Uncertainty in 2-point correlation function estimators

Antoine Labatie[a], Jean-Luc Starck[a], Marc Lachièze-Rey[b], Pablo Arnalte-Mur[c,d]

[a]*Laboratoire AIM (UMR 7158), CEA/DSM-CNRS-Université Paris Diderot, IRFU, SEDI- SAP, Service dAstrophysique, Centre de Saclay, F-91191 Gif-Sur-Yvette cedex, France.*
[b]*Astroparticule et Cosmologie (APC), CNRS-UMR 7164, Université Paris 7 Denis Diderot 10, rue Alice Domon et Léonie Duquet F-75205 Paris Cedex 13, France.*
[c]*Observatori Astronòmic, Universitat de València, Apartat de Correus 22085, E-46071 València, Spain.*
[d]*Departament d'Astronomia i Astrofísica, Universitat de València, 46100-Burjassot, València, Spain.*

Abstract

We study the uncertainty in different two-point correlation function estimators in currently available galaxy surveys. This is motivated by the active subject of using the BAO feature in the correlation function as a tool to constrain cosmological parameters, which requires a fine analysis of the statistical significance.

We discuss how estimators are affected by both the uncertainty on the mean density $\bar{n}$ and the integral constraint $\frac{1}{V^2}\int_{V^2}\hat{\xi}(r)d^3r = 0$ which necessarily causes a bias. We quantify both effects for currently available galaxy samples using simulated mock SDSS catalogues that follow a lognormal model, with a ΛCDM correlation function and similar properties as the samples (number density, mean redshift for the ΛCDM correlation function, survey geometry, mass-luminosity bias). We look at the variance and bias of the different estimators in order to compare their quality and know if they are affected by the integral constraint.

Introduction

The correlation function is the most popular tool for analyzing the distribution of galaxies [16]. Any model, like in particular the standard ΛCDM, predicts a certain shape for $\xi(r)$ with a dependence on the cosmic parameters. Given this dependence we can constrain cosmic parameters when measuring the correlation function with enough precision.

In this context a very active field is the study of the Baryon Acoustic Oscillations (BAO) which should imprint the matter correlation function. It is a relic of the sound waves in the early Universe when baryon and photons where coupled in a relativistic plasma before recombination which caused the wave propagation to end ([3]). It can be seen as a small peak in the correlation function at a scale r_s corresponding to the comoving distance of the sound horizon.

The main difficulty for detecting and analyzing the BAO's in large scale structures comes from the low statistical significance of the signal. It can only be seen on the widest galaxy surveys and has only been clearly detected in the most extended samples that include Luminous Red Galaxies. In addition to the statistical uncertainty the signal is affected by observational effects that may not be taken into account correctly, such as redshift distortions, mass-luminosity bias in the population of galaxies or wrong redshift to distance conversion.

Our goal here is to focus on the statistical uncertainty in the correlation function estimation, in particular at the BAO scale . There are two types of statistical uncertainties. The first one comes from cosmic fluctuations due to limited sample volume, and the other one from the finite number of galaxies which do not trace exactly the underlying field (i.e. shot noise).

There are various estimators of the correlation function. Their bias expresses the difference between their expected value and the value of the physical quantity of concern. Estimators are also subject to variance. In practice there is no way to evaluate the bias of the estimator if it exists, and it must be considered itself as a source of uncertainty, in addition to the estimator's variance.

To study these quantities on the estimators, we use simulations with ΛCDM power spectrum on the same volume as the data and with the same estimated parameters (density of galaxies, mass-luminosity bias, mean redshift for the power spectrum). Our simulations assume a lognormal model for the density field as proposed by [1], which has proven to be valid for a good range of scales.

A sufficient number of simulations provides the bias of the estimator by looking at the difference between its empiric mean value (i.e. average over all simulations) and the input correlation function. The variance of the estimators can be estimated by looking at the empiric 1σ fluctuation. This allows us to compare the different estimator's qualities and gives an idea of the uncertainty in the estimation.

The plan of the paper is as follows. In 1.1 we present the different estimators of the correlation function that we consider. We recall some of their properties, in particular their sensibility to the uncertainty in the mean density $\bar{n}$ in 1.2 and the bias imposed by the integral constraint in 1.3. In 2 we present the SDSS samples that we want to mimic with our simulations (one LRG sample and one main sample) and in 3 the lognormal model and our procedure for fitting simulation parameters to the data. Finally in 4 we perform the analysis of the uncertainty in the ξ estimation. We compare the quality of the different estimators in 4.1 and look at the effect of the integral constraint in the simulations in 4.2.

1. 2PCF estimators and bias

1.1. 2PCF estimators

The two-point correlation function is a second order statistic that describes the clustering of a field or a point process. More precisely $\xi(r)$ measures the excess of probability to find a pair of points in two volumes dV_1 and dV_2 at distance r compared to a random distribution.

$$dP_{12} = \bar{n}^2[1 + \xi(r)]dV_1 dV_2 \tag{1}$$

where $\bar{n}$ is the expected density of the distribution.

There are various estimators of the correlation function, most using random catalogues with identical geometry to measure this excess of probability. Let us define $DD(r)$, $RR(r)$ and $DR(r)$ as the number of pairs at a distance in $[r \pm dr/2]$ of respectively data-data, random-random and data-random points. We also define N_{DD}, N_{RR} and N_{DR} as the total number of corresponding pairs in the (real or random) catalog.

In this paper we will use 4 different estimators, Peebles-Hauser ([17]), Davis-Peebles ([2]), Hamilton ([10]) and Landy-Szalay ([12]), which have the following expressions:

$$\begin{aligned}
\hat{\xi}_{PH}(r) &= \frac{N_{RR}}{N_{DD}}\frac{DD(r)}{RR(r)} - 1 \\
\hat{\xi}_{DP}(r) &= \frac{N_{DR}}{N_{DD}}\frac{DD(r)}{DR(r)} - 1 \\
\hat{\xi}_{HAM}(r) &= \frac{{N_{DR}}^2}{N_{DD}N_{RR}}\frac{DD(r)RR(r)}{[DR(r)]^2} - 1 \\
\hat{\xi}_{LS}(r) &= 1 + \frac{N_{RR}}{N_{DD}}\frac{DD(r)}{RR(r)} - 2\frac{N_{RR}}{N_{DR}}\frac{DR(r)}{RR(r)}
\end{aligned}$$

Estimating ξ would be easier knowing the exact number of points in the volume expected from the distribution. In practice we can only estimate it with the empiric quantities N_D and N_{DD}. We show in section 1.2 that Hamilton and Landy-Szalay only depend on the second order on this uncertainty in the mean density and thus perform better. Moreover in [12] Landy-Szalay has been proven to be nearly of minimal variance for a random distribution (i.e. Poisson with no correlation).

1.2. Uncertainty in the mean density

We show the calculations given in [10] in a simple case where the sample is volume-limited so that the optimal strategy is to weight all galaxies equally. The empiric density in the catalogue n is a sum of Dirac functions on the galaxies of the catalogue. If $\bar{n}$ is the expected density then δ is the relative fluctuation in the sample:

$$\delta = \frac{n - \bar{n}}{\bar{n}} \tag{2}$$

We write W the indicator function of the sample volume and $<>$ the integration on the volume. For example $< W(\mathbf{x})\, n(\mathbf{x}) >$ is the integration of the empiric density and thus equals the number of points in the sample. We introduce the following quantities that have statistical expectations of 0:

$$\bar{\delta} = \frac{< W(\mathbf{x})\, \delta(\mathbf{x}) >}{< W(\mathbf{x}) >} \tag{3}$$

$$\Psi(r) = \frac{< \delta(\mathbf{x})\, W(\mathbf{x}) W(\mathbf{y}) >_r}{< W(\mathbf{x}) W(\mathbf{y}) >_r} \tag{4}$$

$$\hat{\xi}(r) = \frac{< \delta(\mathbf{x}) \delta(\mathbf{y})\, W(\mathbf{x}) W(\mathbf{y}) >_r}{< W(\mathbf{x}) W(\mathbf{y}) >_r} \tag{5}$$

where $<>_r$ means a double integration in the volume, restricted to $\mathbf{x}$ and $\mathbf{y}$ separated by a distance in $[r \pm dr/2]$. $\hat{\xi}$ is an unbiased estimator of the real ξ but we cannot calculate it since we do not know $\bar{n}$ and δ.

With short calculations it is possible to express the different estimators with the quantities $\hat{\xi}$, $\bar{\delta}$ and Ψ ([10]):

$$\hat{\xi}_{PH}(r) = \frac{\hat{\xi}(r) + 2\Psi(r) - 2\bar{\delta} - \bar{\delta}^2}{[1 + \bar{\delta}]^2}; \tag{6}$$

$$\hat{\xi}_{DP}(r) = \frac{\hat{\xi}(r) + \Psi(r) - \bar{\delta} - \Psi(r)\, \bar{\delta}}{[1 + \bar{\delta}]\, [1 + \Psi(r)]}; \tag{7}$$

$$\hat{\xi}_{H}(r) = \frac{\hat{\xi}(r) - \Psi(r)^2}{[1 + \Psi(r)]^2}; \tag{8}$$

$$\hat{\xi}_{LS}(r) = \frac{\hat{\xi}(r) - 2\bar{\delta}\Psi(r) + \bar{\delta}^2}{[1 + \bar{\delta}]^2}. \tag{9}$$

These formulas explain the superiority of Hamilton and Landy-Szalay estimators with Ψ and $\bar{\delta}$ terms at the second order in the numerator. Terms in the denominator are not important since they generate a small relative error, whereas terms in the numerator can generate a high relative error when their values become non negligible compared to $\hat{\xi}$. For Hamilton and Landy-Szalay estimators the error is dominated by the one of $\hat{\xi}$ and not really affected by Ψ and $\bar{\delta}$ which are linked to the uncertainty in $\bar{n}$.

With these formula we see that the estimators are biased in the general case. Indeed $\bar{\delta}$ and $\Psi(r)$ have expected value 0 and $\hat{\xi}(r)$ has expected value $\xi(r)$ but the terms are combined in multiplications and division. So we do not get the expected value of the left-hand side by replacing each term by its expected value in the right-hand side of equations 6, 7, 8, 9.

1.3. The integral constraint

The random catalogue is used to measure an excess of pairs compared to a random distribution. Equivalently it can be seen as a tool to calculate volumes. Let V be the domain of the sample, if we take the limit $N_R \to \infty$:

$$\frac{RR(r)}{N_{RR}} = \frac{\#\text{ pairs at distance } r' \in [r \pm dr/2]}{\#\text{ pairs}}$$
$$f(r) \overset{def}{=} \lim_{N_R \to \infty} \frac{RR(r)}{N_{RR}} = \frac{1}{|V|^2} \int_V d^3\mathbf{x} \int_V d^3\mathbf{y} \;\; 1_{|\mathbf{y}-\mathbf{x}| \in [r \pm dr/2]} \tag{10}$$

To simplify the text we define I and $\hat{I}_{PH}$, $\hat{I}_{DP}$, $\hat{I}_H$, $\hat{I}_{LS}$ ($\hat{I}$ when refering to any estimator) as the values of the integration against $f(r)$ for the real correlation and for the different estimators:

$$I \overset{def}{=} \int_0^{r_{\max}} f(r)\xi(r) \tag{11}$$
$$\hat{I}_{PH} \overset{def}{=} \int_0^{r_{\max}} f(r)\hat{\xi}_{PH}(r) \tag{12}$$

with $r_{\max}$ the maximum distance between 2 points in the volume.

We will show that there is a constraint on the Peebles-Hauser estimator $\hat{\xi}_{PH}(r)$ imposing the following equality, regardless of the real function $\xi(r)$ that is estimated:

$$\hat{I}_{PH} = 0 \tag{13}$$

For a smooth sample and small separation r, the inner integral in equation 10 equals for nearly all $\mathbf{y}$ the volume of the spherical envelope $\mathbf{x} \in V_r$ with $|\mathbf{y} - \mathbf{x}| \in [r \pm dr/2]$. So for small r we get $f(r) \approx \frac{|V_r|}{|V|} = \frac{4\pi r^2 dr}{|V|}$ and if it was the case for all r the constraint 13 would become:

$$\int_{\mathbb{R}^3} \hat{\xi}(\mathbf{r}) d^3\mathbf{r} = 0. \tag{14}$$

But when r becomes non negligible compared to the sample size, $f(r) \neq \frac{|V_r|}{|V|}$, and so the constraint 13 is different from 14 and depends on the sample volume and geometry.

Let us show the relation 13 for the Peebles-Hauser estimator:

$$\hat{\xi}_{PH}(r) = \frac{N_{RR}}{N_{DD}} \frac{DD(r)}{RR(r)} - 1 \approx \frac{1}{f(r)} \frac{1}{N_{DD}} DD(r) - 1 \tag{15}$$

In practice the integration consists in making the sum over all bins r_i of the correlation function estimated up to r_{max}:

$$\begin{aligned} \hat{I}_{PH} = \sum_i f(r_i)\hat{\xi}_{PH}(r_i) &\approx \sum_i f(r_i) \left[\frac{1}{f(r_i)} \frac{1}{N_{DD}} DD(r_i) - 1 \right] \\ &= \frac{1}{N_{DD}} \sum_i DD(r_i) - \sum_i f(r_i) = 1 - 1 = 0 \end{aligned}$$

It is possible to show that the same constraint $\hat{I} = 0$ should be approximately verified for the other estimators. For this we need to simplify $DR(r)$ in the limit $N_R \to \infty$.

$$\frac{1}{N_D N_R} DR(r) = \frac{1}{N_D} \sum_{\mathbf{d_j}} \frac{\#\text{ random points } \mathbf{r_i} \text{ s.t. } |\mathbf{r_i} - \mathbf{d_j}| \in [r \pm dr/2]}{\#\text{ random points}}$$
$$g(r) = \lim_{N_R \to \infty} \frac{1}{N_D N_R} DR(r) = \frac{1}{N_D} \sum_{\mathbf{d_j}} \frac{1}{V} \int_V d^3\mathbf{y} \;\; 1_{|\mathbf{y}-\mathbf{d_j}| \in [r \pm dr/2]} \tag{16}$$

This functions $g(r)$ depends on the point positions in the catalogue. We can make another approximation if the size of the correlation is small compared to the volume and if there are enough data

points. Then data points are approximately uniformly distributed in the volume and we can replace the mean on data positions by the mean on the volume:

$$g(r) \approx \frac{1}{V}\int_V d^3\mathbf{x}\left(\frac{1}{V}\int_V d^3\mathbf{y}\ \ 1_{|\mathbf{y}-\mathbf{x}|\in[r\pm dr/2]}\right) = f(r) \tag{17}$$

Under this approximation all estimators are equivalent and verify the integral constraint. But the last approximation is not as good as for Peebles-Hauser estimator and the constraint should be less tight.

We see again that the estimators are biased in the general case. The real correlation function does need not to verify the constraint, whereas the estimators do verify it and thus their expected value also.

1.4. Effect of the integral constraint on the bias

To show how the constraint can affect the correlation function estimation we generated realizations of segment Cox process (see [15]). The field consists in segments of length l randomly distributed in the volume and points randomly distributed on each segment. The intensity of the point process λ is equal to the mean length of segments per unit volume L_V times the mean number of points per unit length λ_l. This process is easy to sample and its correlation function is known analytically ([21]):

$$\xi(r) = \begin{cases} \frac{1}{2\pi r^2 L_V} - \frac{1}{2\pi r l L_V} & \text{for} \quad r \leq l \\ 0 & \text{for} \quad r \geq l \end{cases} \tag{18}$$

It is always positive so the integral constraint forces false negative values for the estimators. We considered the process with segment length $l = 10$ (units here are arbitrary), a mean length by unit volume $L_V = 0.1$ and a mean number of points per unit length $\lambda_l = 1.8$. We calculated the correlation function estimators on 2000 cubes of sizes $a = 10$, 2000 cubes of size $a = 20$ and 512 cubes of size $a = 50$. We plot figure 1 the mean value of the estimators on the samples and the empiric σ value. To exemplify the presence of the bias we show in the insets the empiric σ divided by $\sqrt{N}$ with N the number of realizations which gives the uncertainty in the empiric mean. A difference between the mean value and the real ξ much larger than $\frac{\sigma}{\sqrt{N}}$ means a bias is present in the estimators.

We observe that a bias is present for all estimators and for all sizes of cubes. As expected it becomes smaller when the sample size increases just like the variance decreases. For Landy-Szalay and Hamilton estimators the bias also decreases faster than the estimators' variances. The bias approximately equals half of the standard deviation σ in a large region for $a = 10$ and $a = 20$ whereas it is very small compared to the variance for $a = 50$. Biases are similar for the different estimators for $a = 10$ and $a = 20$ although Landy-Szalay and Hamilton have smaller variances than Peebles-Hauser and Davis-Peebles.

The effect of the bias is to force negative values at intermediate scales, so that the weighted sum in $\hat{I}$ approaches 0. Figure 2 shows the weighted estimators $f(r_i)\hat{\xi}(r_i)$ and how the integral cancels for the estimators and not for the real ξ. The effect is clear for $a = 20$ and $a = 10$ (not shown because results for $a = 10$ and for $a = 20$ have similar trends). For $a = 50$ however the bias comes not entirely from the integral constraint as the weighted function takes alternatively negative and positive values. So the small bias that is still present could come from other effects (e.g. finite number of random points).

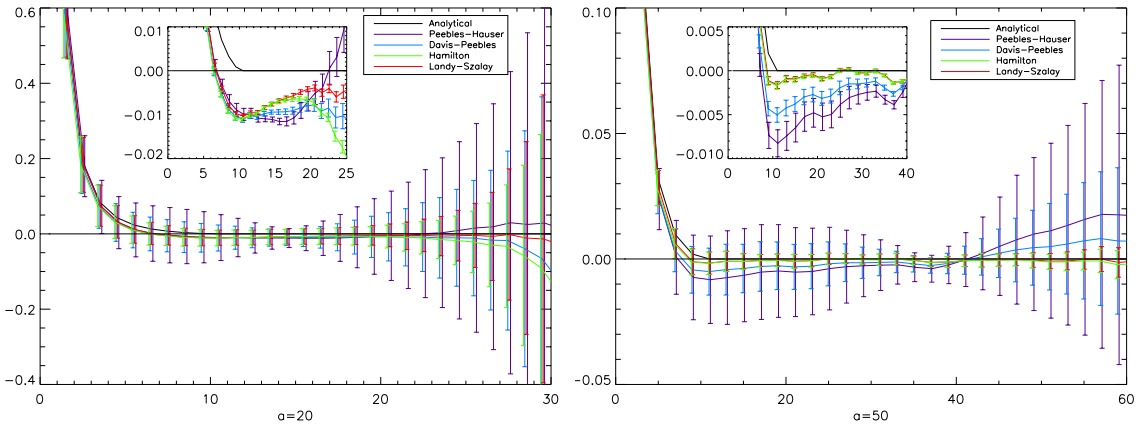

Figure 1: Mean and 1σ for the different estimators on 2000 Cox realizations for cubes size $a = 20$ (left panel) and 512 realizations for $a = 50$ (right panel). We plot the analytic function (black), Peebles-Hauser (purple), Davis-Peebles (light blue), Hamilton (green), Landy-Szalay (red). In inset we zoom over the biased region with error bars $\frac{\sigma}{\sqrt{N}}$ which is approximately the standard deviation of the mean on N realizations.

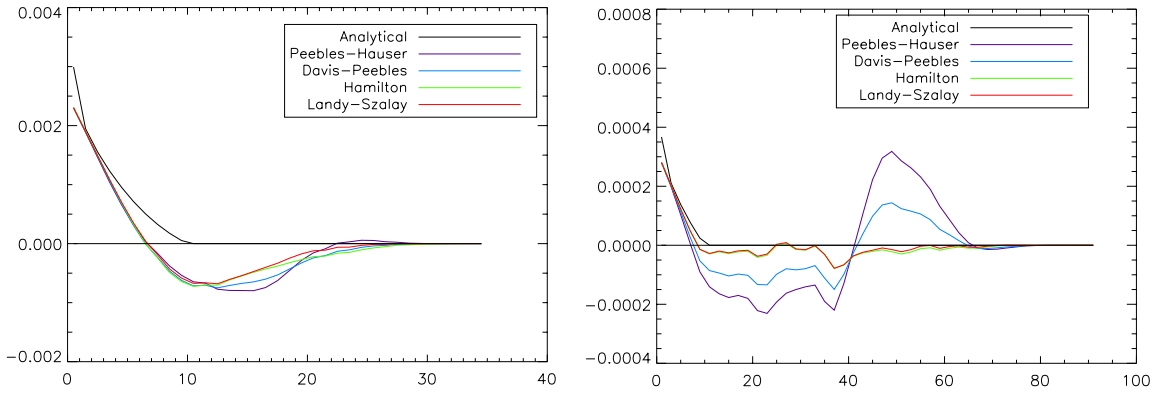

Figure 2: Weighted estimators $f(r_i)\hat{\xi}(r_i)$ for $a = 20$ (left panel) and $a = 50$ (right panel). We plot it for the analytic function (black), Peebles-Hauser (purple), Davis-Peebles (light blue), Hamilton (green), Landy-Szalay (red).

We show table 1 the value of I for the real ξ and $\hat{I}$ for the estimators' means. The constraint is close to be verified ($\hat{I} \approx 0$), especially for Peebles-Hauser, even when the real ξ does not verify it ($I \gg \hat{I}$).

The weight function f sums to 1 (see equation 10) so a the difference in $\sum_i f(r_i)\xi(r_i)$ and $\sum_i f(r_i)\hat{\xi}(r_i)$ (I and $\hat{I}$) implies in average a similar difference between ξ and $\hat{\xi}$. Negative bias may compensate positive bias in the integral, so it can be an underestimation.

For the Landy-Szalay and Hamilton estimators the constraint gets weaker between $a = 20$ and $a = 50$. These values of a correspond to values of I for the real ξ of approximately 0.01 and 0.001. A quantity which is more intuitive than I is the normalized mass variance inside a sample V:

$$\sigma^2(V) = \frac{\mathrm{Var}\left[M(V)\right]}{\mathbb{E}\left[M(V)\right]^2} \tag{19}$$

$\sigma(V)$ represents the fluctuation of mass in the sample. It can be shown that I is equal to $\sigma^2(V)$ up to the shot noise variance (see [9]), which can be usually neglected. Thus we can express conditions for the constraint to be weak or negligible in terms of the $\sigma(V)$ value. The cubic samples with $a = 20$ and $a = 50$ correspond respectively to $\sigma(V) \approx 0.1$ and $\sigma(V) \approx 0.03$. So the constraint still affects the

estimation for a 10% homogeneity level and starts to be weak for a 3% homogeneity level for this Cox process.

	I	$\sigma(V)$	$\hat{I}_{PH}$	$\hat{I}_{DP}$	$\hat{I}_{H}$	$\hat{I}_{LS}$
$a = 10$	0.059	0.24	2.87×10^{-5}	9.24×10^{-5}	7.16×10^{-3}	0.0125
$a = 20$	9.6×10^{-3}	0.098	1.88×10^{-5}	-5.54×10^{-4}	2.23×10^{-4}	1.47×10^{-3}
$a = 50$	7.8×10^{-4}	0.027	3.8×10^{-6}	-7.04×10^{-5}	-1.86×10^{-5}	1.02×10^{-4}

Table 1: Values of I and $\hat{I}$ for different estimators on cube sizes $a = 10$, $a = 20$ and $a = 50$

2. Samples and simulations

2.1. SDSS galaxy samples

We want to test the reliability of the correlation function estimation on current galaxy surveys. The largest survey up to date is the Sloan Digital Sky Survey (SDSS) with a final version in Data Release 7 (DR7,[6]). It contains a magnitude-limited sample of galaxies (main) and a nearly-volume-limited sample of luminous red galaxies (LRG).

To create volume-limited samples of the main we used the catalogue available in Mangle's webpage[1]. This catalogue is based on the New York University Value-Added Galaxy Catalog ([8]). It contains r-band absolute magnitudes (M_r) for each galaxy that are already K-corrected and corrected for evolution at a fiducial redshift of $z = 0.1$ following [7].

We also used a volume-limited sample of LRGs drawn directly from SDSS-DR7. LRGs are early-type galaxies selected using different luminosity and colour cuts [4], and extending to higher redshift. We computed the K-corrected g-band absolute magnitudes (M_g), and corrected for evolution at a fiducial redshift of $z = 0.3$, following the method described in [4].

In both cases, we obtained volume-limited samples by dividing the survey in different galaxy populations (according to the absolute magnitude in each case) and then cutting the sample at a minimum and a maximum redshift so that the density remains approximately constant. The selected volume-limited samples from the main catalogue are similar to those used by [5], while the LRG one is the same as used by [14].

Finally we restricted the samples to a region of the sky that is nearly complete except for small areas masked by bright stars. For this we cut the sample in the survey coordinate system (η, λ) with limits $-31.25^\circ < \eta < 28.75^\circ$ and $-54.8^\circ < \lambda < 51.8^\circ$. Because of this the samples are smaller and we have less statistics for correlation function estimation, but it is simpler for obtaining simulations in the same volume.

We give in Table 2 the magnitude and redshift limits used to construct the four volume-limited samples. We also give their total number of galaxies (N_g), volume (V) and mean density ($\bar{n}$).

In this paper only serve as reference and are used to adjust our simulation parameters.

Name	Magnitude Limits	Redshift Limits	Distance Limits $(h^{-1}\text{Mpc})$	N_g	V $(h^{-1}\text{ Mpc})^3$	$\bar{n}$ $(h^{-1}\text{ Mpc})^{-3}$
main1	$M_r < -20$	$0.038 < z < 0.119$	$112.94 < d < 346.99$	126733	22.571×10^6	5.615×10^{-3}
main2	$M_r < -21$	$0.059 < z < 0.168$	$174.50 < d < 484.06$	66327	60.492×10^6	1.096×10^{-3}
main3	$M_r < -21.5$	$0.071 < z < 0.205$	$209.40 < d < 585.27$	29576	107.041×10^6	2.763×10^{-4}
LRG	$-23.544 < M_g < -21.644$	$0.14 < z < 0.42$	$410 < d < 1140$	34347	790.4×10^6	4.345×10^{-5}

Table 2: Characteristics of the SDSS samples

2.2. Simulations

2.2.1. The lognormal model

The usual paradigm for the distribution of galaxies n_g is the Cox process, i.e. a Poisson process with an intensity given by a continuous field $\rho_g(\mathbf{x})$, which itself is a statistical process. Knowing $\rho_g(\mathbf{x})$

[1] http://space.mit.edu/∼molly/mangle/

the number of galaxies in a volume dV around $\mathbf{x}$ is a Poisson variable with intensity $\rho_g(\mathbf{x})dV$. It can be verified that the correlation function of the point process is the same as the underlying continuous process ρ_g plus a weighted Dirac function $\frac{1}{\bar{n}}\delta_0$ due to the discreteness.

The process ρ_g is linked to the underlying matter density field ρ_m since galaxies form in matter over-densities, but is not supposed to be identical. Indeed it has been observed that correlation is higher in the galaxy distribution than in the matter field, and also depends on galaxy population. The ratio of the two is the square of the mass-luminosity bias b. Note that the term bias here has a different meaning than when we speak about the bias of estimators. The mass-luminosity bias quantifies how fluctuations are amplified in the distribution of galaxies whereas the bias of an estimator is the difference between its expected value and the quantity to estimate.

In general b should depend on the scale but here we will simplify and consider it constant:

$$\xi_g(r) = b^2\ \xi_m(r). \tag{20}$$

We will consider a galaxy field ρ_g following a lognormal model as proposed in [1]. A lognormal field Y with an expected value of 1 is obtained from a gaussian field X by:

$$Y = e^{X - \frac{\sigma_X^2}{2}}. \tag{21}$$

This model has been successfully applied to density field reconstruction in [11], where it enters as a prior model for the matter field. The lognormal model is quite simple and has other interesting properties (see [1]):

- It describes well the distribution of galaxies as found by Hubble (1934) and recently in [11] when the galaxy field is smoothed on scales between 10 and 30 Mpc
- The positivity of the field is ensured unlike in a gaussian model
- It is completely described by its correlation function as the gaussian field (and numerous quantities can be calculated as easily, i.e. statistics of the peaks, genus)
- It is arbitrarily close to a gaussian field at early times where $\sigma \approx 0$
- It is the solution of the equations of evolution of ρ when supposing that the initial density field peculiar velocities are gaussian

2.2.2. Adjusting simulation parameters

We adopt for our simulations a ΛCDM power spectrum $P_{\Lambda\mathrm{CDM}}$ given by the iCosmo software ([19]) with the WMAP7 cosmological parameters. We decided to reproduce the main2 sample, given in section 2.1, which is an average main sample, and the LRG sample. We take the power spectrum at the mean redshift for each sample, i.e. at redshift $z = 0.1$ for the main2 sample, and at $z = 0.3$ for the LRG sample.

The simulations give the continuous field ρ_g on a discrete grid with a size 700 by 700 by 700 with a physical size of $(1200\ h^{-1}\,\mathrm{Mpc})^3$ for the main2 sample and $(1600\ h^{-1}\,\mathrm{Mpc})^3$ for the LRG sample, i.e. with elementary cells respectively of $1.71\ h^{-1}\,\mathrm{Mpc}$ and $2.29\ h^{-1}\,\mathrm{Mpc}$. We then place in each cell a number of galaxies which is a Poisson realization of intensity ρ_g at this cell. Overall this will have the effect of smoothing the correlation function approximately with the cell size.

We choose a mean density of points in the volume that gives on average the same number of points as in the SDSS samples.

A last step is to choose the mass-luminosity bias b between the samples and the ΛCDM matter correlation function. For estimating this factor we fit the ΛCDM correlation function to the one estimated on the data $\hat{\xi}$:

$$b^2\,\xi_{\Lambda\mathrm{CDM}} \approx \hat{\xi} \tag{22}$$

By this method we find a bias for the main samples (the variation of b is rather small between the different main samples) and for the LRG sample compared to ΛCDM at redshift $z = 0.1$ and $z = 0.3$. We find respectively $b = 1.5$ and $b = 2.5$: as usually observed, the bias increases with luminosity

([13],[22]). The bias obtained for the LRG is a bit larger, than the one usually found for LRG, $b \approx 2$ (e.g. in [20]). This probably comes from the fact that we selected only brightest galaxies of the LRG population.

3. Uncertainty in estimating ξ

3.1. Bias and variance of the estimators

For each sample (main2 and LRG) we use 200 lognormal simulations with the parameters described before and compute the different estimators for each realization. Each time we use respectively 100 000 and 150 000 random points which is enough so that the corresponding error is small. Each time a different random catalogue so when we take the mean over all realizations for the analysis of the bias, the effect of finite number of random points is completely negligible.

Yet on individual realizations the fluctuation due to finite number of random points can increase a little bit the variance of the estimators. For a given contribution to the variance the number of required points is related to the volume size and geometry, and to the size of the bins for estimating ξ (in all our tests we took bin sizes of 10). More precisely the condition is that $\frac{1}{N_{RR}} RR(r)$ approximates with a given precision $\frac{1}{V^2} \int_V d^3\mathbf{x} \int_V d^3\mathbf{y} \;\; 1_{|\mathbf{y}-\mathbf{x}| \in [r \pm dr/2]}$.

We show in figure 3 the estimator's means compared to the theoretical ΛCDM correlation function. For clarity the curves have been translated by ± 1 h^{-1}Mpc. A bias can be seen for the estimation in the main sample as the mean differs by approximately half of the variance from the true value for $r > 90$ h^{-1}Mpc. However on the LRG, sample estimator's means are nearly indistinguishable from the theoretical values.

This also validates our simulation process which gives an output correlation function fitting very well the one in input. There is a small difference at the scale of the BAO (in addition to the bias) that we attribute to the smoothing introduced by grid discretization described section 2.2.2. The BAO is a local maximum so the function decreases after smoothing.

Concerning the estimator's variances there are much smaller on the LRG sample than on the main since the volume is bigger and the Poisson fluctuations remain small for a number of galaxies $N_D \approx 34000$.

We also see that Hamilton and Landy-Szalay estimators are much better than the 2 others in terms of variance. This agrees with previous studies ([18]) showing a superiority of these estimators. It also agrees with the analysis in [12] considering a Poisson process with no correlation. In the latter case Landy-Szalay and Hamilton estimators have second order variances decay in $\frac{1}{n^2}$ with n the number of data points (i.e. a $\frac{1}{|V|^2}$ decay with $|V|$ the volume size) whereas Peebles-Hauser and Davis-Peebles have first order decay in $\frac{1}{n}$.

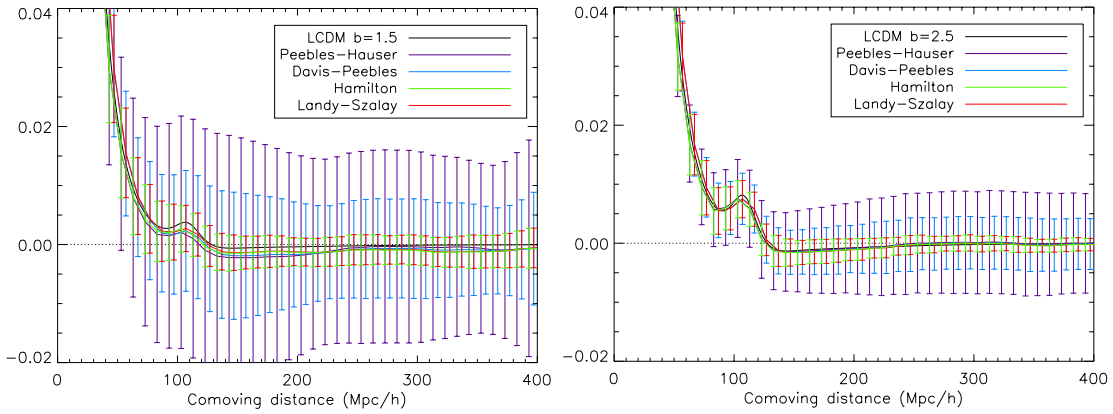

Figure 3: Left Panel: Different estimators' means and 1σ for 200 main2 realizations, Peebles-Hauser (purple), Davis-Peebles (light blue), Hamilton (green), Landy-Szalay (red) and ΛCDM at $z = 0.1$ with $b = 1.5$ (black). Right Panel: Same for the LRG sample except ΛCDM at $z = 0.3$ with $b = 2.5$.

3.2. Effect of the integral constraint

We are interested here in the influence of the constraint studied in section 1.3. The constraint writes $\int_0^{r_{max}} f(r)\hat{\xi}(r) = 0$ with $f(r) \approx \frac{4\pi r^2 dr}{|V|}$ for small r. Assuming $\int_0^\infty r^2\ \xi(r)\ dr$ is finite the value of $\int_0^{r_{max}}\ f(r)\ \xi(r)dr$ vanishes as $\frac{1}{|V|}$ at large volumes. In usual ΛCDM models the power spectrum verifies $P(0) = 0$ and thus the correlation function verifies $\int_0^\infty\ r^2\ \xi(r)dr = 0$ which makes the constraint even more easy to be satisfied.

Table 3 shows the value of the constrained integral for theoretical $\xi_{\Lambda\mathrm{CDM}}$ and for the measured $\hat{\xi}$, respectively I and $\hat{I}$. For the main2 sample $\hat{I}$ is significantly closer to 0 than I meaning that the constraint has an effect on the estimation. The effect is negligible for the LRG sample.

The value of I gives the mean bias of $\hat{\xi}$ caused by the integral constraint: it is of order 10^{-3} for the main2 sample and of order 10^{-4} for the LRG sample. Comparing to the values of ξ at the scales of interest (i.e. usually between 50 and $150h^{-1}$Mpc) the bias is significant for the main2 sample but it is negligible for the LRG sample.

We can also make a parallel with the Cox model of section 1.3 where the effect of the constraint becomes very small for $\sigma(V) \approx 0.03$. The main2 sample has the same value $\sigma(V) \approx 0.03$ but the effect is still important. In the LRG sample the value is 3 times smaller, $\sigma(V) \approx 0.01$ so it is not surprising that the effect is negligible.

	I	$\sigma(V)$	$\hat{I}_{PH}$	$\hat{I}_{DP}$	$\hat{I}_H$	$\hat{I}_{LS}$
main2	8.85×10^{-4}	≈ 0.03	5.93×10^{-6}	$-1.65 \times 10-4$	-1.42×10^{-4}	4.95×10^{-5}
LRG	1.10×10^{-4}	≈ 0.01	1.05×10^{-4}	8.16×10^{-5}	6.61×10^{-5}	7.41×10^{-5}

Table 3: Values of I and $\hat{I}$ for different estimators

Acknowledgements

We acknowledge the use of the Sloan Digital Sky Survey data (http://www.sdss.org) and of the NYU Value-Added Galaxy Catalog (http://sdss.physics.nyu.edu/vagc/).

This research was supported by the European Research Council grant ERC-228261.

Conclusion

We have studied statistical properties, in particular uncertainties, of the correlation function estimators.

For this we simulated lognormal mock galaxy catalogues; the different parameters of the simulations are adjusted to those of the SDSS samples: mean redshift of the ΛCDM input power spectrum, density of galaxies in the sample, mass-luminosity factor bias. Using enough realizations, we quantified the uncertainty in ξ coming from both estimators' variances and biases.

We compared different estimators of the correlation function, in particular regarding their sensibility to the fluctuation in the number of galaxies n (i.e. the uncertainty in the mean density): Peebles-Hauser and Davis-Peebles depend at first order on that fluctuation; whereas Hamilton and Landy-Szalay have a second order dependence. As a consequence the variances of the first two estimators have only a first order decay in the volume size whereas the two latter estimators have a second order decay. We confirmed with the simulations that Hamilton and Landy-Szalay have much smaller variances.

We evaluated the effect of the integral constraint: it can affect the estimation for small volumes, but this effect is negligible when the real ξ itself is close to to verify the constraint. For the Cox process the effect becomes very small when fluctuations in the volume are less than 3% ($\sigma(V) < 0.03$). This homogeneity level is achieved for one of the main galaxy sample. Yet for this sample the integral constraint still affects the estimation with a bias in $\hat{\xi}(r)$ of approximately $0.5\,\sigma$ for $r > 90\ h^{-1}$Mpc. For the LRG sample with $\sigma(V) \approx 0.01$, the estimators are unbiased and the estimation of ξ is reliable up to variance.

References

[1] P. Coles and B. Jones. A lognormal model for the cosmological mass distribution. *Mon. Not. R. Astron. Soc.*, 248:1–13, 1991.

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

[3] D. J. Eisenstein and W. Hu. Baryonic features in the matter transfer function. *The Astrophysical Journal*, 496(2):605, 1998.

[4] D. J. Eisenstein et al. Spectroscopic Target Selection for the Sloan Digital Sky Survey: The Luminous Red Galaxy Sample. *Astronomical Journal*, 122:2267–2280, November 2001. SDSS.

[5] I. Zehavi et al. The Luminosity and Color Dependence of the Galaxy Correlation Function. *Astrophysical Journal*, 630:1–27, September 2005. SDSS.

[6] K. N. Abazajian et al. The Seventh Data Release of the Sloan Digital Sky Survey. *Astrophysical Journal Supplement Series*, 182:543–558, June 2009.

[7] M. R. Blanton et al. The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1. *Astrophysical Journal*, 592:819–838, August 2003.

[8] M. R. Blanton et al. New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. *Astronomical Journal*, 129:2562–2578, 2009.

[9] A. Gabrielli, M. Joyce, and F. S. Labini. Glass-like universe: Real-space correlation properties of standard cosmological models. *Phys. Rev. D*, 65(8):083523, Apr 2002.

[10] A. J. S. Hamilton. Toward better ways to measure the galaxy correlation function. *Astrophysical Journal*, 417:19, November 1993.

[11] F. S. Kitaura, J. Jasche, and R. B. Metcalf. Recovering the nonlinear density field from the galaxy distribution with a poisson-lognormal filter. Technical Report arXiv:0911.1407, Nov 2009. Comments: 17 pages, 9 figures, 1 table.

[12] S. D. Landy and A. S. Szalay. Bias and variance of angular correlation functions. *Astrophysical Journal*, 412:64–71, July 1993.

[13] C. Li, G. Kauffmann, Y. P. Jing, S. D. M. White, G. Boerner, and F. Z. Cheng. The dependence of clustering on galaxy properties, 2005.

[14] V. J. Martínez, P. Arnalte-Mur, E. Saar, P. de la Cruz, M. J Pons-Bordería, S. Paredes, A. Fernández-Soto, and E. Tempel. Reliability of the Detection of the Baryon Acoustic Peak. *Astrophysical Journal*, 696:L93–L97, 2009.

[15] V. J. Martínez and E. Saar. *Statistics of the Galaxy Distribution*. Chapman & Hall/CRC, 2002.

[16] P. J. E. Peebles. *The Large-Scale Structure of the Universe*. Princeton University Press, 1980.

[17] P. J. E. Peebles and M. G. Hauser. Statistical analysis of catalogs of extragalactic objects. iii. the shane-wirtanen and zwicky catalogs. *The Astrophysical Journal Supplement Series*, 28:19–+, November 1974.

[18] M. Pons-Bordería, V. J. Martínez, D. Stoyan, H. Stoyan, and E. Saar. Comparing estimators of the galaxy correlation function. *Astrophysical Journal*, 523:480–491, October 1999.

[19] A. Refregier, A. Amara, T. Kitching, and A. Rassat. iCosmo: an Interactive Cosmology Package. 2008.

[20] U. Sawangwit, T. Shanks, F. B. Abdalla, R. D. Cannon, S. M. Croom, A. C. Edge, N. P. Ross, and D. A. Wake. Angular correlation function of 1.5 million lrgs: clustering evolution and a search for bao. 2009.

[21] D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, Chichester, 1995.

[22] I. Zehavi, D. J. Eisenstein, R. C. Nichol, M. R. Blanton, D. W. Hogg, J. Brinkmann, J. Loveday, A. Meiksin, D. P. Schneider, and M. Tegmark. The intermediate-scale clustering of luminous red galaxies, 2004.