Experimental cosmic statistics - I. Variance

Stéphane Colombi, István Szapudi, Adrian Jenkins, Jörg Colberg

Mon. Not. R. Astron. Soc. **313**, 711–724 (2000)

Experimental cosmic statistics – I. Variance

Stéphane Colombi,[1] István Szapudi,[2,3]★† Adrian Jenkins[2] and Jörg Colberg[4]
[1]*Institut d'Astrophysique de Paris, CNRS, 98bis bd Arago, F-75014 Paris, France*
[2]*University of Durham, Department of Physics, South Road, Durham DH1 3LE*
[3]*Canadian Institute of Theoretical Astrophysics, 60 St George St, Toronto, Ontario M5S 3H8, Canada*
[4]*Max-Planck-Institut für Astrophysik, D-85740 Garching, Germany*

Accepted 1999 November 17. Received 1999 November 17; in original form 1999 May 24

ABSTRACT

Counts-in-cells are measured in the τCDM Virgo Hubble Volume simulation. This large N-body experiment has 10^9 particles in a cubic box of size $2000\,h^{-1}$ Mpc. The unprecedented combination of size and resolution allows, for the first time, a realistic numerical analysis of the cosmic errors and cosmic correlations of statistics related to counts-in-cells measurements, such as the probability distribution function P_N itself, its factorial moments F_k and the related cumulants $\bar{\xi}$ and S_Ns. These statistics are extracted from the whole simulation cube, as well as from 4096 subcubes of size $125\,h^{-1}$ Mpc, each representing a virtual random realization of the local universe.

The measurements and their scatter over the subvolumes are compared to the theoretical predictions of Colombi, Bouchet & Schaeffer for P_0, and of Szapudi & Colombi and Szapudi, Colombi & Bernardeau for the factorial moments and the cumulants. The general behaviour of experimental variance and cross-correlations as functions of scale and order is well described by theoretical predictions, with a few per cent accuracy in the weakly non-linear regime for the cosmic error on factorial moments. On highly non-linear scales, however, all variants of the hierarchical model used by SC and SCB to describe clustering appear to become increasingly approximate, which leads to a slight overestimation of the error, by about a factor of two in the worst case. Because of the needed supplementary perturbative approach, the theory is less accurate for non-linear estimators, such as cumulants, than for factorial moments.

The cosmic bias is evaluated as well, and, in agreement with SCB, is found to be insignificant compared with the cosmic variance in all regimes investigated.

While higher order statistics were previously evaluated in several simulations, this work presents textbook quality measurements of S_Ns, $3 \leq N \leq 10$, in an unprecedented dynamic range of $0.05 \lesssim \bar{\xi} \lesssim 50$. In the weakly non-linear regime the results confirm previous findings and agree remarkably well with perturbation theory predictions including the one-loop corrections based on spherical collapse by Fosalba & Gaztañaga. Extended perturbation theory is confirmed on all scales.

Key words: methods: numerical – methods: statistical – galaxies: clusters: general – large-scale structure of Universe.

1 INTRODUCTION

Measurements of higher order statistics in galaxy catalogues test theories of structure formation, the nature of the initial fluctuations and the processes of galaxy formation. The power of such measurements to constrain theories, however, depends crucially on the detailed understanding of the errors. Usually it is tacitly assumed that the underlying distribution of events is Gaussian and thus the term 'errors' becomes synonymous with the 'variance'. Knowledge of the variance is sufficient only when the error distribution is Gaussian.

For statistics related to counts-in-cells a rigorous theory for the cosmic errors was presented in a suite of papers by Szapudi & Colombi (1996, hereafter SC), Colombi, Szapudi & Szalay (1998) and Szapudi, Colombi & Bernardeau (1999a, hereafter SCB). Nevertheless these calculations relied on approximations, for

★ E-mail: szapudi@cita.utoronto.ca
† Present address: Canadian Institute of Theoretical Astrophysics, 60 St George St, Toronto, Ontario M5S 3H8, Canada.

which the domain of validity could not be checked extensively until the arrival of the Virgo Hubble Volume simulations. Moreover, the regime where the underlying cosmic distribution is Gaussian could not be examined previously. This paper addresses the first problem by studying the statistical errors and cross-correlations numerically, while a companion paper, Szapudi et al. (2000, hereafter Paper II, in this issue), discusses the underlying distributions of statistics in their full splendour.

Let us consider a statistic A measured in a galaxy catalogue of volume V. The corresponding indicator is denoted by $\tilde{A}$. In practice, only one sample of our local universe is accessible. However, a frequentist numerical experiment can be performed in a large numerical simulation if a sufficient number $C_{\mathcal{E}}$ of galaxy catalogues $\mathcal{E}_i$ can be extracted from it. In each of them a value $\tilde{A}_i$, $1 \leq i \leq C_{\mathcal{E}}$, can be measured.

For any statistic A the cosmic distribution function $\Upsilon(\tilde{A})$ is the probability density of measuring the value $\tilde{A}$ in a particular finite realization. This distribution function can be approximately extracted from the $C_{\mathcal{E}}$ subsamples under the ergodic hypothesis. For simplicity, we dispense with the (logical) notation $\tilde{\Upsilon}$ and replace it in what follows with Υ. This expresses the fact that we do not wish to enter one more level of complexity by considering the 'error on the error' problem (SC) in greater detail. The smoothness and regularity of our measurements suggest that the number of realizations, which represent a two orders of magnitude improvement over any previous work, is large enough to provide an adequate determination of the quantities measured.

While in practice the function $\Upsilon(\tilde{A})$ is the fundamental quantity underlying all measurements, this paper concentrates on its first two moments; Paper II examines its shape and skewness in detail.

In the following definitions, integrals are to be understood as summations of the estimator over the distribution function. The first moment of $\Upsilon(\tilde{A})$ is the spatial average

$$\int \tilde{A}\Upsilon(\tilde{A})\,\mathrm{d}\tilde{A} = \langle\tilde{A}\rangle \equiv A, \tag{1}$$

where it is assumed that the estimator $\tilde{A}$ is *unbiased*. The bias is negligible compared to the relative cosmic error in most meaningful cases (SCB) as illustrated later by practical examples. For completeness, however, the definition of the *cosmic bias* is

$$b_A \equiv \frac{\langle\tilde{A}\rangle - A}{A}. \tag{2}$$

The second (centred) moment of the cosmic distribution is called the cosmic error,

$$\int (\tilde{A} - A)^2\Upsilon(\tilde{A})\,\mathrm{d}\tilde{A} = \langle(\tilde{A} - A)^2\rangle \equiv (\Delta A)^2. \tag{3}$$

For a biased statistic, the variance should be centred around the biased average and not the true value. It can however be shown formally (SCB) that the above definition is valid to second order in $\Delta A/A$ for any biased statistic.[1]

Finally, the cosmic covariance can be defined analogously to the variance as $\langle(\tilde{A} - A)(\tilde{B} - B)\rangle$.

The theoretical results for the errors and cross-correlations are summarized below. If v and V are the cell and catalogue volumes respectively, the cosmic error can be approximately separated into three components to leading order in v/V (SC).

(1) The discreteness or shot-noise error which is the result of the finite number of objects N_{obj} in the catalogue, increases towards small scales and with the order of the statistics considered, but becomes negligible when N_{obj} is very large.

(2) The edge effect error is the result of the uneven weight given to galaxies near the edges of the survey compared to those near the centre. It is especially significant on large scales, comparable to the size of the catalogue.

(3) The finite volume error is the result of fluctuations of the underlying density field on scales larger than the characteristic size of the catalogue.

The next to leading order correction in v/V is proportional to the perimeter of the catalogue ∂V. At this level of accuracy there are also correlations between the three sources of error (e.g. Colombi et al., in preparation, hereafter CCDFS).

Colombi, Bouchet & Schaeffer (1995, hereafter CBS) investigated in detail the cosmic error on the void probability function. The groundwork for error calculations of statistics related to counts-in-cells is based on SC where the cosmic error for factorial moments[2] was evaluated *analytically*. SCB extended the work of SC to cross-correlations, including perturbation theory predictions (e.g. Bernardeau 1996, hereafter B96). The cosmic errors, biases (see also Hui & Gaztañaga 1999, hereafter HG) and covariances for cumulants[2] $\bar{\xi}$ and S_N were calculated as well. The main goal of this paper is to compare the analytical predictions of CBS, SC and SCB to measurements made in the VIRGO τCDM Hubble Volume simulation.

The exhaustive nature of the comparison that follows warrants the questions: is it meaningful to thrive for the detailed numerical understanding of the theory? How much of it is practically useful? Can it accurately estimate the errors on measurements in future surveys? While some of these questions were addressed in SCB, a brief account of supporting arguments is given next.

The analytics do take into account all possible theoretical errors, but systematics, such as those resulting from cut out holes, incompleteness from fibre separation, possible magnitude errors in the case of the 2dF, etc., could in principle corrupt the theory and introduce biases. These effects might even require detailed simulation of the survey. In the case of the UKST (United Kingdom Schmidt Telescope) and Stromlo surveys such simulations were performed and compared with the predictions: the spectacular agreement surprised even the present authors (Hoyle, Szapudi & Baugh, in preparation). Thus systematics do not dominate in all surveys; for another example, where cut out holes were found to have an insignificant effect on the cosmic probability distribution of the two-point correlations function see Kerscher, Szapudi & Szalay (in preparation).

Moreover, the wide theoretical framework is flexible enough to incorporate all systematics, which have the effect of altering certain parameters, such as the factorial moments. In such a case any bias can be corrected for.

There might be unforeseen systematics which have such a complicated non-linear effect that they cannot even be modelled by the appropriate alteration of a set of parameters. While it would be difficult to anticipate whether these could dominate for a particular survey, it is still instructive to investigate the potential results in an ideal case, especially during the design phase of the survey. Error calculations help in optimizing geometry, sampling and other parameters. During the design of the Visible (Near) IR

[1] More precisely, to first order in $\langle(\tilde{x}_i - x_i)(\tilde{x}_j - x_j)\rangle$ where $\tilde{x}_i$ denote the unbiased estimators from which $\tilde{A}$ is constructed in a non-linear fashion.

[2] For example, see Appendix A for definitions and notations.

Table 1. The scales for which we measured the CPDF.

$\ell(h^{-1}\,\mathrm{Mpc})$	0.24	0.49	0.98	1.95	3.91	7.8	15.6	31.3	62.5	125	250
$\mathcal{E}$					√	√	√	√	√	√	√
$\mathcal{E}_i$		√	√	√	√	√	√	√	√		

Multi-Object Spectrograph (VIRMOS) survey such considerations were taken into account (Colombi et al., in preparation). These calculations, as well as maximum likelihood analyses, need to explore such a large region in parameter space that they would typically be impractical to carry out with simulations.

In addition to applications to surveys, the theory can be applied reliably to assess significance of measurements in simulations where multiple runs would be too costly (e.g. Szapudi et al. 1999b). All these present and potential future applications motivate the detailed investigations performed in this article.

The exposition is organized as follows. Section 2 describes the N-body data used for the purpose of our study. Section 3 analyses the count-in-cells distribution function P_N, its cumulants $\bar{\xi}$ and S_Ns, and the scaling function of the void probability distribution $\sigma \equiv -\ln(P_0)/F_1$. These quantities are measured in the full simulation as well as in $C_\mathcal{E} = 4096$ subsamples. The accuracy of the simulation is assessed by comparing the measurements to the non-linear Ansatz of Hamilton et al. (1991) improved by Peacock and Dodds (1996, hereafter PD), and to perturbation theory (hereafter PT) predictions. The model of Fosalba & Gaztañaga (1998) and extended perturbation theory (hereafter EPT, see Colombi et al. 1997) are considered as well. Section 4 extends these investigations to the cosmic error and the variance of the cosmic distribution function. A preliminary investigation of the cross-correlations is done for factorial moments and cumulants. The measurements are compared where possible to the theoretical predictions of SC, SCB and CBS, including extended perturbation theory. Finally Section 5 recapitulates the results and discusses their implications. In addition, Appendix A gives a summary of the definitions and notations used in this paper for counts-in-cells statistics. It will be useful for the reader unfamiliar with these concepts.

2 THE N-BODY DATA

The τCDM Hubble Volume simulation (e.g. Evrard et al., in preparation) was carried out using a parallel P^{3}M code described in MacFarland et al. (1998). The code was run on 512 processors of the Cray T3E-600 at the Rechenzentrum in Garching.

Initial conditions were laid down by imposing perturbations on an initially uniform state represented by a 'glass' distribution of particles generated by the method of White (1996). Because of the size of the simulation, a glass file of 10^6 particles was tiled 10 times in each direction. As the initial glass file was created with periodic boundary conditions tiling does not create any non-uniformities at the interface between the tiles.

A Gaussian random density field was set up by perturbing the positions of the particles and assigning velocities to them according to the growing mode linear theory solutions, using the algorithm described by Efstathiou et al. (1985). Individual modes were assigned random phases and the power for each mode was selected at random from an exponential distribution with mean power corresponding to the desired power spectrum $\langle|\delta_k^2|\rangle$. Unlike Efstathiou et al. (1985), however, the initial velocities were set up exactly proportional to the initial displacements, according to the Zel'dovich (1970) approximation. As shown by Scoccimarro (1998) this leads to larger initial transients. To compensate for this the simulation was started at a high redshift, $z = 29$.

The cosmological model used for the simulation τCDM is described in more detail in Jenkins et al. (1998). The approximation to the linear cold dark matter (CDM) power spectrum (Bond & Efstathiou 1984) was used

$$\langle|\delta_k^2|\rangle = \frac{Ak}{\{1 + [aq + (bq)^{3/2} + (cq)^2]^\nu\}^{2/\nu}}, \tag{4}$$

where $q = k/\Gamma$, $a = 6.4\,h^{-1}\,\mathrm{Mpc}$, $b = 3\,h^{-1}\,\mathrm{Mpc}$, $c = 1.7\,h^{-1}$ Mpc and $\nu = 1.13$. The value of Γ was set equal to 0.21. The normalization constant, A, is chosen by fixing the value of σ_8^2 (the linear variance of the matter distribution in a sphere of radius $8\,h^{-1}\,\mathrm{Mpc}$ at $z = 0$). A value of $\sigma_8 = 0.6$ was motivated by estimates based on cluster abundances (White, Efstathiou & Frenk 1993; Eke, Cole & Frenk 1996).

The simulation was integrated using a leapfrog scheme as described in Hockney & Eastwood (1981), section 11-4-3. The simulation was completed in 500 equal steps in time. The softening used was $100\,\mathrm{kpc}\,h^{-1}$ comoving Plummer equivalent – see Jenkins et al. (1998).

3 COUNTS-IN-CELLS ANALYSIS: THE UNDERLYING STATISTICS

The count probability distribution function (CPDF) P_N is defined as the probability of finding N objects in a cell of volume v thrown at random in the catalogue. CPDF was measured in the whole simulation $\mathcal{E}$ for cubic cells of size $L_\mathrm{box}/512 \leq \ell \leq L_\mathrm{box}/8$, where $L_\mathrm{box} = 2000\,h^{-1}\,\mathrm{Mpc}$ is the size of the simulation cube (see Table 1). Then the simulation cube was divided into 16^3 contiguous cubic subsamples $\mathcal{E}_i$ of size $L = 125\,h^{-1}\,\mathrm{Mpc}$. P_N was evaluated in each of these for $L/512 \leq \ell \leq L/2$ (see Table 1). The successive convolution algorithm of Szapudi et al. (1999b, hereafter SQSL) allowed the determination of the CPDF on all scales simultaneously in only a few minutes of CPU on a workstation[3] with 512^3 sampling cells. The accuracy is thus $P_N \geq P_{\mathrm{min},1} = 1/512^3 \simeq 7.45 \times 10^{-9}$ for the measurement in $\mathcal{E}$ and for each individual $\mathcal{E}_i$; the accuracy increases by averaging over all subsamples: $P_N \geq P_{\mathrm{min},2} = 1/(512 \times 16)^3 \simeq 1.82 \times 10^{-12}$. For $4 \lesssim \ell \lesssim 63\,h^{-1}\,\mathrm{Mpc}$ the measurements in $\mathcal{E}$ and $\mathcal{E}_i$ overlap (Table 1). This is illustrated by Fig. 1, displaying P_N as a function of N: the figure presents the CPDF extracted from both the full cube and averaged over all the subcubes. In the overlap region, the difference can be detected as slight irregularities of the high-N tail from the full cube measurements. The figure suggests that at least on the smallest scales considered in $\mathcal{E}$ (or each $\mathcal{E}_i$), our sampling is probably insufficient by the standards of SC. However, this does not affect significantly the calculations as indicated by the agreement of the moments measured in $\mathcal{E}$ and those calculated from averages obtained from the subsamples. Therefore measurement errors will be neglected in what follows, i.e. *infinite*

[3] This estimate does not include the reading in of the file.

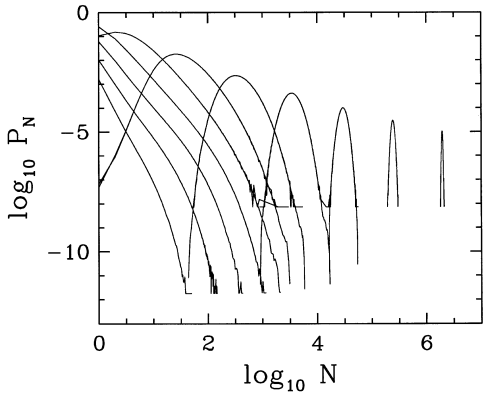

Figure 1. The measured CPDF as a function of N. Various scales are plotted as described in the text and in Table 1. The curves shift to the right as ℓ increases.

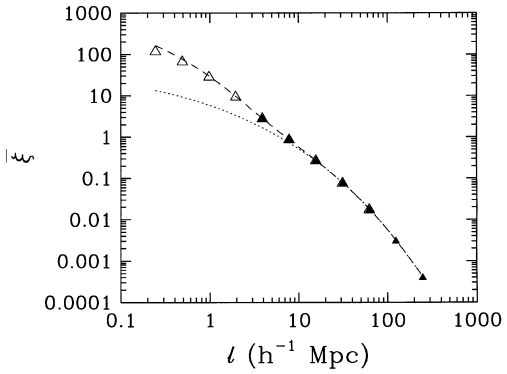

Figure 2. The averaged two-point correlation function $\bar{\xi}$ as a function of scale. It is compared with linear theory (dots) and with the non-linear Ansatz of Hamilton et al. (1991) with the recipe of Peacock & Dodds (1996) (dashes). The open symbols correspond to the $\bar{\xi}$ obtained from the CPDF averaged over all the subsamples $\mathcal{E}_i$ and the filled symbols to the measurement in $\mathcal{E}$.

sampling is assumed. Note that this ideal can be achieved in practice for two-dimensional and small three-dimensional catalogues via the algorithm of Szapudi (1998), however, the present simulation is too large for this method.

The smallest scale considered is only 2.4 times larger than the softening length $\lambda_\epsilon = 100\,h^{-1}$ kpc. As discussed extensively in Colombi, Bouchet & Hernquist (1996), contamination by softening restricts the validity of the simulation on small scales. For spherical cells of radius R, at least $R \gtrsim 4\lambda_\epsilon$ should hold. For the cubic cells of the present simulation this condition translates to $\ell \gtrsim 6.5\lambda_\epsilon \simeq 0.65\,h^{-1}$ Mpc. Thus the two smallest cell sizes, i.e. the two leftmost points could be contaminated by softening, a fact that should be borne in mind, especially when comparing with theoretical calculations which employ models motivated by dynamics. On the other hand, for statistical purposes the dynamics can be ignored and the simulation can be regarded as a set with prescribed statistics. Then the possible contamination is irrelevant at the level of the approximations taken in the next sections.

Another possible source of contamination could be, in principle, the anticorrelation introduced by the glass initial positions. The effect of this is, however, extremely small as evidenced by the measurement of $\bar{\xi}$ shown below.

Fig. 2 displays the average correlation function $\bar{\xi}$ as a function of scale. By definition

$$\bar{\xi} \equiv \frac{1}{v^2}\int_v \mathrm{d}^3 r_1\,\mathrm{d}^3 r_2\,\xi(|r_1 - r_2|), \tag{5}$$

where $\xi(r)$ is the two-point correlation function. In practice, it is obtained as the variance of the counts-in-cells, corrected for discreteness effects automatically via the use of factorial moments (e.g. see SQSL and Appendix A for the detailed description of the method used in this paper to obtain the cumulants including the variance from counts-in-cells). The measured $\bar{\xi}$ is compared with linear theory (dots) and with the non-linear Ansatz of Hamilton et al. (1991) improved by PD (dashes). As expected, the agreement with linear theory in the regime $\bar{\xi} \lesssim 1$ is excellent, even on the largest scales where the anticorrelations introduced by the glass initial condition could cause contamination. The two leftmost points are slightly below the dashes, because of softening effects as discussed above, otherwise the results are in perfect accord with theory.

Fig. 3 plots the extracted cumulants, S_Ns, against $\bar{\xi}$. They are compared with predictions of various models, including perturbation theory (PT, dots). By definition (e.g. Balian & Schaeffer 1989a)

$$S_N = N^{N-2} Q_N \equiv \bar{\xi}_N / \bar{\xi}^{N-1}, \tag{6}$$

where $\bar{\xi}_N$ is the N-point correlation function averaged over a cell:

$$\bar{\xi}_N = \frac{1}{v^N}\int_v \mathrm{d}^3 r_1 \cdots \mathrm{d}^3 r_N \xi_N(r_1, \ldots, r_N). \tag{7}$$

Perturbation theory predictions have been calculated for spherical cells by Juszkiewicz, Bouchet & Colombi (1993) for S_3 and extended to arbitrary order by Bernardeau (1994):

$$S_N(\ell) = f_N(\gamma_1, \ldots, \gamma_{N-2}), \tag{8}$$

$$\gamma_i \equiv \frac{\mathrm{d}^i \log \bar{\xi}}{(\mathrm{d}\log \ell)^i}. \tag{9}$$

For example

$$S_3 = \frac{34}{7} + \gamma_1, \tag{10}$$

$$S_4 = \frac{60712}{1323} + \frac{62}{3}\gamma_1 + \frac{7}{3}\gamma_1^2 - \frac{2}{3}\gamma_2. \tag{11}$$

The dots on Fig. 3 assume $\gamma_i = 0$, $i \geq 2$. While this is incorrect, in principle, for a scale-dependent spectrum such as τCDM, the long dashes on the left-hand panels prove that the contribution of γ_2 is insignificant. Higher order γ_i terms, as discussed also by Baugh, Gaztañaga & Efstathiou (1995), have an even smaller effect and can be rightly neglected.

PT predictions are accurately fulfilled in the weakly non-linear regime. This confirms again numerous earlier works (see, e.g. Juszkiewicz et al. 1993, 1995; Bernardeau 1994; Baugh et al. 1995; Gaztañaga & Baugh 1995; SQSL). In fact the textbook quality agreement with PT demonstrates the accuracy of the τCDM Hubble Volume simulation.

The dashes give the predictions obtained from extended perturbation theory (EPT, Colombi et al. 1997; see also Szapudi, Meiksin & Nichol 1996 for EPT applied to galaxy data, and Scoccimarro & Frieman 1998 for 'hyperextended' perturbation theory). EPT assumes that the same forms of the higher order moments are preserved in the highly non-linear regime. There γ_1 above is simply an adjustable parameter without any particular

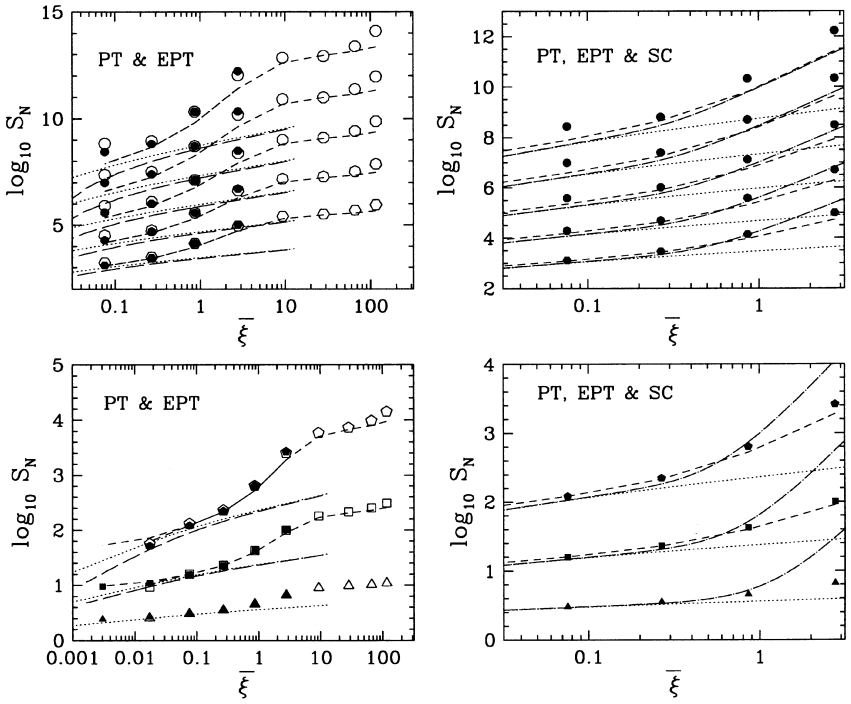

Figure 3. The cumulants $S_N \equiv \bar{\xi}_N/\bar{\xi}^{N-1}$ as functions of $\bar{\xi}$ compared to various theoretical models. The left-hand panels show the full dynamic range, while the right-hand ones concentrate on the transition to the non-linear regime. The models considered are perturbation theory (dots on all panels and long dashes on left panels), extended perturbation theory (short dashes) and one loop perturbation theory based on the spherical model (dots–long dashes on right panels). The upper and the lower panels give S_N for $6 \le N \le 10$ and $3 \le N \le 5$ respectively (the value of S_N increases with order N). The convention for the symbols is the same as in Fig. 2. Note that the right-hand panels show only the measurements in the full simulation $\mathcal{E}$.

meaning, i.e.

$$\gamma_{1,\mathrm{eff}} = \gamma_1(S_3) = S_3 - \frac{34}{7}, \tag{12}$$

where S_3 is the measured one. With this value of γ_1 the S_Ns, $N \ge 4$, can be computed using equation (8) (with $\gamma_i = 0$, $i \ge 2$). The dashed curve matches the measurements quite well even in the highly non-linear regime thereby reconfirming the efficiency of EPT (see also SQSL). The agreement is not expected to be absolutely perfect from this Ansatz: on Fig. 3, EPT tends to underestimate slightly the measured values of S_N when $1 \lesssim \bar{\xi} \lesssim 10$.

The dynamic range in the upper left panel of Fig. 3 is narrower than in the lower left panel: on large scales the agreement between PT and measurement becomes less accurate for the S_Ns, especially if N is large. This might be related to transients owing to the initial setup of the particles on a glass perturbed by using the Zel'dovich approximation. On the one hand, the transients related to pure Zel'dovich should decrease the value of the S_Ns (e.g. Juszkiewicz et al. 1993; Scoccimarro 1998) while, on the other hand, the anticorrelations resulting from the glass could have the opposite effect by decreasing $\bar{\xi}^{N-1}$ more than $\bar{\xi}_N$. Although this problem was not examined in detail, the glass contamination on $\bar{\xi}$ appears to be inconsequential. Alternatively, finite volume effects can degrade the high-N tail of the CPDF (e.g. Colombi, Bouchet & Schaeffer 1994; CBS; Colombi et al. 1996). In addition, it is worth re-emphasizing that the two rightmost points are prone to errors caused by softening as discussed earlier.

The right-hand panels of Fig. 3 zoom in on the transition between the weakly and highly non-linear regime. For comparison,

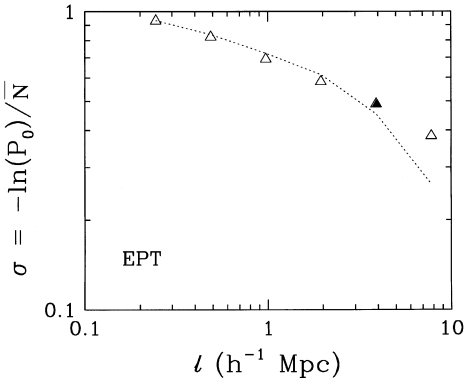

Figure 4. The scaling function $\sigma \equiv -\ln(P_0)/\bar{N}$, compared with extended perturbation theory (dots). The convention for the symbols is the same as in Fig. 2. Note that on the largest scales we measure $P_0 = 0$, and thus no points are plotted. For the direct measurement in $\mathcal{E}$ there is no empty cell with $\ell = 7.8\,h^{-1}$ Mpc because of our insufficient sampling.

PT (with $\gamma_i = 0$, $i \ge 2$, dots), EPT (dashes) and the one loop perturbation theory of Fosalba & Gaztañaga (1998) (dots–long dashes) are displayed. The last model yields agreement with the extracted values of S_N for $\bar{\xi} \lesssim 1$, or even larger when the order N is high enough (see upper right panel). This affirms the success of one-loop perturbation theory (see also Lokas et al. 1996; Scoccimarro et al. 1998). Interestingly, EPT produces almost identical results to the spherical model when $\bar{\xi} \lesssim 1$.

Finally, Fig. 4 shows $\sigma = -\ln(P_0)/\bar{N}$ as a function of scale, compared with EPT predictions. By definition (White 1979;

Balian & Schaeffer 1989a; see also Appendix A)

$$\sigma = \sum_{N=1}^{\infty} (-1)^{N-1} \frac{S_N}{N!} (\bar{N}\bar{\xi})^{N-1}, \tag{13}$$

where $\bar{N}$ is the average count in a cell. This function is thus sensitive to low order statistics when $N_c \equiv \bar{N}\bar{\xi} \ll 1$, and to high-order statistics when $N_c \gg 1$. According to Fig. 4, EPT is an accurate Ansatz on small scales where σ is close to unity and is dominated by low order S_N. It is a less precise approximation on the largest scales probed, as expected. Indeed, the rightmost point of Fig. 4 corresponds to where $\bar{\xi} \simeq 1$ in Fig. 3. There EPT increasingly underestimates the S_Ns when N is high. Note the remarkable power-law behaviour of $\sigma \propto \ell^{-D_0}$, $D_0 \simeq 0.25$, in agreement with the predictions of the scaling model of Balian & Schaeffer (1989a). This reflects a non-trivial (multi)fractal particle distribution (Balian & Schaeffer 1989b) with a Hausdorff dimension D_0. Such behaviour was found in a standard CDM model by Bouchet, Schaeffer & Davis (1991). Subsequently, the fractal distribution with $D_0 \simeq 0.5$ was established by Colombi, Bouchet & Schaeffer (1992).

4 THE COSMIC ERROR

In the previous section we demonstrated that good agreement was obtained comparing measurements made on the τCDM Hubble Volume data set with previous work regarding higher order clustering statistics. Having established the accuracy of the data set this section concentrates on the the determination of cosmic errors and their comparison to the available theoretical predictions, where possible. In Section 4.1 we summarize analytic calculations of the cosmic errors and their cross-correlations. From this follows a systematic study of the experimental cosmic error of low-order statistics, i.e. factorial moments F_k, $1 \le k \le 4$ (Section 4.2) and cumulants $\bar{\xi}$, S_3 and S_4 (Section 4.3) together with a thorough comparison with the theoretical predictions. Also in Section 4.3 we discuss the cosmic bias of the cumulants. Then the void probability and its scaling function σ are explored (Section 4.4) followed by the cosmic error on the CPDF itself (Section 4.5). Finally, in Section 4.6, there is a preliminary investigation of the cosmic cross-correlations of factorial moments and cumulants.

In all subsequent figures, except for the cross-correlations, there are error bars plotted on the symbols corresponding to measurements resulting from the finite number of realizations $C_\mathcal{E} = 4096$. These measurement errors, proportional to $1/\sqrt{C\varepsilon}$ (SC), are negligible for our simulation, and the error bars are smaller than the size of the symbols in most cases. As discussed in Section 1, we neglect the cosmic error on the determination of the cosmic error (which results from the finite size of the Hubble Volume itself) because in practice it is insignificant.

4.1 Cosmic error: theoretical predictions

Before making any comparison with the analytic predictions, we outline the main ideas in CBS, SC and SCB – more details can be found in these papers. Spherical cells of radius ℓ are assumed throughout for simplicity.

The bivariate CPDF $P_{N,M}(\ell, r)$ is the probability of finding N and M points in two cells of size ℓ at distance $r = |r_1 - r_2|$ from each other. According to SC the cosmic error is computed via a double integral of $P_{N,M}(\ell, r)$ over r_1, and r_2, conveniently split according to whether the cells overlap or not.

(i) *Overlapping cells* ($r \lesssim 2\ell$) : give rise to the discreteness and edge effect errors (see Section 1). The locally Poissonian assumption (CBS, SC) enables the approximate representation of the generating function $P(x, y)$ for overlapping cells by using only the monovariate generating function $P(x)$, i.e. the calculation depends on $\bar{\xi}$, S_N, $N \ge 3$ and the average count $\bar{N}$.

(ii) *Disjoint cells* ($r \gtrsim 2\ell$) : generate the finite volume error (see Section 1). To simplify the writing of $P_{N,M}(\ell, r)$, the distance r is assumed to be large enough compared to the cell size such that the bivariate CPDF can be Taylor expanded (to first order) in terms of $\xi(r)/\bar{\xi}$. This approximation is surprisingly accurate even when the cells touch each other (Szapudi, Szalay & Boschán 1992; B96). Three models are used: two particular but still quite general forms of the hierarchical model, SS and BeS, introduced by Szapudi & Szalay (1993a, hereafter SSa, 1993b) and by Bernardeau & Schaeffer (1992), respectively, and PT (B96). See SC and SCB for more details. The former two models depend only on monovariate statistics, i.e. on $\bar{\xi}$ and S_N, $N \ge 3$ and $\bar{N}$. PT on the other hand is expressed in terms of γ_i, $\bar{\xi}$ and $\bar{N}$ (B96). In principle, PT is accurate only in the weakly non-linear regime, for which it was originally designed, but it can be extended to the non-linear regime as well: for monovariate distributions, EPT was proposed by Colombi et al. (1997), as discussed and tested versus measurements in Section 3. This Ansatz can actually be naturally generalized to the bivariate CPDF (Szapudi & Szalay 1997, SCB). Our version, denoted by E^2PT, takes the measured (non-linear) value for $\bar{\xi}$, $\gamma_{1,\mathrm{eff}}$ from equation (12) and it assumes, as EPT, $\gamma_i = 0$ for $i \ge 2$.

Except for the error on the void probability and its scaling function σ detailed in CBS, the theoretical results shown in this section were computed to leading order in v/V, where v is the cell volume and $V = L^3$ is the sample volume.

The calculation of the error on a statistics of order k depends on $\bar{N} \equiv F_1$, $\bar{\xi}$, $\bar{\xi}(\hat{L})$, the average of the correlation function over the survey (see below) and S_N, $3 \le N \le 2k$. PT is determined by γ_i, $i \le 2k - 2$ (Section 3) and E^2PT by $\gamma_{1,\mathrm{eff}}$ as explained above. In all cases, we use the measured value of $\bar{N}$. Other parameters are chosen as follows.

(a) *PT:* linear theory is employed to compute $\bar{\xi}$ and $\bar{\xi}(\hat{L})$ (the catalogue is assumed to be spherical to simplify the calculation of integral 16 below) while higher order statistics are evaluated according to equation (8) with $\gamma_i = 0$, $i \ge 2$.

(b) *Other models:* the experimental $\bar{\xi}$ is used (open symbols on Fig. 2). The quantity $\bar{\xi}(\hat{L})$ is computed numerically with the non-linear Ansatz of PD discussed in Section 3 (assuming that the catalogue is spherical). For the S_Ns, the measurements (open symbols on the left panels of Fig. 3) are used for $\ell \le 15\, h^{-1}$ Mpc. On larger scales, EPT is more appropriate to determine S_N, $N \ge 4$: the increasing inaccuracy of the S_Ns on large scales and for large N require this procedure. It is justified all the more since, when $\bar{\xi} \lesssim 0.27$, EPT matches quite well to the PT predictions (see Fig. 3).

There is a subtlety worth mentioning which concerns the finite volume error, proportional to the integral

$$\bar{\xi}(\hat{L}) = \frac{1}{\hat{V}} \int_{r_{12} \ge 2\ell} \mathrm{d}^3 r_1\, \mathrm{d}^3 r_2\, \xi(|r_1 - r_2|). \tag{14}$$

To leading order in v/V, this integral reads (CCDFS)

$$\bar{\xi}(\hat{L}) = \bar{\xi}_0(\hat{L}) - \frac{8v}{\hat{V}} \bar{\xi}_1(2\ell), \tag{15}$$

with

$$\bar{\xi}_0(\hat{L}) = \frac{1}{\hat{V}^2}\int_{r_1,r_2\in\hat{V}} \mathrm{d}^3 r_1\,\mathrm{d}^3 r_2 \xi(|r_1 - r_2|), \tag{16}$$

$$\bar{\xi}_1(\ell) \equiv \frac{1}{v}\int_{r\le\ell} 4\pi r^2 \xi(r)\,\mathrm{d}r. \tag{17}$$

In the above equations, $\hat{V}$ corresponds to the volume covered by cells of volume v included in the catalogue.

The next to leading order correction, $\bar{\xi}_1$, can be identified as a negligible correction to the edge effects for most practical purposes. Although it did not make a significant difference, we included this correction none the less.

4.2 Cosmic error: factorial moments

Fig. 5 presents the cosmic error measured for the factorial moments F_k, $1 \le k \le 4$. By definition

$$F_k \equiv \langle (N)_k \rangle \equiv \langle N(N-1)\cdots(N-k+1)\rangle = \sum_N (N)_k P_N. \tag{18}$$

The factorial moments directly estimate the moments of the underlying continuous density field: $F_k = \bar{N}^k \langle \rho^k \rangle$ where $\bar{N} = F_1$ is the average count (e.g. SSa). On Fig. 5, the dotted, dash, long dash and dotted-long dash curves correspond to SS, BeS, E^2PT and PT.

All the models converge and agree quite well with the measurements on large scales $\ell \gtrsim \ell_0 \simeq 7.1\,h^{-1}$ Mpc, as expected, since PT predictions should be valid. In contrast, on small scales $\ell < \ell_0$ the models overestimate slightly the numerically obtained error, E^2PT being the most accurate. It is worth remembering that the leftmost two points may be contaminated by smoothing effects and should not be over-interpreted. Nevertheless, the decrease of precision on small scales suggests that our assumptions (i) or (ii) in Section 4.1 are becoming more and more approximate in the non-linear regime, i.e. either the local Poisson assumption or the particular hierarchical decompositions lose their accuracy. To test this idea the contribution of overlapping cells (edge + discreteness effects) were separated from the contribution of disjoint cells (finite volume effects), as shown respectively as solid and dash–long dash curves on Fig. 6, which concentrates on E^2PT (long dashes). Note that the solid curve represents the SS and BeS models as well. Finite volume effects appear to dominate on small scales because our subsamples are dense enough to suppress the discreteness error as expected (SC). This pinpoints assumption (ii) as the source of inaccuracy. Note that naively one would suspect additional loss of precision in the Taylor expansion of the bivariate CPDF. However, the finite volume error is a double integral over all the cells included in the catalogue and separated by more than 2ℓ. The contribution of close cells is small, especially when ℓ/L is small. Thus E^2PT itself appears to break down in the non-linear regime (SS and BeS are even less accurate), at least for the particular experiment we are analysing. Despite that EPT itself fares quite well (Fig. 3), its simplest natural extension to bivariate distributions, E^2PT, is less accurate, as noticed earlier by Szapudi & Szalay (1997) in connection with the cumulant correlators of the APM (automated plate measurement) galaxy catalogue. However, the accuracy of the calculation based on E^2PT should be adequate for most practical uses, and future work on the representation of the bivariate distribution in the highly non-linear regime will result in increased precision.

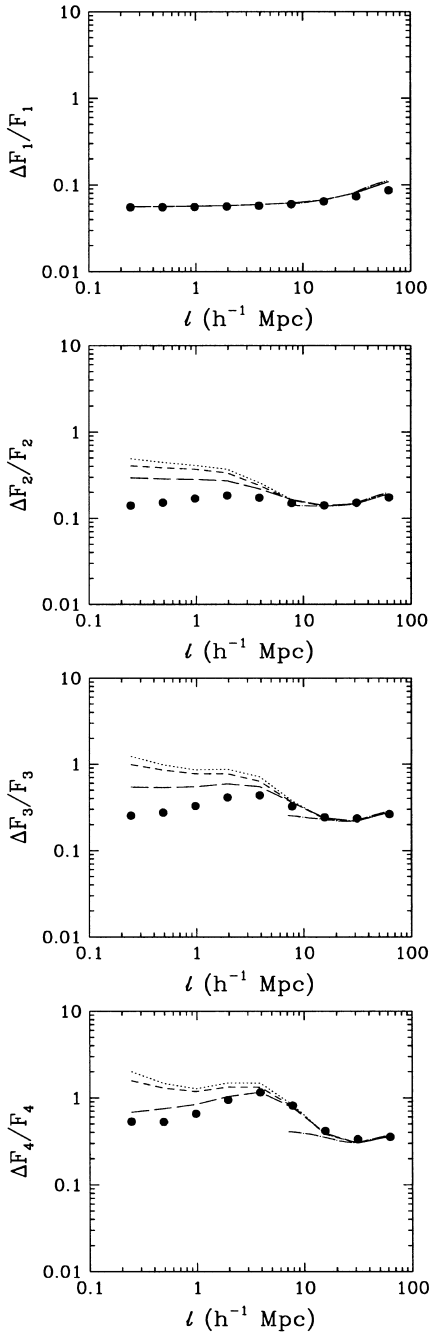

Figure 5. The cosmic error $\Delta F_k/F_k$ as a function of scale. Each panel corresponds to a value of k. The dots, dashes, long dashes, dot–long dashes correspond respectively to the SS, BeS, E^2PT and PT models. PT is shown only in its expected range of validity, $\ell \gtrsim \ell_0$, where ℓ_0 is the correlation length defined by $\bar{\xi}(\ell_0) \equiv 1$. For $k = 1$, all the models give the same result. As discussed in the beginning of Section 4, there are error bars owing to the finite number of realizations $C_\mathcal{E} = 4096$, but they are so small that they do not show.

The solid curves in Fig. 6 represent the main contribution of the cosmic error on large scales. Here, as expected (SC), the cosmic error is dominated by edge effects. Despite the fact that theoretical predictions were determined to leading order in v/V and the largest scale considered is $\ell = L/2$, i.e. $v/V = 1/8$, the agreement

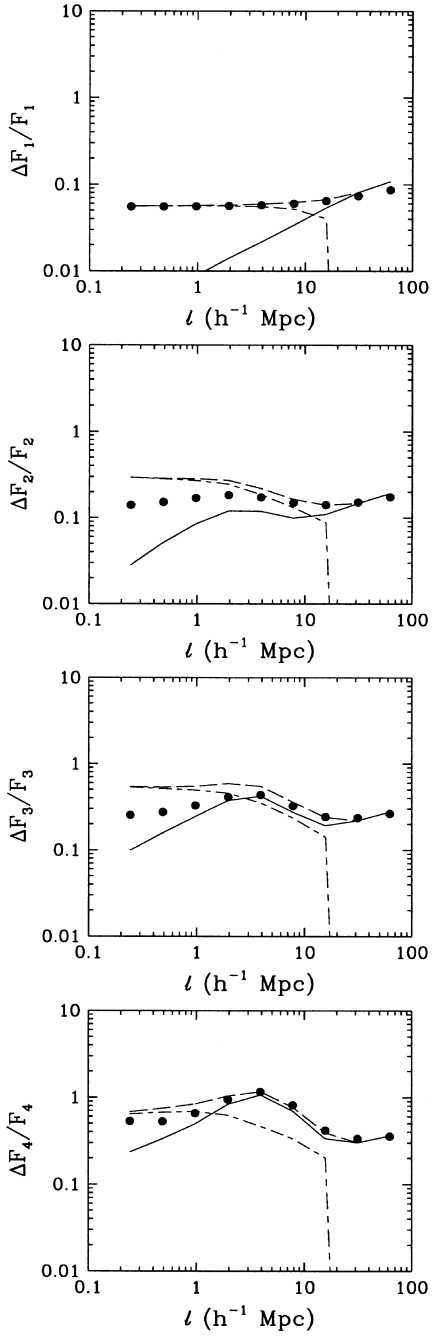

Figure 6. Same as in Fig. 5, but now the long dashed, dashed–long dashed and solid curve correspond respectively to the E^2PT model, the finite volume contribution and the edge + discreteness contribution. Note the sudden cut-off at large scales for the finite volume error, in agreement with equation (15). Without the $8(v/V)\xi_1(2\ell)$ correction, the cut-off would not show up, but this would not significantly change the total error.

between theory and measurement is surprisingly good. CCDFS have computed the next to leading order contribution proportional to the perimeter ∂V of the survey. With this correction, which increases the cosmic error, especially on the largest scales, next to leading order theory would be inferior to the leading order one. The reason is that the calculation of CCDFS assumes a perimetric curvature radius much larger than the cell size. This assumption, which is useful for deep galaxy surveys with small sky coverage, obviously fails for a compact catalogue such as this one where the cell size ℓ becomes comparable to L.

4.3 Cosmic error and cosmic bias: variance and cumulants

So far only the full moments F_k have been examined. The cumulants $\bar{\xi}$ and S_N, however, are the more physically motivated quantities. But the statistics of these is complicated by the fact that they are ratios. For example (see Appendix A)

$$\bar{\xi} = F_2/F_1^2 - 1. \tag{19}$$

As is well known in statistics (e.g. HG, SCB) $\langle A/B\rangle \neq \langle A\rangle/\langle B\rangle$. In other words, the estimator

$$\bar{\tilde{\xi}} = \tilde{F}_2/\tilde{F}_1^2 - 1 \tag{20}$$

is biased. Note that this is a general feature for any statistic constructed from unbiased estimators in a non-linear fashion (e.g. SCB). However, SCB showed theoretically that the *cosmic bias* defined in Section 1, given here by

$$b_{\bar{\xi}} \equiv (\langle\bar{\tilde{\xi}}\rangle - \bar{\xi})/\bar{\xi}, \tag{21}$$

is of same order of $(\Delta\bar{\xi}/\bar{\xi})^2$ in the regime $\Delta\bar{\xi}/\bar{\xi} \ll 1$. Similar reasoning applies to the S_Ns. Thus leading order theoretical calculations neglect the bias. This can be done safely in the domain of validity of the perturbative approach used to expand a non-linear combination of biased estimators. A reasonable criterion proposed by SCB for this domain is that the cosmic bias be small compared to the relative cosmic error which itself should be small compared to unity. For an arbitrary (possibly biased) statistic A this reads as

$$b_A \ll \Delta A/A \ll 1. \tag{22}$$

The left panels of Fig. 7 are analogous to Fig. 5 and show the measured cosmic error as a function of scale for the biased estimators of $\bar{\xi}$, S_3 and S_4. The middle panels show the absolute value of the cosmic bias (open symbols) compared to the cosmic error (filled symbols). For additional clarity, the cosmic bias is plotted in linear coordinates as well in the right-hand panels.

It is interesting first to compare the cosmic error for factorial moments and cumulants of same order. The discreteness error is negligible for the scaling regime and the statistics considered here. The cumulants fare better/worse than the factorial moments in the non-linear/weakly non-linear regimes, respectively. The finite volume error, dominating on small scales, is the limiting factor for factorial moments, while the edge effect error, dominating on large scales, drives the errors of the cumulants. This is in full accord with the predictions of SCB which can be consulted for more details.

The theoretical models on Fig. 7 use the analytic calculations of SCB and are computed analogously to Fig. 5, as explained in Section 4.1. E^2PT only is presented in the middle and right-hand panels. Again, it is worth remembering that the leftmost points are dangerously close to the limit of possible contamination from artificial smoothing effects introduced by the force softening.

For the variance $\bar{\xi}$, the theory systematically overestimates the errors and the cosmic bias, except for the latter on large scales. This is not at all unexpected in light of the previous findings on small scales, where the three models SS, BeS and E^2PT lose precision. In the weakly non-linear regime, $\ell > \ell_0 = 7.1\,h^{-1}$ Mpc, where perturbation theory is valid, this is somewhat disappointing. However, the dynamic range is limited by criterion (22), which is

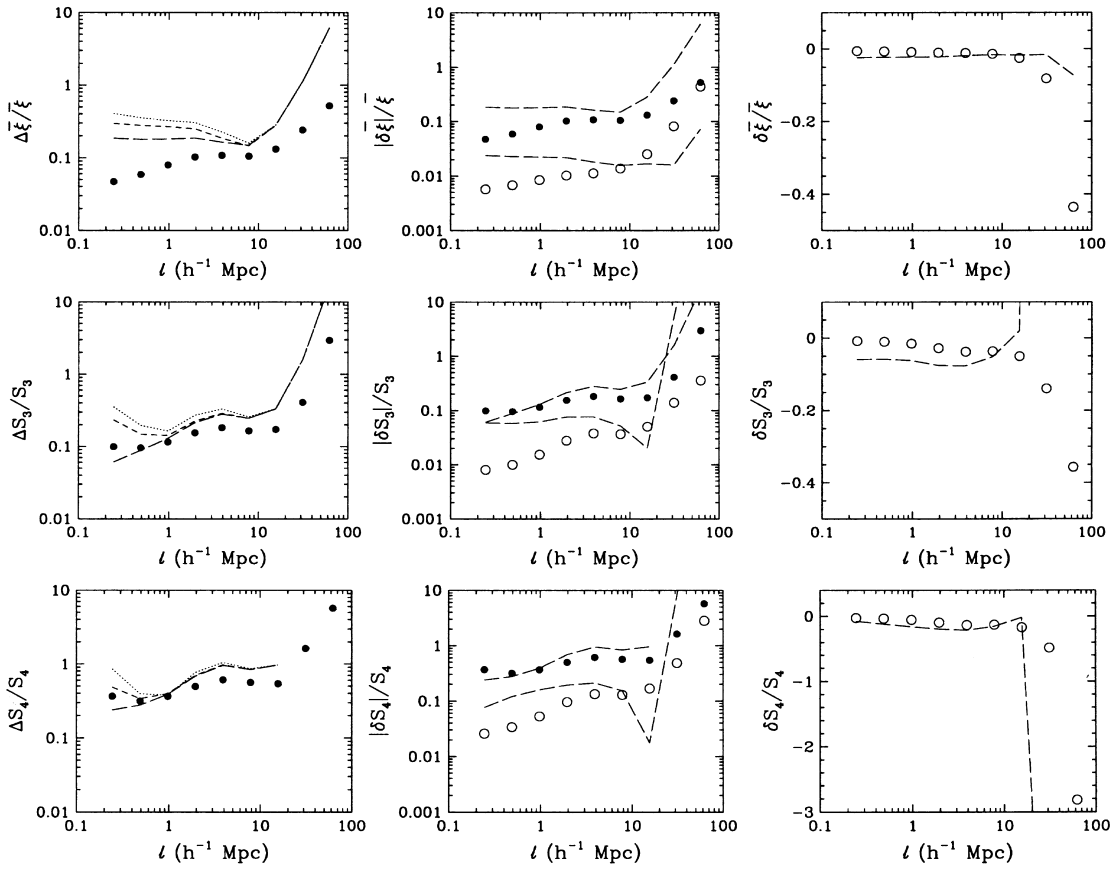

Figure 7. Same as Fig. 5 for the average correlation function (top row of panels), and the cumulants S_3 (middle row of panels) and S_4 (lower row of panels). The cosmic bias is plotted both in logarithmic coordinates (middle column of panels) and linear coordinates (right column of panels). The filled and the open symbols correspond to the cosmic error and the cosmic bias respectively. The theory breaks down on large scales for S_4 shown in the bottom left-hand panel. In this regime, the leading order calculation gives negative $(\Delta S_4/S_4)^2$ (see SCB). The theory result for the cosmic bias is shown for the E^2PT model only. In the middle column panels, there are two long-dashed curves: each one of which should be compared with the closest symbols overall, corresponding either to the cosmic error (filled) or the cosmic bias (open).

hardly, if at all, fulfilled here. Hence the leading order perturbative approach is likely to be insufficient.

For higher order statistics S_3 and S_4, the theory again tends to overestimate the amplitude of the measured cosmic bias on small scales. On large scales, where the predicted $|b_{S_k}|$ presents a sudden turn-up, condition (22) breaks down, thus the theory is inapplicable. The measured cosmic errors, on the other hand, are in accord with the theory within the range of its validity. The agreement on small scales is even better for $\Delta S_k/S_k$ than for $\Delta F_k/F_k$, $k = 3, 4$. This, however, should not be over-interpreted, as it is probably a coincidence owing to cancellation effects of the ratios $S_3 = \bar{\xi}_3/\bar{\xi}^2$ and $S_4 = \bar{\xi}_4/\bar{\xi}^3$.

The cosmic bias is always negative (right-hand panels of Fig. 7), i.e. the biased estimators tend to underestimate real values (SCB; HG). In this particular experiment, the measured cosmic bias is always dominated by the measured cosmic error as predicted by the perturbative approach, except for the largest scales. Here the cosmic bias can become of same order as the cosmic error. HG suggested that the cosmic bias should be corrected for when measuring cumulants. Whether this makes sense depends on the magnitude of the *cosmic skewness*, i.e. the skewness of the cosmic distribution function itself. This will be discussed in more detail by Paper II. However, it is worth noting that function $\Upsilon(\tilde{A})$ is positively skewed and that its maximum corresponds to the most likely measurement. This is in general smaller than the average, $\langle\tilde{A}\rangle$. Thus, as pointed out already by SC, the measured value $\tilde{A}$ in a finite sample is *likely* to underestimate the real value A *even if* $\tilde{A}$ *is unbiased*. If the cosmic skewness and/or the cosmic variance are large compared to the cosmic bias, it is pointless to correct for the cosmic bias. Either of the above is true for most surveys, including the upcoming wide-field surveys such as the 2dF and SDSS, thus bias-corrected estimators are unlikely to be useful in the future.

4.4 Cosmic error and cosmic bias: void probability and scaling function

The upper panel of Fig. 8 shows $\Delta P_0/P_0$ as a function of scale compared to the prediction of CBS (long dashes), with the finite volume error contribution (dashes–long dashes) and with the edge + discreteness contribution (solid curve). The agreement between theory and prediction is excellent.

The lower panel of Fig. 8 corresponds to the scaling function σ. As for $\bar{\xi}$ and S_N, the indicator $\tilde{\sigma} = -\ln(\tilde{P}_0)/\tilde{N}$ is biased. This bias (open symbols) is of order $(\Delta\sigma/\sigma)^2$ and can be neglected.[4] The agreement between theory and measurement is less impressive than for P_0, but this is mostly owing to the difference of dynamic

[4] The theoretical and measured errors displayed on the bottom part of Fig. 8 correspond to the biased indicator.

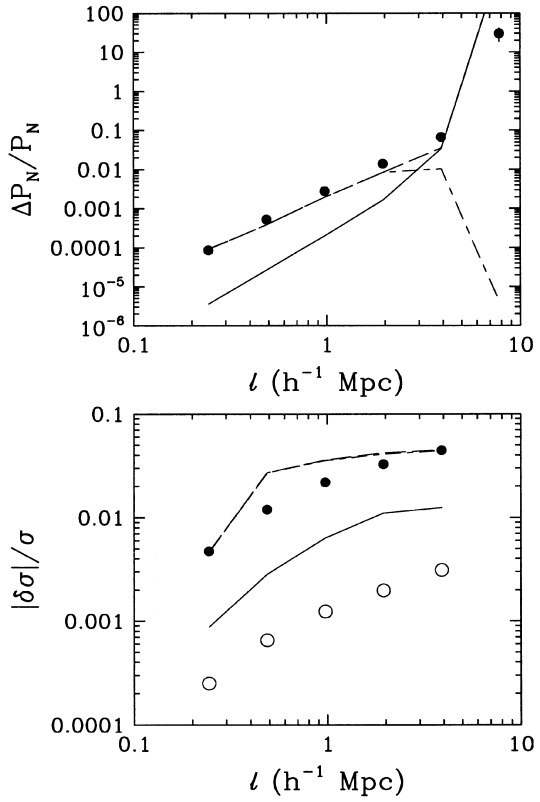

Figure 8. The cosmic error of the void probability function P_0 (upper panel) and on the scaling function $\sigma = -\ln(P_0)/\bar{N}$ (lower panel). The measurements (filled symbols) are compared with the theoretical predictions of CBS (long dashes). The finite volume error contribution is drawn with short dash–long dash and the edge + discreteness effects contribution with solid lines. The available scaling range is limited by the fact that on large scales the measured void probability is zero. For $\ell \simeq 7.8\,h^{-1}$ Mpc (upper right point on upper panel), the void probability cancels from time to time in the subsamples $\mathcal{E}_i$. As a result, it is possible to compute the unbiased function σ but the estimated cosmic error on the biased estimator $\bar{\sigma}$ is infinite. The open symbols in the lower panel correspond to the measured cosmic bias in σ. It is positive and much smaller than the cosmic error. It can be neglected for all the relevant dynamic range in the experiment considered here.

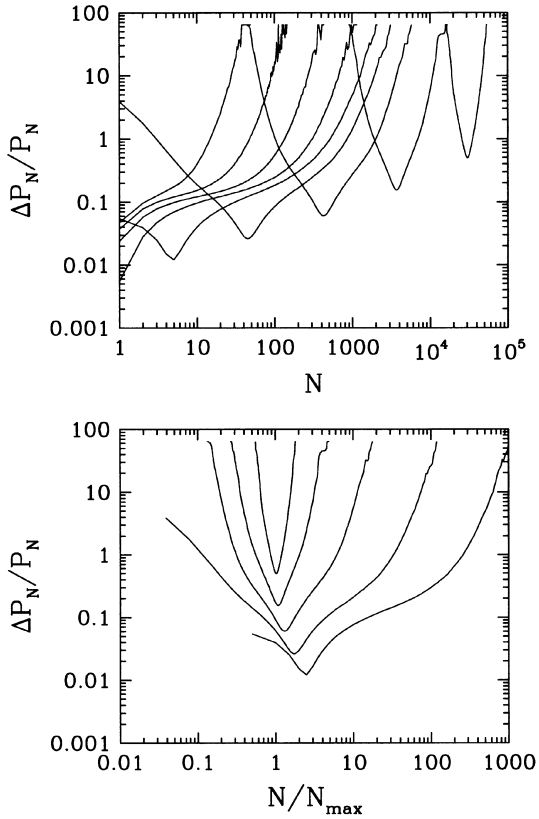

Figure 9. The cosmic error $\Delta P_N/P_N$ in the CPDF as a function of N (upper panel) and as a function of $N/N_{\max}$, where $N_{\max}$ is the value of N for which P_N is maximum (lower panel). In the lower panel, only the scales large enough so that $N_{\max} > 1$ are displayed.

range covered by the error in the upper and the lower panels of Fig. 8. Moreover, the calculation of $\Delta\sigma/\sigma$ by CBS is only approximate and could certainly be improved (see the discussion in CBS).

The error bars about σ are quite small: nearly an order of magnitude smaller than in Figs 5 and 7. According to equation (13), σ reflects the low-order statistics when $N_c = \bar{N}\bar{\xi} \ll 1$ ($\sigma \lesssim 1$ in Fig. 4) and the high order statistics when $N_c \gg 1$ ($\sigma < 1$). From the point of view of the errors, function σ is an excellent higher-order indicator (as discussed earlier by CBS); it is better than the low-order factorial moments or cumulants, at least in the non-linear regime $\ell \lesssim \ell_0$. This fact alone unfortunately does not guarantee the usefulness of this statistic as various models of large scale structure formation could be degenerate with respect to the void probability. The thorough work of Little & Weinberg (1994) suggests that this is indeed the case. It is tempting, although dangerous, to extrapolate the results of their analysis to the function σ.

4.5 Cosmic error: counts-in-cells

The upper panel of Fig. 9 shows the cosmic error in the CPDF as a function of N for the various scales considered in $\mathcal{E}_i$. The scale increases with the x-coordinate of the upper right part of each curve. In the lower panel $\Delta P_N/P_N$ is represented in a similar manner as a function of $N/N_{\max}$, where $N_{\max}$ is the value of N for which P_N is a maximum. [We did not display the (small) scales corresponding to $N_{\max} = 0$ or $N_{\max} = 1$]. In agreement with intuition, the cosmic error reaches its minimum in the vicinity of $N \simeq N_{\max}$ and becomes increasingly large in the tails. Thus the shape of the CPDF near its maximum has the most power to constrain in terms of errors. Kim & Strauss (1998) have measured the cumulants S_3 and S_4 by fitting an Edgeworth expansion convolved with a Poisson distribution to the measured CPDF in the 1.2 Jy *IRAS* galaxy catalogue. According to their recipe, the best determined part of the CPDF near the maximum was kept for the fit. Their maximum likelihood approach uses a simple model for the cosmic error, but their method is promising. Its main weakness is the necessity to make a strong prior assumption for the shape of the CPDF. A natural consequence is that the estimated error bars on the measured cumulants are considerably smaller than with the standard methods.

4.6 Cosmic correlations

So far this section has dealt only with the second moment of the cosmic distribution function, i.e. with the cosmic errors. For a full

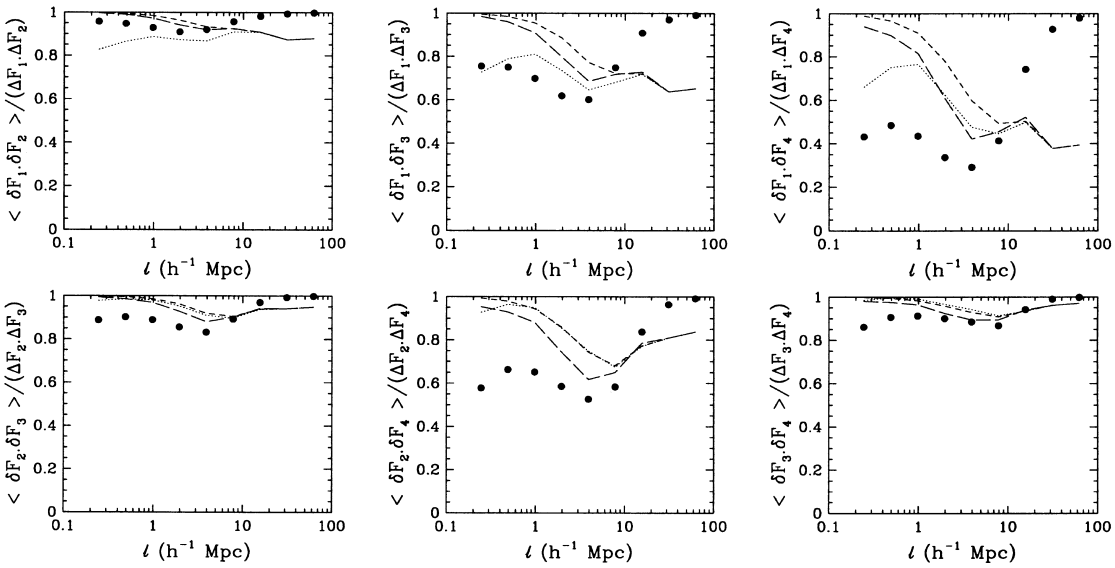

Figure 10. The measured cosmic cross-correlation coefficients of the factorial moments (symbols) are compared with the models SS (dots), BeS (dashes) and E^2PT (long dashes).

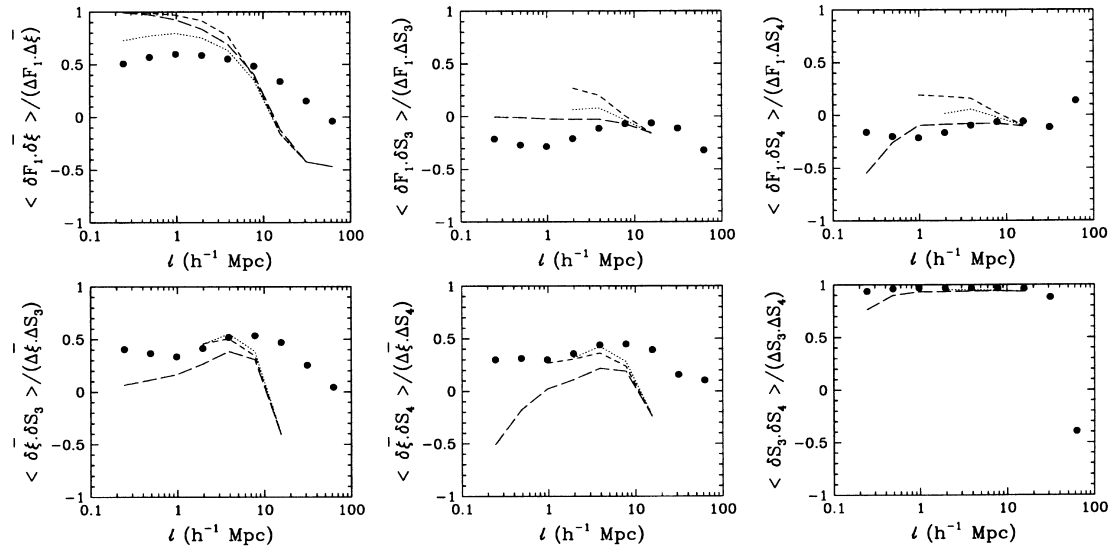

Figure 11. Same as in Fig. 10 but for the cosmic cross-correlation coefficients of the cumulants. The dynamic range for the theory is restrained by condition (24).

description in the Gaussian limit, however, the moments of the joint distribution function are needed. These moments form the cosmic (cross-correlation) matrix (SCB). It is defined as $\langle(\tilde{A}-\langle\tilde{A}\rangle)(\tilde{B}-\langle\tilde{B}\rangle)\rangle$, where $\tilde{A}$ and $\tilde{B}$ are any counts-in-cells related indicators, for example $A = F_k(\mathcal{C})$ and $B = F_{k'}(\mathcal{C}')$, or $A = \bar{\xi}(\mathcal{C})$ and $B = S_N(\mathcal{C}')$, etc. A detailed theoretical investigation can be found in SCB (for $\mathcal{C} = \mathcal{C}'$). By definition, for two statistics A and B, the correlation coefficient $-1 \le \rho \le 1$ reads as

$$\rho \equiv \frac{\langle\delta\tilde{A}\delta\tilde{B}\rangle}{\Delta A\Delta B} \equiv \frac{\langle(\tilde{A}-A)(\tilde{B}-B)\rangle}{\Delta A\Delta B}. \qquad (23)$$

The cosmic cross-correlation coefficient together with the errors form the full correlation matrix. The inverse of this is the central quantity for the joint probability distribution function in the Gaussian limit. As a preliminary numerical analysis, Figs 10 and 11 present the correlation coefficients as functions of scale ($\mathcal{C}' = \mathcal{C}$) for factorial moments and cumulants, respectively. As in Fig. 5, the dots, dashes and long dashes show the theoretical predictions given by the SS, BeS and E^2PT models, respectively, as computed by SCB. The computation of $\langle\delta\tilde{A}\delta\tilde{B}\rangle$ in equation (23) is analogous to that of the cosmic error (see SCB for more details). (For ΔA and ΔB, and to have completely self-consistent calculations, we take the theoretical results as well in equation 23).

The agreement between theory and measurement is less convincing for the cosmic cross-correlations than for the cosmic error. This appearance is partly results from the linear coordinates of the figures which emphasize deviations, but none the less are real.

On Fig. 10 there is a significant discord between theory and measurements for factorial moments in the middle-top, middle-bottom and top-right panels. On small scales, this result is quite natural: it is probably owing to the inaccuracy of the models SS,

BeS and E^2PT employed to describe the underlying bivariate distributions (Section 4.2). In the weakly non-linear regime, this discrepancy is apparently puzzling, since the predicted cosmic error matches perfectly the measurements (Fig. 5). The disagreement increases with $|k-l|$, where k and l are the corresponding orders. On large scales, the cross-correlations are dominated by edge effects leading to the suspicion that the local Poisson approximation (SC, Section 4.1) is becoming increasingly inaccurate with $|k-l|$.[5] Another although less likely possibility, is that the leading order approach in v/V is insufficient and higher order corrections are necessary to calculate cross-correlations. It would go beyond the scope of this paper to analyse in detail these effects which are left for future research.

For the cumulants, in addition to the above arguments, our perturbative approach to compute cross-correlation allows only a narrow dynamic range for analytic predictions, defined by criterion (22). In Fig. 11, this condition is chosen for practical purposes to be

$$|b_A| \leq \Delta A/A \leq 1. \tag{24}$$

This is necessary but not sufficient: the theory appears to disagree significantly with the measurements on large scales at the top-left, lower-left and lower-middle panels of Fig. 11.

Despite some of the discrepancies, the general features of the cross-correlations are well described by the theoretical predictions. For instance the cross-correlation between two statistics A_k and A_l decreases with the difference between the orders $|k-l|$ as predicted (SCB). In our particular experiment $\bar{N}$ is significantly correlated with $\bar{\xi}$, but only weakly (anticorrelated) with S_k, $k=3,4$. Similarly, $\bar{\xi}$ and S_3 are weakly, but S_3 and S_4 are strongly correlated. A detailed discussion on these effects can be found in SCB.

5 SUMMARY AND DISCUSSION

In this paper we have studied experimentally the properties of the moments of the cosmic distribution function of measurements $\Upsilon(\tilde{A})$, where $\tilde{A}$ is an indicator of a counts-in-cells statistic. For a thorough examination of $\Upsilon(\tilde{A})$ itself the reader is referred to Paper II also in this volume.

We examined the factorial moments F_k, the cumulants $\bar{\xi}$ and S_Ns, the void probability P_0, its scaling function, $\sigma \equiv -\ln(P_0)/F_1$ and the count-in-cells themselves P_N. $\Upsilon(\tilde{A})$ was measured in the largest available τCDM simulation divided into 4096 cubical subsamples. In each of these many subsamples, $\tilde{A}$ was extracted and its probability distribution function Υ was estimated with great accuracy. The main results of our analysis are the following.

(i) The measured count-in-cells in the whole simulation, in particular the cumulants S_N, are in excellent agreement with perturbation theory predictions in the weakly non-linear regime. This confirms the results of numerous previous investigations in an unprecedented dynamic range. The textbook quality agreement demonstrates the state-of-the-art accuracy of the simulation. Similarly, the measurements confirm extended perturbation theory (EPT) in the full available dynamic range $0.05 \lesssim \bar{\xi} \lesssim 50$, for S_N, $N \leq 10$. In addition one loop perturbation theory predictions based on the spherical model (Fosalba & Gaztañaga 1998) were found to be an excellent description of the measured S_N up to $\bar{\xi} \lesssim 1$.

(ii) The variance of Υ is the square of the expected cosmic error, ΔA, in the measurement of A in a subsample, identified with a realization of the local observed universe. The measurement of $\Delta A/A$, for $A = P_0, \sigma$, F_k and S_N appears to be globally in good accord with the theoretical predictions of Colombi et al. (1995), SC (1996) and SBC (1998a).

In the highly non-linear regime, the theoretical predictions of SC and SCB tend to overestimate the cosmic error slightly, except for the ratios $S_3 = \bar{\xi}_3/\bar{\xi}^2$ and $S_4 = \bar{\xi}_4/\bar{\xi}^3$. In the latter case, there are some cancellations and the agreement between theory and measurement is good, even on small scales, but this is probably a coincidence. Thus it appears that none of the three variants of the hierarchical model in SC and SCB, can give an accurate enough account of the non-linear behaviour of gravitational dynamics for the bivariate distribution functions.[6]

In the weakly non-linear regime, agreement between theory and predictions is excellent for the factorial moments, but less good for the cumulants, owing to the limitations of the perturbative approach used to expand such ratios.

None the less EPT yields the most precise overall agreement with theory for our particular experiment. On small scales $1\,h^{-1}\,\mathrm{Mpc} \lesssim \ell \lesssim 4\,h^{-1}\,\mathrm{Mpc}$, EPT overestimates the errors perhaps by a factor of two in the worst case.

(iii) In addition to the cosmic errors, the *cosmic bias*, b_A, was studied in detail as well. An estimator is biased when its ensemble average is different from the real value: $b_A \equiv \langle \tilde{A} \rangle / A - 1 \neq 0$. This is always the case when unbiased estimators are combined in a non-linear fashion to form a new estimator (SCB; HG), such as the cumulants.

In agreement with SCB, the measured cosmic bias is of order $(\Delta A/A)^2$ and thus negligible when the cosmic error is small. However, as for the errors, the theory tends to overestimate the bias in the non-linear regime. On large scales, where the cosmic bias becomes significant because of edge effects, the perturbative approach used by SBC to compute theoretical predictions is then outside of its domain of validity.

Note that in the regime where the cosmic bias is significant, the cosmic error is likely to be large. For instance, in the particular numerical experiment used in this paper, the cosmic bias was always smaller than the cosmic errors and in most cases negligible. Moreover, in the regime where the bias could be significant, the cosmic distribution function $\Upsilon(\tilde{A})$ is significantly positively skewed (Paper II). This implies that the measured $\tilde{A}$ is likely to underestimate the true value even for an unbiased estimator. The result is an *effective* cosmic bias, at most of order $\Delta A/A$. As already shown by SC, this effective bias can contaminate even unbiased estimators such as $\tilde{F}_k$ and $\tilde{P}_N$. As a consequence, it is pointless correcting for the cosmic bias, in contrast with the proposition of HG, unless it is done in the framework of a maximum likelihood approach which takes into account fully the effects of the shape of the cosmic distribution function.

(iv) To complete the analysis of second moments, a preliminary investigation of the cosmic correlation coefficients for factorial moments and cumulants was conducted. Together with the cosmic errors, these coefficients form the cosmic cross-correlation matrix

[5] This is not surprising: this approximation neglects local correlations. This is all the more inaccurate as the difference between the 'weights' given to two overlapping cells, i.e. $(N)_k$ and $(N)_l$ for factorial moments, increases.

[6] As discussed in Section 4.1, the analysis of the cosmic error indirectly probes the bivariate probability distribution function $P_{N,M}(r,\ell)$ of having N and M galaxies, respectively, in two cells of size ℓ separated by distance r (see, e.g. SC).

which underlies maximum likelihood analysis in the Gaussian limit.

Theoretical predictions of SBC give good qualitative account of the measured correlation coefficients, although they become increasingly approximate with the difference between the corresponding orders. This is likely to be a consequence of the local Poisson assumption (SC) employed for analytic predictions.

Provided that the Gaussian limit is reached in terms of the error distribution, the formalism of SC and SBC allows for a maximum likelihood analysis of the CPDF measured in three-dimensional galaxy catalogues. Two preliminary investigations are currently being undertaken. Szapudi, Colombi & Bernardeau (in preparation) reanalyse already existing joint measurements of F_1 and $\bar{\xi}$, and Bouchet, Colombi & Szapudi (in preparation) perform a likelihood analysis of the count-in-cells measured in the 1.2 Jy *IRAS* survey (Bouchet et al. 1993).

Paper II probes the domain of the Gaussian approximation for the cosmic distribution function, together with preliminary investigations for the bivariate cosmic distributions $\Upsilon(\tilde{A}, \tilde{B})$. As shown there, the Gaussian limit is reached when the relative cosmic error is small compared to unity. This is expected to hold for a large dynamic range in future large galaxy surveys such as the 2-degree Field Survey (2dF) and the Sloan Digital Sky Survey (SDSS) (Colombi et al. 1998).

Statistical analyses of weak lensing surveys are similar to counts-in-cells measurements (e.g. Bernardeau, Van Waerbeke & Mellier 1997; Jain, Seljak & White 1999; Mellier 1999). As a result, slight modification of the formalism of SC and SCB is fruitful to compute theoretical cosmic errors and cross-correlations (Bernardeau, Colombi & Szapudi, in preparation).

Finally, it is worth mentioning a few questions which were not addressed so far by the investigations presented in this paper. As light might not trace mass, the distribution of galaxies may be biased (not to be confused with the cosmic bias), and also realistic galaxy surveys are subject to redshift distortion. While the above results were obtained for the mass, note that the theory which served as a basis of comparison is quite general and was formulated to describe phenomenologically either the mass or the galaxies. It appears that there should be no qualitative changes introduced by biasing or redshift distortions (e.g. Szapudi et al., in preparation), thus the same theory can be used for the galaxies as for the mass, except perhaps with slightly different parameters, or underlying statistical models. In fact, two of the models (SS, BeS) were entirely motivated by the galaxy and not by the mass distribution; they are expected to be more accurate for realistic catalogues if used in a self-consistent fashion. The scaling properties underlying these models is even more accurate in redshift space, as is well known. EPT, on the other hand, was originally motivated by theoretical considerations of the mass distribution and numerical simulations (Colombi et al. 1997), and therefore it is no wonder that it is the most successful model for the mass (but see also Scoccimarro & Frieman 1998). None the less, even EPT was found to be a fairly good model for the galaxy distribution, and at least in the Edinburgh–Durham Southern Galaxy Catalogue (EDSGC) survey (Szapudi et al. 1996), a possible indication that galaxies approximately trace mass after all. In addition, it is worth mentioning that biasing models can be non-deterministic, i.e. stochastic in nature, but this again does not introduce anything new qualitatively which could not be handled in the framework of the theory of SCB. Finally, the theory outlined in this paper was contrasted against measurements in a τCDM simulation. However, the analytical framework is general enough to accommodate any cosmological model and there are no qualitative differences in this respect between different cosmologies with Gaussian initial conditions and hierarchical clustering. Thus repeating the same analysis for a different CDM-like cosmogony would be superfluous and inconsequential.

ACKNOWLEDGMENTS

The FORTRAN routine for computing S_N, $3 \leq N \leq 10$, using one loop perturbation theory predictions based on the spherical model was provided by P. Fosalba (see the right-hand panels of Fig. 3). We thank F. Bernardeau, P. Fosalba, C. Frenk, R. Scoccimarro, A. Szalay and S. White for useful discussions. It is a pleasure to acknowledge support for visits by IS and SC to the MPA, Garching and by SC to the Department of Physics, Durham, during which part of this work was completed. IS and AJ were supported by the PPARC rolling grant for Extragalactic Astronomy and Cosmology at Durham.

The Hubble Volume simulation data was made available by the Virgo Supercomputing Consortium (http://star-www.dur.ac.uk/~frazerp/virgo/virgo.html). The simulation was performed on the T3E at the Computing Centre of the Max-Planck Society in Garching. We would like to give our thanks to the many staff at the Rechenzentrum who have helped us to bring this project to fruition. The FORCE package (FORtran for Cosmic Errors) used for the error calculations in this paper is available on request from its authors SC and IS.

REFERENCES

Balian R., Schaeffer R., 1989a, A&A, 220, 1
Balian R., Schaeffer R., 1989b, A&A, 226, 373
Baugh C. M., Gaztañaga E., Efstathiou G., 1995, MNRAS, 274, 1049
Bernardeau F., 1994, A&A, 291, 697
Bernardeau F., 1996, A&A, 312, 11 (B96)
Bernardeau F., Schaeffer R., 1992, A&A, 255, 1 (Be5)
Bernardeau F., Van Waerbeke L., Mellier Y., 1997, A&A, 322
Bond J. R., Efstathiou G., 1984, ApJ, 285, L45
Bouchet F. R., Schaeffer R., Davis M., 1991, ApJ, 383, 19
Bouchet F. R., Strauss M. A., Davis M., Fisher K. B., Yahil A., Huchra J. P., 1993, ApJ, 417, 36
Colombi S., Bouchet F. R., Schaeffer R., 1992, A&A, 263, 1
Colombi S., Bouchet F. R., Schaeffer R., 1994, A&A, 281, 301
Colombi S., Bouchet F. R., Schaeffer R., 1995, ApJS, 96, 401 (CBS)
Colombi S., Bouchet F. R., Hernquist L., 1996, ApJ, 465, 14 (CBH)
Colombi S., Bernardeau F., Bouchet F. R., Hernquist L., 1997, MNRAS, 287, 241
Colombi S., Szapudi I., Szalay A. S., 1998, MNRAS, 296, 253
Efstathiou G., Davis M., Frenk C. S., White S. D. M., 1985, ApJS, 57, 241
Eke V. R., Cole S., Frenk C. S., 1996, MNRAS, 282, 263
Fosalba P., Gaztañaga E., 1998, MNRAS, 301, 503
Gaztañaga E., Baugh C. M., 1995, MNRAS, 273, L1
Hamilton A. J. S., Kumar P., Lu E., Matthews A., 1991, ApJ, 374, L1
Hockney R. W., Eastwood J. W., 1981, Computer Simulation Using Particles, McGraw Hill, New York
Hui L., Gaztañaga E., 1999, ApJ, 519, 622
Jain B., Seljak U., White S. D. M., 1999, preprint (astro-ph/9901191)
Jenkins A. et al., 1998, ApJ, 499, 20
Juszkiewicz R., Bouchet F. R., Colombi S., 1993, ApJ, 412, L9
Juszkiewicz R., Weinberg D. H., Amsterdamski P., Chodorowski M., Bouchet F. R., 1995, ApJ, 442, 39
Kim R. S., Strauss M. A., 1998, ApJ, 493, 39
Little B., Weinberg D. H., 1994, MNRAS, 267, 605

Lokas E. L., Juszkiewicz R., Bouchet F. R., Hivon E., 1996, ApJ, 467, 1
MacFarland T., Couchman H. M. P., Pearce F. R., Pichlmeier J., 1998, New Astron., 3, 687
Mellier Y., 1999, ARA&A, 37, 127
Peacock J. A., Dodds S. J., 1996, MNRAS, 280, 19P (PD)
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton University Press, Princeton, p. 147, 149–150
Scoccimarro R., 1998, MNRAS, 299, 1097
Scoccimarro R., Frieman J. A., 1999, ApJ, 520, 35
Scoccimarro R., Colombi S., Fry J. N., Frieman J. A., Hivon E., Melott A., 1998, ApJ, 496, 586
Szapudi I., 1998, ApJ, 497, 16
Szapudi I., Colombi S., 1996, ApJ, 470, 131 (SC)
Szapudi I., Szalay A. S., 1993a, ApJ, 408, 43 (SSa)
Szapudi I., Szalay A. S., 1993b, ApJ, 414, 493
Szapudi I., Szalay A. S., 1997, ApJ, 481, L1
Szapudi I., Szalay A. S., Boschán P., 1992, ApJ, 390, 350
Szapudi I., Meiksin A., Nichol R., 1996, ApJ, 473, 15
Szapudi I., Colombi S., Bernardeau F., 1999a, MNRAS, 310, 428 (SCB)
Szapudi I., Quinn T., Stadel J., Lake G., 1999b, ApJ, 517, 54 (SQSL)
Szapudi I., Colombi S., Jenkins A., Colberg J., 2000, MNRAS, 313, 725 (Paper II, this issue)
White S. D. M., 1979, MNRAS, 186, 145
White S. D. M., 1996, in Schaeffer R., Silk J., Spiro M., Zinn-Justin J., eds, Cosmology and Large-Scale Structure. Elsevier, Dordrecht
White S. D. M., Efstathiou G., Frenk C. S., 1993, MNRAS, 262, 1023
Zel'dovich Ya. B., 1970, A&A, 5, 84

APPENDIX A: DEFINITIONS AND NOTATIONS

The count probability distribution function (CPDF) P_N, gives the probability of finding N objects in a cell of volume v thrown at random in the catalogue.

Factorial moments, F_k, are defined as follows

$$F_k \equiv \langle (N)_k \rangle \equiv \langle N(N-1)\cdots(N-k+1)\rangle = \sum_N (N)_k P_N, \qquad \text{(A1)}$$

where the falling factorial $(N)_k$ is defined in the first part of the equation. The F_k are proportional to the moments of the underlying density field ρ smoothed over the cell of volume v: $F_k = \bar{N}^k \langle \rho^k \rangle$ (SSa; assuming the normalization $\langle \rho \rangle = 1$), where $\bar{N}$ is the average count in a cell:

$$\bar{N} \equiv \langle N \rangle = F_1. \qquad \text{(A2)}$$

Counts-in-cells are related to quantities of dynamical interest, such as the (connected) N-point correlation functions, ξ_N (e.g. Peebles 1980). The averaged N-point correlation function over a cell is given by

$$\bar{\xi}_N \equiv \frac{1}{v^N} \int_v \mathrm{d}^3 r_1 \cdots \mathrm{d}^3 r_N \xi(r_1, \ldots, r_N). \qquad \text{(A3)}$$

This is the connected moment of the smoothed density field, $\bar{\xi}_N = \langle \delta^N \rangle_c$ (with $\delta \equiv \rho - 1$). The connected moments or cumulants of a Gaussian field are identically zero for $N \geq 3$. In this paper, normalized cumulants are defined as

$$S_N \equiv \frac{\bar{\xi}_N}{\bar{\xi}^{N-1}}, \qquad \text{(A4)}$$

with the short-hand notation $\bar{\xi} \equiv \bar{\xi}_2$. By definition, $S_1 \equiv S_2 \equiv 1$, thus for second order $\bar{\xi}$ is used.

The quantities S_3 and S_4 are often called skewness and kurtosis in the astrophysical literature, although their definition differs slightly from the original usage in statistics. The reason for normalization in equation (A4) is dynamical. The S_Ns exhibit a weak scale dependence only owing to the scale-free nature of gravity. In the highly non-linear regime stable clustering is expected to set in (e.g. Peebles 1980) and in the weakly non-linear regime perturbation theory predicts approximate scaling depending on the initial fluctuation spectrum (e.g. Juszkiewicz et al. 1993; Bernardeau 1994).

The counts-in-cells generating function,

$$P(x) \equiv \sum_{N=0}^{\infty} x^N P_N, \qquad \text{(A5)}$$

writes (White 1979; Balian & Schaeffer 1989a; SSa)

$$P(x) = \exp\{ -\bar{N}(1-x)\sigma[N_c(1-x)]\}, \qquad \text{(A6)}$$

where

$$N_c \equiv \bar{N}\bar{\xi} \qquad \text{(A7)}$$

is the typical number of objects in an overdense cell (e.g. Balian & Schaeffer 1989a) and

$$\sigma(N_c) = \sum_{N=1}^{\infty} (-1)^{N-1} \frac{S_N}{N!} N_c^{N-1}. \qquad \text{(A8)}$$

It is worth noticing that (White 1979; Balian & Schaeffer 1989a; SSa)

$$P(x) = P_0[\bar{N}(1-x)], \qquad \text{(A9)}$$

if the void probability is expressed in terms of average counts $\bar{N}$. The measurement of P_0 is particularly interesting since it probes directly the count probability generating function:

$$\sigma(N_c) = -\ln(P_0)/\bar{N}. \qquad \text{(A10)}$$

The exponential generating function for factorial moments,

$$F(x) = \sum_{k \geq 0} F_k \frac{x^k}{k!} \qquad \text{(A11)}$$

is directly related to $P(x)$ (SSa) through

$$F(x) = P(x+1). \qquad \text{(A12)}$$

Combining equations (A6), (A8) and (A12), one can obtain a useful relation between cumulants and factorial moments (SSa)

$$S_N = \frac{\bar{\xi} F_N}{N_c^N} - \frac{1}{N} \sum_{k=1}^{N-1} \binom{N}{k} \frac{(N-k) S_{N-k} F_k}{N_c^k}. \qquad \text{(A13)}$$

The state of the art practical recipe consists of measuring the CDPF with high oversampling (Section 3), computing the factorial moments from equation (A1) and finally calculating the cumulants from the above recursion equation (A13). This procedure eliminates the need for an explicit discreteness correction.

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