\xi)$: Transforming the Gaussian from $y$- to $\xi$-space yields a quasi-Gaussian likelihood which agrees very well with simulated correlation functions. We presented some attempts to improve the calculation of the quasi-Gaussian PDF, in particular by obtaining the $\xi_0$-dependent mean and covariance matrix analytically. Additionally, we compared our approach to other well-established techniques of Gaussianizing, namely to Box-Cox and copula methods, and showed that the quasi-Gaussian method is superior.

As a next step, we tested the performance of the quasi-Gaussian likelihood by actually using it in a Bayesian analysis, though we limited our studies to a toy-model analysis: Namely, we simulated a sample of correlation function as data, and used both the quasi-Gaussian and the classical Gaussian likelihood to constrain the parameters of the underlying power spectrum. While both methods manage to reproduce the input parameters quite well, the resulting posterior distributions have different shapes – this is a strong hint that the quasi-Gaussian likelihood potentially impacts the results of cosmological parameter estimation from measured correlation functions.

Finally, we shifted our focus closer to real-life applications, namely by measuring correlation functions from the Millennium Simulation. The large volume of the simulation allowed us to compute $\xi$ in a high number of subcubes, thus obtaining a sample of correlation functions and facilitating a statistical analysis. We discussed challenges in the measurement of $\xi$, for example the choice of estimator, random catalog size, and lag separation, as well as the question of how to overcome the integral constraint. By computing the bounds for each realization of $\xi$ using the numerical methods previously explained, we showed that they are very clearly obeyed, independent of the choice of lags.

Lastly, we also tested $\xi$ and the transformed quasi-Gaussian quantity $y$ for Gaussianity by computing moments of the distributions – the fact that the distributions in $y$-space are far more Gaussian than the original correlation functions and in some cases even yield a skewness and kurtosis comparable to Gaussian samples gives further support to the validity of the quasi-Gaussian approach.

## 7.2 Outlook

While we established a solid groundwork for the quasi-Gaussian approach in this thesis, many possible next steps come to mind: On the one hand, there is still room for improvement on the theoretical side, since some attempts we presented in this work did not give usable results, such as the analytical computation of the mean and covariance matrix for the Gaussian distribution in $y$-space and in particular their $\xi_0$-dependences. While this is a minor caveat, there are more crucial challenges, in particular the expensive computation of the constraints for two- and three-dimensional fields. This area merits some attention, since improving the performance would allow for a higher accuracy of the constraints – most importantly, however, getting rid of the current limitation to eight lags seems desirable, since modern astronomical observations usually measure $\xi$ at far more lags. Unfortunately, this is not only a matter of computation costs, although they do scale strongly with the number of lags (recall that the determination of the constraints requires the calculation of a convex hull in a space of dimensionality equal to the number of lags) – due to limitations in the qhull algorithm used in this work, computing convex hulls of dimensionality larger than eight would require an all-new implementation. On a less technical note, in this context it might also be interesting to generalize the constraints to non-equidistant lags.

Most of the potential next steps mentioned so far are testable on simulations, and the studies of the correlation functions measured from the Millennium Simulation presented in this work are worth revisiting: In particular, a Bayesian analysis using the realizations of $\xi$ measured from the simulation (which we omitted since it would require a measurement of the power spectrum) would be a useful way to study the performance of the quasi-Gaussian likelihood. However, the most crucial advancement would be the application of our methods to real data. Of course, this poses many new challenges – aside from the current limitation to only eight (and equidistant) lags, numerous other effects have to be taken into account: When making use of $N$-body simulations as done in this work, the absolute halo positions in three-dimensional space are known, but in an actual survey, one instead measures redshifts. Thus, one has to deal with redshift-space distortions, which can impose different constraints on the correlation function measured along and perpendicular to the line-of-sight. In other cases, one might have to deal with different coordinate systems (e.g. spherical ones), which would also require further work in order to compute the constraints.

An additional long-term goal might be to also move away from correlation functions and consider the power spectrum instead: While from a theoretical point of view, constraints on the power spectrum seem simple (i.e. it has to be non-negative), in practice, one has to use an estimator, which is always effected by window functions. Thus, similar constraints must also exist for power spectrum estimators, but these have yet to be derived.

Nonetheless, as a final remark, the constraints on correlation functions of three-dimensional random fields are in principle treatable despite open challenges and room for improvements – thus, this work opens up a vast field of applications where Gaussian likelihoods for $\xi$ have previously been used.

APPENDIX A

# Acknowledgments

There are a lot of people I would like to thank for supporting me over the course of my PhD. The top spot on this long list definitely belongs to Peter Schneider, who gave me the opportunity to do further research on this project after my diploma thesis – while he granted me the freedom to work on my own, he still always made time for me (even in the last weeks, after certain life-changing good news landed in his world), and provided input and ideas. I also thank my co-supervisor Cristiano Porciani, who gave important hints for the final steps in the project, as well as Ian Brock and Andreas Bartels for agreeing to be part of my Promotionskommission.

My thanks also go to the other collaborators in this project, namely to my former office mate David Keitel for all his work on the (even more) mathematical parts of the project and his code for computing the analytical uni- and bivariate distributions, to Randolph Roeseler for his code for the numerical computation of the constraints, and to Benjamin Joachimi, who kindly provided his Box-Cox code.

In my time at AIfA, I had the chance to pursue various activities (scientific and otherwise, some of which even involved an actual telescope), who probably had at least as much impact on me as my actual research – so naturally, there are many institute members (both current and past ones) I would like to thank. While I cannot list all of them, I will name a few and hope nobody feels left out: first, Christina Stein-Schmitz and Ellen Vasters, who always helped me with administrative things and without whom everyday life at AIfA would probably break down; of course, my office mates Alex and Zeinab, who provided a nice work atmosphere, and cookies whenever needed; my fellow Astroclub organizers; the public outreach seminar; Daniela Wuttke, my former office mate, fellow SPP organizer, and fellow ice hockey fan (in this context, thanks to Chris for taking over!). And since a regular work day would hardly have been complete without a match of table soccer, a big shout-out to all my kicker friends (and occasional enemies). Likewise, I thank all Bruce Night participants, and especially Bharad for Local Nerd News, too many chats and fun evenings to remember, and one of the most awesome travels of my life.

Finally, this work would never have seen the light of day without many other personal friends and family to whom I owe a great deal of gratitude: CFM, whose wisdom and power of observation accompanied me in my process of becoming a Bonner Bürger; all the DSA people; my parents who always support me in all ways possible; of course, Go-Kiwi! And most importantly Nina for, well, simply everything.

APPENDIX B

# Bibliography


C. P. Ahn, R. Alexandroff, C. Allende Prieto, F. Anders, S. F. Anderson, T. Anderton, B. H. Andrews, É. Aubourg, S. Bailey, F. A. Bastien, and et al. The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment. ApJS, 211:17, April 2014, arXiv:1307.7735 [astro-ph.IM].

R. Allison and J. Dunkley. Comparison of sampling techniques for Bayesian parameter estimation. MNRAS, 437:3918–3928, February 2014, arXiv:1308.2675 [astro-ph.IM].

R. E. Angulo and S. D. M. White. One simulation to fit them all - changing the background parameters of a cosmological N-body simulation. MNRAS, 405:143–154, June 2010, arXiv:0912.4277 [astro-ph.CO].

S. Arrenberg, H. Baer, V. Barger, L. Baudis, D. Bauer, J. Buckley, M. Cahill-Rowley, R. Cotta, A. Drlica-Wagner, J. L. Feng, S. Funk, J. Hewett, D. Hooper, A. Ismail, M. Kaplinghat, K. Kong, A. Kusenko, K. Matchev, M. McCaskey, D. McKinsey, D. Mickelson, T. Rizzo, D. Sanford, G. Shaughnessy, W. Shepherd, T. M. P. Tait, X. Tata, S. Tulin, A. M. Wijangco, M. Wood, J. Yoo, and H.-B. Yu. Dark Matter in the Coming Decade: Complementary Paths to Discovery and Beyond. *ArXiv e-prints*, October 2013, arXiv:1310.8621 [hep-ph].

C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software, 22(4):469–483, 1996.

F. Bernardeau, S. Colombi, E. Gaztañaga, and R. Scoccimarro. Large-scale structure of the Universe and cosmological perturbation theory. Phys. Rep., 367:1–248, September 2002, astro-ph/0112551.

BICEP2 Collaboration, P. A. R. Ade, R. W. Aikin, D. Barkats, S. J. Benton, C. A. Bischoff, J. J. Bock, J. A. Brevik, I. Buder, E. Bullock, C. D. Dowell, L. Duband, J. P. Filippini, S. Fliescher, S. R. Golwala, M. Halpern, M. Hasselfield, S. R. Hildebrandt, G. C. Hilton, V. V. Hristov, K. D. Irwin, K. S. Karkare, J. P. Kaufman, B. G. Keating, S. A. Kernasovskiy, J. M. Kovac, C. L. Kuo, E. M. Leitch, M. Lueker, P. Mason, C. B. Netterfield, H. T. Nguyen, R. O'Brient, R. W. Ogburn, IV, A. Orlando, C. Pryke, C. D. Reintsema, S. Richter, R. Schwarz, C. D. Sheehy, Z. K. Staniszewski, R. V. Sudiwala, G. P. Teply, J. E. Tolan, A. D. Turner, A. G.

Vieregg, C. L. Wong, and K. W. Yoon. BICEP2 I: Detection Of B-mode Polarization at Degree Angular Scales. *ArXiv e-prints*, March 2014, arXiv:1403.3985 [astro-ph.CO].

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

A. Boyarsky, O. Ruchayskiy, D. Iakubovskyi, and J. Franse. An unidentified line in X-ray spectra of the Andromeda galaxy and Perseus galaxy cluster. *ArXiv e-prints*, February 2014, arXiv:1402.4119 [astro-ph.CO].

E. Cameron and A. N. Pettitt. Approximate Bayesian Computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. MNRAS, 425:44–65, September 2012, arXiv:1202.1426 [astro-ph.IM].

J. Carron. On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices. A&A, 551:A88, March 2013, arXiv:1204.4724 [astro-ph.CO].

S. Crandall and B. Ratra. Median statistics cosmological parameter values. *ArXiv e-prints*, November 2013, arXiv:1311.0840 [astro-ph.CO].

R. A. C. Croft and M. Dailey. On the measurement of cosmological parameters. *Submitted to* MNRAS, December 2011, arXiv:1112.3108 [astro-ph.CO].

N. Dalal, O. Doré, D. Huterer, and A. Shirokov. Imprints of primordial non-Gaussianities on large-scale structure: Scale-dependent bias and abundance of virialized objects. Phys. Rev. D, 77(12):123514, June 2008, arXiv:0710.4560.

M. Davis and P. J. E. Peebles. A survey of galaxy redshifts. V - The two-point position and velocity correlations. ApJ, 267:465–482, April 1983.

T. M. Davis. Cosmological constraints on dark energy. *ArXiv e-prints*, April 2014, arXiv:1404.7266.

S. Dorn, N. Oppermann, and T. A. Enßlin. Diagnostics for insufficiencies of posterior calculations in Bayesian signal inference. Phys. Rev. E, 88:053303, November 2013, arXiv:1307.3889 [astro-ph.IM].

S. Dorn, N. Oppermann, R. Khatri, M. Selig, and T. A. Enßlin. Fast and precise way to calculate the posterior for the local non-Gaussianity parameter $f_{\rm nl}$ from cosmic microwave background observations. Phys. Rev. D, 88:103516, November 2013, arXiv:1307.3884 [astro-ph.CO].

W. L. Freedman, B. F. Madore, B. K. Gibson, L. Ferrarese, D. D. Kelson, S. Sakai, J. R. Mould, R. C. Kennicutt, Jr., H. C. Ford, J. A. Graham, J. P. Huchra, S. M. G. Hughes, G. D. Illingworth, L. M. Macri, and P. B. Stetson. Final Results from the Hubble Space Telescope Key Project to Measure the Hubble Constant. ApJ, 553:47–72, May 2001, astro-ph/0012376.

M. Frigo and S. G. Johnson. *FFTW: Fastest Fourier Transform in the West*, 2003. http://www.fftw.org.

M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. *GNU Scientific Library Reference Manual – Third Edition*. Network Theory Ltd., 2009. http://www.gnu.org/software/gsl/.

A. J. S. Hamilton. Toward Better Ways to Measure the Galaxy Correlation Function. ApJ, 417:19–35, November 1993.

J. Hartlap, T. Schrabback, P. Simon, and P. Schneider. The non-Gaussianity of the cosmic shear likelihood – or: How odd is the Chandra Deep Field South? A&A, 504:689–703, September 2009, arXiv:0901.3269 [astro-ph.CO].

N. Henze. Invariant tests for multivariate normality: a critical review. *Statistical Papers*, 43(4):467–506, October 2002.

P. C. Hewett. The estimation of galaxy angular correlation functions. MNRAS, 201:867–883, December 1982.

G. Hinshaw, D. Larson, E. Komatsu, D. N. Spergel, C. L. Bennett, J. Dunkley, M. R. Nolta, M. Halpern, R. S. Hill, N. Odegard, L. Page, K. M. Smith, J. L. Weiland, B. Gold, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, E. Wollack, and E. L. Wright. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. ApJS, 208:19, October 2013, arXiv:1212.5226 [astro-ph.CO].

R. V. Hogg, J. W. McKean, and A. T. Craig. *Introduction to Mathematical Statistics, Sixth Edition*. Pearson Education, Upper Saddle River, NJ, USA, 2005.

H. Jeffreys. *Theory of probability, 3rd edition, Oxford Classics series (reprinted 1998)*. Oxford University Press, Oxford, UK, 1961.

B. Joachimi, A. N. Taylor, and A. Kiessling. Cosmological information in Gaussianised weak lensing signals. MNRAS, 418:145–169, April 2011, arXiv:1104.1399 [astro-ph.CO].

S. Joe and F. Y. Kuo. Constructing sobol sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30(5):2635–2654, 2008.

E. Jones Baxter and E. Rozo. A Maximum Likelihood Approach to Estimating Correlation Functions. ApJ, 779:62, December 2013, arXiv:1305.4613 [astro-ph.CO].

D. Keitel and P. Schneider. Constrained probability distributions of correlation functions. A&A, 534:A76, October 2011, arXiv:1105.3672 [astro-ph.CO].

M. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 1: Distribution Theory. Macmillan, New York, NY, USA, 4th edition, 1977.

M. Kerscher, I. Szapudi, and A. S. Szalay. A Comparison of Estimators for the Two-Point Correlation Function. ApJ, 535:L13–L16, May 2000, astro-ph/9912088.

M. Kilbinger, D. Wraith, C. P. Robert, K. Benabed, O. Cappé, J.-F. Cardoso, G. Fort, S. Prunet, and F. R. Bouchet. Bayesian model comparison in cosmology with Population Monte Carlo. MNRAS, 405:2381–2390, July 2010, arXiv:0912.1614 [astro-ph.CO].

E. Komatsu, B. D. Wandelt, D. N. Spergel, A. J. Banday, and K. M. Górski. Measurement of the Cosmic Microwave Background Bispectrum on the COBE DMR Sky Maps. ApJ, 566:19–29, February 2002, astro-ph/0107605.

P. Kroupa, M. Pawlowski, and M. Milgrom. The Failures of the Standard Model of Cosmology Require a New Paradigm. *International Journal of Modern Physics D*, 21:30003, December 2012, arXiv:1301.3907 [astro-ph.CO].

M. Kuhlen, M. Vogelsberger, and R. Angulo. Numerical simulations of the dark universe: State of the art and the next decade. *Physics of the Dark Universe*, 1:50–93, November 2012, arXiv:1209.5745 [astro-ph.CO].

J. Kurchan. Elementary constraints on autocorrelation function scalings. Phys. Rev. E, 66(1):017101, July 2002, cond-mat/0110628.

A. Labatie, J.-L. Starck, and M. Lachièze-Rey. Detecting Baryon Acoustic Oscillations. ApJ, 746:172, February 2012, arXiv:1112.0980 [astro-ph.CO].

A. Labatie, J.-L. Starck, and M. Lachièze-Rey. Effect of model-dependent covariance matrix for studying Baryon Acoustic Oscillations. ApJ, 760:97, 2012, arXiv:1210.0878 [astro-ph.CO].

O. Lahav and A. R. Liddle. The Cosmological Parameters 2014. *ArXiv e-prints*, January 2014, arXiv:1401.1389 [astro-ph.CO].

S. D. Landy and A. S. Szalay. Bias and variance of angular correlation functions. ApJ, 412:64–71, July 1993.

F. Leclercq, A. Pisani, and B. D. Wandelt. Cosmology: from theory to data, from data to theory. *ArXiv e-prints*, March 2014, arXiv:1403.1260 [astro-ph.CO].

G. Lemson and the Virgo Consortium. Halo and Galaxy Formation Histories from the Millennium Simulation: Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the LambdaCDM cosmogony. *ArXiv e-prints*, August 2006, arXiv:astro-ph/0608019 [astro-ph].

Z. Li, P. Wu, H. Yu, and Z.-H. Zhu. A possible resolution of tension between Planck and Type Ia supernova observations. *Science China Physics, Mechanics and Astronomy*, 57(2):381–386, 2014, arXiv:1311.3467 [astro-ph.CO].

A. R. Liddle. Statistical Methods for Cosmological Parameter Selection and Estimation. *Annual Review of Nuclear and Particle Science*, 59:95–114, November 2009, arXiv:0903.4210 [hep-th].

D. MacKenzie and T. Spears. 'The formula that killed Wall Street': The Gaussian copula and modelling practices in investment banking. *Social Studies of Science*, 44(3):393–417, 2014.

K. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

K. Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā. Series B. Methodological*, 36:115–128, 1974.

J. Martin, C. Ringeval, and V. Vennin. Encyclopaedia Inflationaris. *ArXiv e-prints*, March 2013, arXiv:1303.3787 [astro-ph.CO].

A. Mead and J. Peacock. Remapping dark matter halo catalogues between cosmological simulations. MNRAS, 440:1233–1247, May 2014, arXiv:1308.5183 [astro-ph.CO].

J. F. Navarro, C. S. Frenk, and S. D. M. White. The Structure of Cold Dark Matter Halos. ApJ, 462:563, May 1996, astro-ph/9508025.

S. Nesseris and J. García-Bellido. Is the Jeffreys' scale a reliable tool for Bayesian model comparison in cosmology? JCAP, 8:36, August 2013, arXiv:1210.7652 [astro-ph.CO].

D. Parkinson and A. R. Liddle. Bayesian Model Averaging in Astrophysics: A Review. *Statistical Analysis and Data Mining*, 6(1):pp. 3–14, February 2013, arXiv:1302.1721 [astro-ph.IM].

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

P. J. E. Peebles and M. G. Hauser. Statistical Analysis of Catalogs of Extragalactic Objects. III. The Shane-Wirtanen and Zwicky Catalogs. ApJS, 28:19, November 1974.

P. J. E. Peebles. Statistical analysis of catalogs of extragalactic objects. VI - The galaxy distribution in the Jagellonian field. ApJ, 196:647–651, March 1975.

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. I. Overview of products and scientific results. A&A, 571:A1, November 2014, arXiv:1303.5062 [astro-ph.CO].

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XVI. Cosmological parameters. A&A, 571:A16, November 2014, arXiv:1303.5076 [astro-ph.CO].

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 Results. XXIV. Constraints on primordial non-Gaussianity. A&A, 571:A24, November 2014, arXiv:1303.5084 [astro-ph.CO].

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0, http://www.R-project.org.

H. P. Robertson. Kinematics and World-Structure. ApJ, 82:284, November 1935.

R. Röseler. *The probability density of correlation functions of random fields in multiple dimensions*. Master thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, April 2013.

M. Sato, K. Ichiki, and T. T. Takeuchi. Precise Estimation of Cosmological Parameters Using a More Accurate Likelihood Function. Phys. Rev. Lett., 105(25):251301, December 2010, arXiv:1011.4996 [astro-ph.CO].

M. Sato, K. Ichiki, and T. T. Takeuchi. Copula cosmology: Constructing a likelihood function. Phys. Rev. D, 83(2):023501, January 2011, arXiv:1011.4997 [astro-ph.CO].

R. J. Scherrer, A. A. Berlind, Q. Mao, and C. K. McBride. From Finance to Cosmology: The Copula of Large-Scale Structure. ApJ, 708:L9–L13, January 2010, arXiv:0909.5187 [astro-ph.CO].

P. Schneider and J. Hartlap. Constrained correlation functions. A&A, 504:705–717, September 2009, arXiv:0905.0577 [astro-ph.CO].

P. Schneider and D. Sluse. Mass-sheet degeneracy, power-law models and external convergence: Impact on the determination of the Hubble constant from gravitational lensing. A&A, 559:A37, November 2013, arXiv:1306.0901 [astro-ph.CO].

P. Schneider. *Extragalactic Astronomy and Cosmology: An Introduction*. Springer, Heidelberg, Germany, 2006.

U. Seljak and E. Bertschinger. Maximum-Likelihood Analysis of the COBE Angular Correlation Function. ApJ, 417:L9, November 1993, arXiv:astro-ph/9309003 [astro-ph.CO].

V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. Nature, 435:629–636, June 2005, arXiv:astro-ph/0504097 [astro-ph].

V. Springel. The cosmological simulation code GADGET-2. MNRAS, 364:1105–1134, December 2005, arXiv:astro-ph/0505010 [astro-ph].

L. Sun, Q. Wang, and H. Zhan. Likelihood of the Power Spectrum in Cosmological Parameter Estimation. ApJ, 777(1):75, November 2013, arXiv:1306.5064 [astro-ph.CO].

S. H. Suyu, M. W. Auger, S. Hilbert, P. J. Marshall, M. Tewes, T. Treu, C. D. Fassnacht, L. V. E. Koopmans, D. Sluse, R. D. Blandford, F. Courbin, and G. Meylan. Two Accurate Time-delay Distances from Strong Lensing: Implications for Cosmology. ApJ, 766:70, April 2013, arXiv:1208.6010 [astro-ph.CO].

R. Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49:71–104, March 2008, arXiv:0803.4089.

M. Vargas-Magaña, J. E. Bautista, J.-C. Hamilton, N. G. Busca, É. Aubourg, A. Labatie, J.-M. Le Goff, S. Escoffier, M. Manera, C. K. McBride, D. P. Schneider, and C. N. A. Willmer. An optimized correlation function estimator for galaxy surveys. A&A, 554:A131, June 2013, arXiv:1211.6211 [astro-ph.CO].

M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. F. Snyder, D. Nelson, and L. Hernquist. Introducing the Illustris Project: Simulating the coevolution of dark and visible matter in the Universe. MNRAS, 444(2):1518–1547, August 2014, arXiv:1405.2921.

A. G. Walker. On Milne's Theory of World-Structure. Proc. Lond. Math. Soc., s2-42(1):90–127, 1937.

M. D. Weinberg. Computational statistics using the Bayesian Inference Engine. MNRAS, 434:1736–1755, September 2013, arXiv:1203.3816 [astro-ph.IM].

A. Weyant, C. Schafer, and W. M. Wood-Vasey. Likelihood-free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty. ApJ, 764:116, February 2013, arXiv:1206.2563 [astro-ph.CO].

P. Wilking and P. Schneider. A quasi-Gaussian approximation for the probability distribution of correlation functions. A&A, 556:A70, August 2013, arXiv:1304.4781 [astro-ph.CO].

T. Williams, C. Kelley, et al. *Gnuplot Version 4.2*, 2008. http://www.gnuplot.info.

Wolfram Research, Inc. *Mathematica Edition: Version 7.0.* Champaign, IL, USA, 2008. http://www.wolfram.com/mathematica.

M. Wyman, D. H. Rudd, R. A. Vanderveld, and W. Hu. Neutrinos Help Reconcile Planck Measurements with the Local Universe. *Physical Review Letters*, 112(5):051302, February 2014, arXiv:1307.7715 [astro-ph.CO].

APPENDIX C

# List of figures


2.1 Different cosmological models classified by their expansion behavior and spatial geometry (adapted from Peacock 1999) . . . 10

3.1 Volume fraction of the correlation coefficients obeying the constraints and linear dimension of the allowed region . . . 23
3.2 Two examples of the curve $\boldsymbol{c}(\lambda)$ for a two-dimensional random field . . . 25
3.3 Direct comparison of the analytically and numerically obtained bounds, averaged over 500 realizations . . . 26
3.4 The difference between the analytically and numerically obtained bounds, averaged over 500 realizations . . . 27
3.5 Illustration of the effect of the convex hull sampling . . . 28
3.6 Direct comparison of the $y_n$, computed using the analytically and numerically obtained bounds, averaged over 500 realizations . . . 29
3.7 The difference between the $y_n$ computed using the analytically and the numerically obtained bounds, averaged over 500 realizations . . . 29

4.1 Iso-probability contours of $p(\xi_3, \xi_6)$ and $p(y_3, y_6)$ for 400 000 realizations of a field with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$, as well as the best fitting Gaussian approximations . . . 33
4.2 $p(\xi_1, \xi_2)$ for a field with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$, as well as the transformed Gaussian approximation from $y$-space . . . 35
4.3 Mardia's skewness and kurtosis of $n$-variate $\{\xi\}$- and $\{y\}$-samples for a Gaussian power spectrum with $Lk_0 = 80$ . . . 36
4.4 The mean of $\xi_n$ for different $n$ as function of $\xi_0$, determined from simulations and analytically to different orders . . . 39
4.5 Different elements of the covariance matrix $C(\{\xi_n\})$, determined from simulations and analytically to different orders . . . 41
4.6 Results of the test setup used to recover the covariance matrix of an 8-variate Gaussian distribution . . . 42
4.7 $p(\xi_1, \xi_2)$ and $p(\xi_3, \xi_6)$ for a field with $N = 32$ grid points and a Gaussian power spectrum with $Lk_0 = 80$, as well as the approximation obtained from a Gaussian copula . . . 44

5.1 The components of the quasi-Gaussian likelihood, namely the Gaussian in $y$-space and the $p(\xi_0)$-term . . . 50
5.2 Posterior probability for the power spectrum parameters $(A, k_0)$ using the Gaussian and the quasi-Gaussian likelihood and flat priors . . . 51
5.3 Marginalized posterior probabilities for the power spectrum parameters $A$ and $k_0$ from the Gaussian likelihood and the quasi-Gaussian analysis . . . 51

6.1 The DM density obtained from the Millennium Simulation at different scales (from Springel *et al.* 2005) . . . 55
6.2 Example of an SQL query to access the output of the Millennium Run . . . 56
6.3 The correlation functions from 1000 subcubes of the Millennium Simulation, computed using different random catalog sizes . . . 57
6.4 Histograms of the distributions of overdensities in the subcubes of the simulation volume for different numbers of subcubes . . . 58
6.5 The mean correlation function from the subcubes of the Millennium Simulation, computed using different numbers of subcubes . . . 60
6.6 The correlation function from the Millennium Simulation as a function of the pair separation $s$, measured with different estimators . . . 61
6.7 Scatter- and contour-plot of the correlation coefficients $r_1$ and $r_2$ measured from the halo catalogs in a varying number of subcubes of the Millennium Simulation, as well as the upper and lower bounds . . . 62
6.8 Scatter-plots of the realizations of the transformed correlation coefficients $x_n$ at lowest and highest lag measured from the 125 subcubes of the Millennium Simulation . . . 63
6.9 Box-and-whisker plots for $r_n$ and $x_n$ illustrating the width and position of the distributions within the regions allowed by the constraints . . . 64
6.10 Box-and-whisker plots of the transformed correlation coefficients at highest and lowest lag, using different estimators . . . 64
6.11 Box-and-whisker plots of the transformed correlation coefficients at highest and lowest lag for varying lag separation and bin width . . . 65
6.12 Test for the univariate Gaussianity of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation . . . 67
6.13 Multivariate Mardia's skewness and kurtosis of the $n$-variate distributions of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation and of corresponding Gaussian samples . . . 67
6.14 Multivariate skewness and kurtosis of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation and of corresponding Gaussian samples, using a broader $\xi_0$-bin than before . . . 68
6.15 Multivariate skewness and kurtosis of the $\{\xi\}$-and $\{y\}$-samples obtained from the Millennium Simulation using different lags than before, and of corresponding Gaussian samples . . . 69

APPENDIX D

# List of tables

2.1 Cosmological parameters for the best-fit ΛCDM model (adapted from Planck Collaboration *et al.* 2014a,b) . . . . . 11
2.2 Different correlation function estimators (adapted from Vargas-Magaña *et al.* 2013) 15

6.1 Cosmological parameters and simulation parameters used for the Millennium Simulation . . . . . 55

APPENDIX E

# List of acronyms

**ABC** Approximate Bayesian Computation.

**AGN** active galactic nucleus.

**BAO** baryon acoustic oscillations.

**BBN** Big Bang nucleosynthesis.

**BIE** Bayesian Inference Engine.

**CDF** cumulative distribution function.

**CDM** Cold Dark Matter.

**CMB** cosmic microwave background.

**COBE** Cosmic Background Explorer.

**DE** Dark Energy.

**DIP** Diagnostics for Insufficiencies of Posterior calculations.

**DM** Dark Matter.

**FFT** fast Fourier transform.

**FLRW** Friedmann-Lemaître-Robertson-Walker.

**GAVO** German Astrophysical Virtual Observatory.

**GR** General Relativity.

**GSL** GNU Scientific Library.

**GUT** Grand Unified Theory.

**IGM** intergalactic medium.

**KL** Kullback-Leibler.

**LHC** Large Hadron Collider.

**LS** Landy-Szalay.

**LSS** large-scale structure.

**MAP** maximum a posteriori.

**MCMC** Markov chain Monte Carlo.

**MOND** Modified Newtonian Dynamics.

**NFW** Navarro–Frenk–White.

**PDF** probability distribution function.

**PH** Peebles and Hauser.

**QSO** quasi-stellar object.

**RMS** root mean square.

**SDSS** Sloan Digital Sky Survey.

**SN** supernova.

**SQL** Structured Query Language.

**VM** Vargas-Magaña.

**WIMP** weakly interacting massive particle.

**WMAP** Wilkinson Microwave Anisotropy Probe.