Mon. Not. R. Astron. Soc. 000, 1–11 (1998) Printed 13 December 1999 (MN LaTeX style file v1.4)

Experimental Cosmic Statistics II : Distribution

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

Submitted to MNRAS

ABSTRACT

Colombi et al. 1999 (paper I) investigated the counts-in-cells statistics and their respective errors in the τCDM Virgo Hubble Volume simulation. This extremely large N-body experiment also allows a numerical investigation of the *cosmic distribution function*, $\Upsilon(\tilde{A})$ itself for the first time. For a statistic A, $\Upsilon(\tilde{A})$ is the probability density of measuring the value $\tilde{A}$ in a finite galaxy catalog. Υ was evaluated for the distribution of counts-in-cells, P_N, the factorial moments, F_k, and the cumulants, $\bar{\xi}$ and S_N's, using the same subsamples as paper I.

While paper I concentrated on the first two moments of Υ, i.e. the mean, the cosmic error and the cross-correlations, here the function Υ is studied in its full generality, including a preliminary analysis of joint distributions $\Upsilon(\tilde{A}, \tilde{B})$. The most significant, and reassuring result for the analyses of future galaxy data is that the cosmic distribution function is nearly Gaussian provided its variance is small. A good practical criterion for the relative cosmic error is that $\Delta A/A \lesssim 0.2$. This means that for accurate measurements, the theory of the cosmic errors, presented by Szapudi & Colombi (1996) and Szapudi, Colombi & Bernardeau (1999), and confirmed empirically by paper I, is sufficient for a full statistical description and thus for a maximum likelihood rating of models. As the cosmic error increases, the cosmic distribution function Υ becomes increasingly skewed and is well described by a generalization of the lognormal distribution. The cosmic skewness is introduced as an additional free parameter. The deviation from Gaussianity of $\Upsilon(\tilde{F}_k)$ and $\Upsilon(\tilde{S}_N)$ increases with order k, N, and similarly for $\Upsilon(\tilde{P}_N)$ when N is far from the maximum of P_N, or when the scale approaches the size of the catalog. For our particular experiment, $\Upsilon(\tilde{F}_k)$ and $\Upsilon(\tilde{\bar{\xi}})$ are well approximated with the standard lognormal distribution, as evidenced by both the distribution itself, and the comparison of the measured skewness with that of the lognormal distribution.

Key words: large scale structure of the universe – galaxies: clustering – methods: numerical – methods: statistical

1 INTRODUCTION

Precision higher order statistics will become a reality when the new wide field surveys, such as the SDSS and the 2dF, become available in the near future. These prospective measurements contain information relating to the regime of structure formation, to the nature of initial conditions, and to the physics of galaxy formation. The ability of such measurements to constrain models, in a broad sense, is inversely proportional to the overlap between the distribution of statistics predicted by different theories for a finite galaxy survey. More precisely, maximum likelihood methods give the probability of the particular measurements for each theory, or after inversion, the likelihood of the theories themselves. This is an especially natural and fruitful procedure for a Gaussian distribution, where the first two moments are sufficient for a full statistical description. This simple case is assumed for most analyses in the literature, and it motivates the special attention given to the investigation of the errors, or standard deviations. In general, however, the underlying distribution of measurements can be strongly non-Gaussian, in which case the correct shape for the distribution has to be employed for a maximum likelihood analysis. As a consequence, terms such as "1-σ measurement" loose their usual meaning: a few σ deviation from the average can be quite likely for a non-Gaussian distribution with a long tail. Therefore it is of utmost importance to ask two important questions:

(i) In what regime is the Gaussian approximation valid for the distribution of the measured statistical quantities?

(ii) If the Gaussian limit is violated, is there any reasonably simple, practical assumption which would enable a maximum likelihood analysis?

This paper attempts to answer these questions by studying numerically the underlying distribution function of measurements for estimators of higher order statistics based on counts-in-cells. This complements the thorough numerical investigation of the errors undertaken by Colombi et al. (1999, hereafter paper I), and the theoretical investigation of the errors exposed in a suite of papers by Szapudi & Colombi (1996, hereafter SC), Colombi, Szapudi, & Szalay (1998, hereafter CSS), and Szapudi, Colombi, & Bernardeau (1999, hereafter SCB).

For a particular statistic A, $\Upsilon(\tilde{A})$ denotes the probability density of measuring a value $\tilde{A}$ in a finite galaxy catalog. We consider the following counts-in-cells statistics: factorial moments F_k, cumulants $\bar{\xi}$ and S_N, void probability P_0 and its corresponding scaling function $\sigma \equiv -\ln(P_0)/F_1$, as well the counts-in-cells distribution itself, P_N. A large τCDM N-body experiment, $\mathcal{E}$, generated by the VIRGO consortium (e.g., Evrard et al. 1999) was divided into $C_{\mathcal{E}} = 4096$ cubic subsamples, $\mathcal{E}_i$, $i = 1, \ldots, C_{\mathcal{E}}$ for estimating numerically the cosmic distribution function, $\Upsilon(\tilde{A})$. This was rendered possible by the fact that this "Hubble Volume" simulation involves 10^9 particles in a cubic box of size $2000h^{-1}$ Mpc. A detailed description of the simulation and the method we used to extract count-in-cells statistics in the full box $\mathcal{E}$ and its each of subsamples $\mathcal{E}_i$ can be found in paper I.

Paper I concentrated entirely on the first two moments of $\Upsilon(\tilde{A})$, the average

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

and the cosmic error

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

In the equations above, the mean $\langle\tilde{A}\rangle$ can differ from the true value. The cosmic bias is defined as

$$b_A \equiv \frac{\langle\tilde{A}\rangle}{A} - 1. \tag{3}$$

It is always present when indicators are constructed from unbiased estimators in a nonlinear fashion, such as cumulants (e.g., SBC; Hui & Gaztañaga 1998, hereafter HG).

The most relevant results of paper I are summarized next:

(i) *The measured average* $\langle\tilde{A}\rangle$ is in excellent agreement with perturbation theory, one-loop perturbation theory and extended perturbation theory (EPT) in their respective range of applicability. These tests demonstrate the quality of our numerical experiment.

(ii) *The measured cosmic error* $\Delta A/A$ is in accord with the theoretical predictions of SC and SBC in their respective domain of validity. A few percent accuracy is achieved in the weakly non-linear regime for the factorial moments. On small scales the theory tends to overestimate the errors, perhaps by a factor of two in the worst case, due to the approximate nature of the hierarchical models representing the joint moments (SCB).

(iii) *The cosmic bias* is negligible compared to the errors in the full dynamic range, as predicted by theory (SCB, see also HG for an opposing view).

(iv) *Cross-correlations between statistics of order k and l* are in general agreement with theory considering the preliminary nature of the measurements. The precision of the predictions, however, decreases with increasing difference of orders, $|k - l|$. This suggests that the local Poisson model (SC) looses accuracy, as expected.

The theory of the errors confirmed by paper I provides an excellent basis for future maximum likelihood analyses of data whenever Υ is Gaussian. While this was tacitly assumed by most previous works, this article examines for the first time the range of validity of this assumption. To this end the cosmic distribution function $\Upsilon(\tilde{A})$ is examined numerically. In particular, one of the parameters determining its shape, the cosmic skewness

$$S \equiv \langle(\tilde{A} - \langle\tilde{A}\rangle)^3\rangle/(\Delta A)^3, \tag{4}$$

is calculated as well. When Gaussianity is no longer a good approximation, new Ansätze are proposed for characterizing $\Upsilon(\tilde{A})$. In addition we perform a preliminary analysis of the bivariate cosmic distributions $\Upsilon(\tilde{A}, \tilde{B})$.

The next section presents the estimates of Υ for the factorial moments, the cumulants (including the variance of the counts), the void probability distribution and its scaling function, and the counts-in-cells themselves. A universal shape is found for $\Upsilon(\tilde{A})$ which is well described in all regimes by a generalized version of the lognormal distribution. In addition to the mean (1) and variance (2), this depends on a third parameter, the cosmic skewness (4). This is also investigated along with the resulting *effective cosmic bias.* Section 3 presents the measured bivariate distributions, with explicit comparison to theoretical predictions of SCB. Finally, section 4 discusses the results in the context of maximum likelihood analysis of future surveys. Readers unfamiliar with counts-in-cells statistics can consult Appendix A in paper I for a concise summary of definitions and notation.

2 THE COSMIC DISTRIBUTION FUNCTION

The main results of this section are displayed in figures 1–6. For simplicity figures 1, 3, and 5 will be referred to as type D, displaying distributions, while figures 2, 4, and 6 as type S, showing skewness. A general description of each type is followed by results obtained for the cosmic distribution of the factorial moments (§ 2.1), cumulants (§ 3.2), counts-in-cells (§ 2.3), and void probability with its scaling function σ (§ 2.4). The cosmic skewness and the resulting effective bias are discussed in § 2.5.

In all figures of type D, the results are displayed in a convenient system of coordinates. For any statistic $\tilde{A}$ the normalized quantity

$$\tilde{x}_A \equiv \frac{\delta\tilde{A}}{\Delta A} = \frac{\tilde{A} - A}{\Delta A} \tag{5}$$

is considered where $A = \langle\tilde{A}\rangle$ to simplify notations. The average of $\tilde{x}_A$ is zero and its variance is unity by definition which facilitates the comparison of the plots. The disadvantage of this coordinate system is that the cosmic error $\Delta A/A$ is not directly shown.

For reference, each figure of type D displays a Gaussian (solid curve), and lognormal distribution with the same variance and average (dots, e.g. Coles & Jones 1991):

$$\Upsilon(\tilde{A}) = \frac{1}{\tilde{A}\sqrt{2\pi\kappa}} \exp\left\{-\frac{[\ln(\tilde{A}/A)+\kappa/2]^2}{2\kappa}\right\}, \quad (6)$$

with

$$\kappa = \ln[1+(\Delta A/A)^2]. \quad (7)$$

The skewness of this distribution is given by

$$S = (\Delta A/A)^3 + 3\Delta A/A. \quad (8)$$

For comparison, the skewness of the lognormal assumption is plotted with dotted lines on figures of type S. The amount of skewness of the lognormal is a function of the cosmic error, i.e. more skewness on the figures indicates a larger cosmic error which is hidden by the choice of the coordinate system.

In addition, a "generalized lognormal distribution" is introduced (dashes on figures of type D):

$$\Upsilon(\tilde{A}) = \frac{s}{\Delta A[s(\tilde{A}-A)/\Delta A+1]\sqrt{2\pi\eta}} \times \exp\left(-\frac{\{\ln[s(\tilde{A}-A)/\Delta A+1]+\eta/2\}^2}{2\eta}\right), \quad (9)$$

$$\eta = \ln(1+s^2), \quad (10)$$

where s is an adjustable parameter. It is fixed by the requirement that the analytical function (9) have identical average, variance, and skewness, $S = s^3 + 3s$, with the measured $\Upsilon(\tilde{A})$. It has more parameters, thus form (9) characterizes the shape of function $\Upsilon(\tilde{A})$ better than the other two functions, especially for the large $\delta\tilde{A}$ tail. As will be shown next, it is an excellent approximation for the underlying probability distribution *in all regimes for all statistics.* This robust universality is the most striking result of this article.

The cosmic distribution function, as with any measurement from finite data, is subject to both measurement and cosmic errors (the "error on the error problem", cf. SC). The measurement error on Υ, due to the finite number of subsamples extracted from the whole simulation, can be calculated via straightforward error propagation. It essentially corresponds to the usual $1/\sqrt{C_{\mathcal{E}}}$ factor, where $C_{\mathcal{E}}$ is the number of subsamples. This is plotted on all figures of type D as errorbars. On figures of type S no errorbars are shown, since this would require an accurate estimate up to the 6th moment of the cosmic distribution $\Upsilon(\tilde{A})$. The excellent agreement between cosmic error measurements and theory (paper I) indicates that the number of subsamples is sufficient and thus the resulting errorbars should be fairly small. Similar arguments suggest that the simulation volume was sufficient large to render the cosmic error on the cosmic distribution negligible.

2.1 Factorial Moments

Figure 1 displays $\Upsilon(\tilde{F}_k)$ for $1 \le k \le 4$ and various scales $\ell = 1,\ 7.8,\ 62.5h^{-1}$ Mpc.

The agreement with the generalized lognormal distribution is excellent, but even the lognormal gives an adequate description. The deviation from a Gaussian is pronounced whenever the relative cosmic error $\Delta F_k/F_k$ is significantly larger than unity. While the figures do not show the cosmic error directly, the skewness of $\Upsilon(\tilde{F}_k)$ is a reliable indication. It increases with the order k since $\Delta F_k/F_k$ also increases with k. Figure 2 shows directly the quantity S measured for $\Upsilon(F_k)$ along with the lognormal value (8). The agreement shows that the lognormal model yields an excellent approximation.

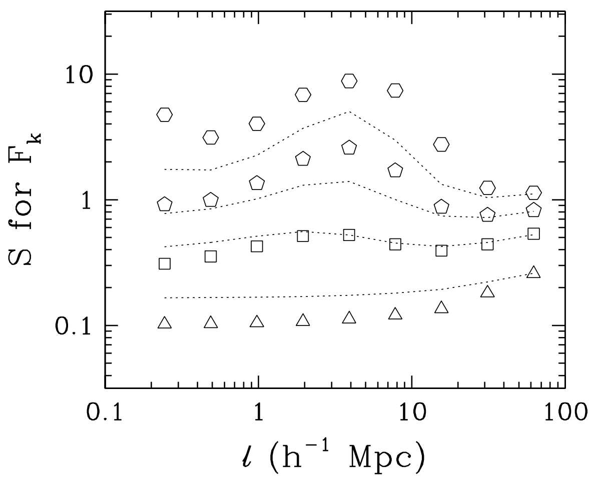

Figure 2. The skewness $S \equiv \langle(\tilde{F}_k - F_k)^3\rangle/(\Delta F_k)^3$ as a function of scale. The triangles, squares, pentagons and hexagons respectively correspond to $k = 1, 2, 3$ and 4. There are also dotted lines corresponding to an underlying lognormal distribution (8); the orders increase from bottom to top. The errors on the measurement have not been estimated since it would require a complicated calculation depending on the estimate of up to the 6$^{\text{th}}$ moment of $\Upsilon(\tilde{F}_k)$.

Fig. 1 in conjunction with the measurements of the cosmic error in Paper I suggests that

$$\Delta A/A \lesssim \Delta_{\rm crit}, \quad \Delta_{\rm crit} = 0.2, \quad (11)$$

is a practical criterion for the validity of the Gaussian approximation.

2.2 Cumulants

Figure 3 is analogous to Fig. 1, showing functions $\Upsilon(\tilde{\bar{\xi}})$, $\Upsilon(\tilde{S}_3)$ and $\Upsilon(\tilde{S}_4)$ for the biased estimators. As was shown in paper I, the bias is negligible compared to the cosmic errors, thus correction is not necessary. The agreement with the lognormal is more approximate than for $\Upsilon(\tilde{F}_k)$, except for the variance $\bar{\xi}$. Indeed, the skewness of $\Upsilon(\tilde{S}_N)$ is in general different from the lognormal prediction, as illustrated by Fig. 4. On small scales it is larger than predicted by equation (8) while on large scales where edge effects dominate it is much smaller. The generalized lognormal (9) can still account for the shape of $\Upsilon(\tilde{S}_N)$ quite well, especially for the large $\tilde{S}_N$ tail.

The cosmic skewness of $\Upsilon(\tilde{S}_k)$ is fairly small on large scales. This is a natural consequence of the fact that cumulants are not subject to the positivity constraint $\tilde{S}_k \ge 0$, as it is the case for factorial moments. On large scales, the measured $\tilde{S}_k$ may well be positive or negative, similarly with $\bar{\xi}$ on extremely large scales. As a result, the left-hand tail

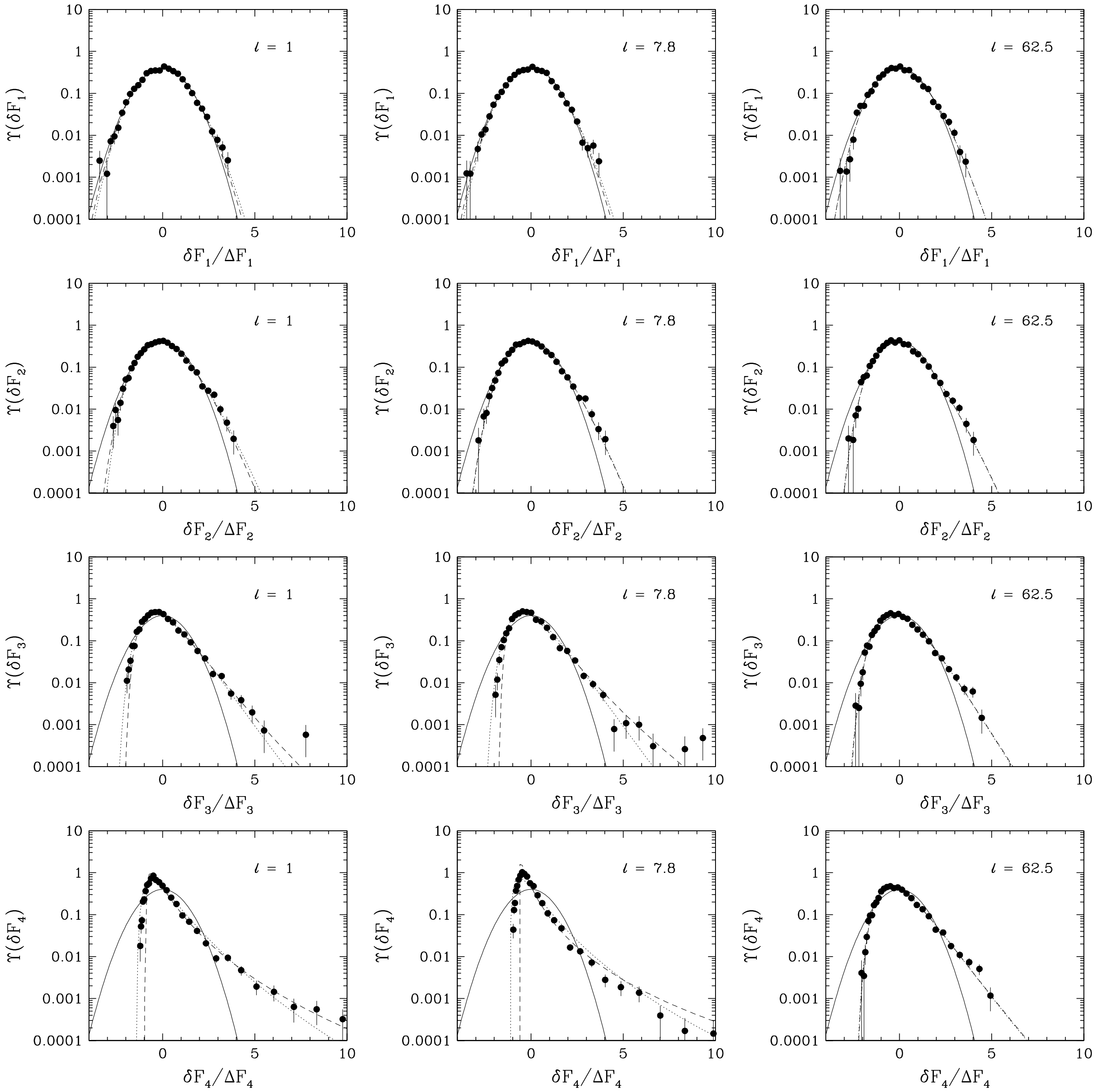

Figure 1. The cosmic distribution function of measurements $\Upsilon(\tilde{F}_k)$ shown as a function of $\delta\tilde{F}_k/\Delta F_k$ as explained in the text. The scale of the measurements $\ell = 1$, 7.8 and $62.5h^{-1}$ Mpc is indicated on each panel. The order $k = 1, 2, 3, 4$ increases from top to bottom. The solid, dotted, and dash curves correspond to the Gaussian, lognormal, and generalized lognormal [eq. (9)] distributions, respectively. While the coordinate system of the Figure does not display the value of the cosmic error directly, the amount of skewness of the lognormal distribution is an indicator of the magnitude $\Delta F_k/F_k$. The errorbars show the measurement error as discussed in beginning of § 2.

of the distribution is more pronounced in both lower right panels of Fig. 3 than the corresponding figure for factorial moments, and $\Upsilon(S_3)$ is almost Gaussian in the middle right panel.

Rule (11) for the Gaussian limit still applies, at least for $\bar{\xi}$, and perhaps a slightly more stringent condition should be chosen for cumulants of higher order. $\Upsilon(\tilde{S}_3)$ is fairly skewed even though the measured cosmic error is slightly below the threshold value for $\ell = 1h^{-1}$ Mpc and $\ell = 7h^{-1}$ Mpc (see paper I).

2.3 Counts-in-cells

Figure 5 shows the function $\Upsilon(\tilde{P}_N)$ in various cases. The upper panels focus on a small scale $\ell \simeq 1h^{-1}$ Mpc. In this regime, the CPDF and $-\Delta P_N/P_N$ are decreasing functions of N as demonstrated in paper I. Once again, the validity of the Gaussian approximation depends on the size of cosmic error. As a result, $\Upsilon(\tilde{P}_N)$ is nearly Gaussian for $N = 1$ and becomes more and more skewed as N increases. The lognormal approximation appears to be adequate within the errors, although it is slightly too skewed as illustrated by Fig. 6.

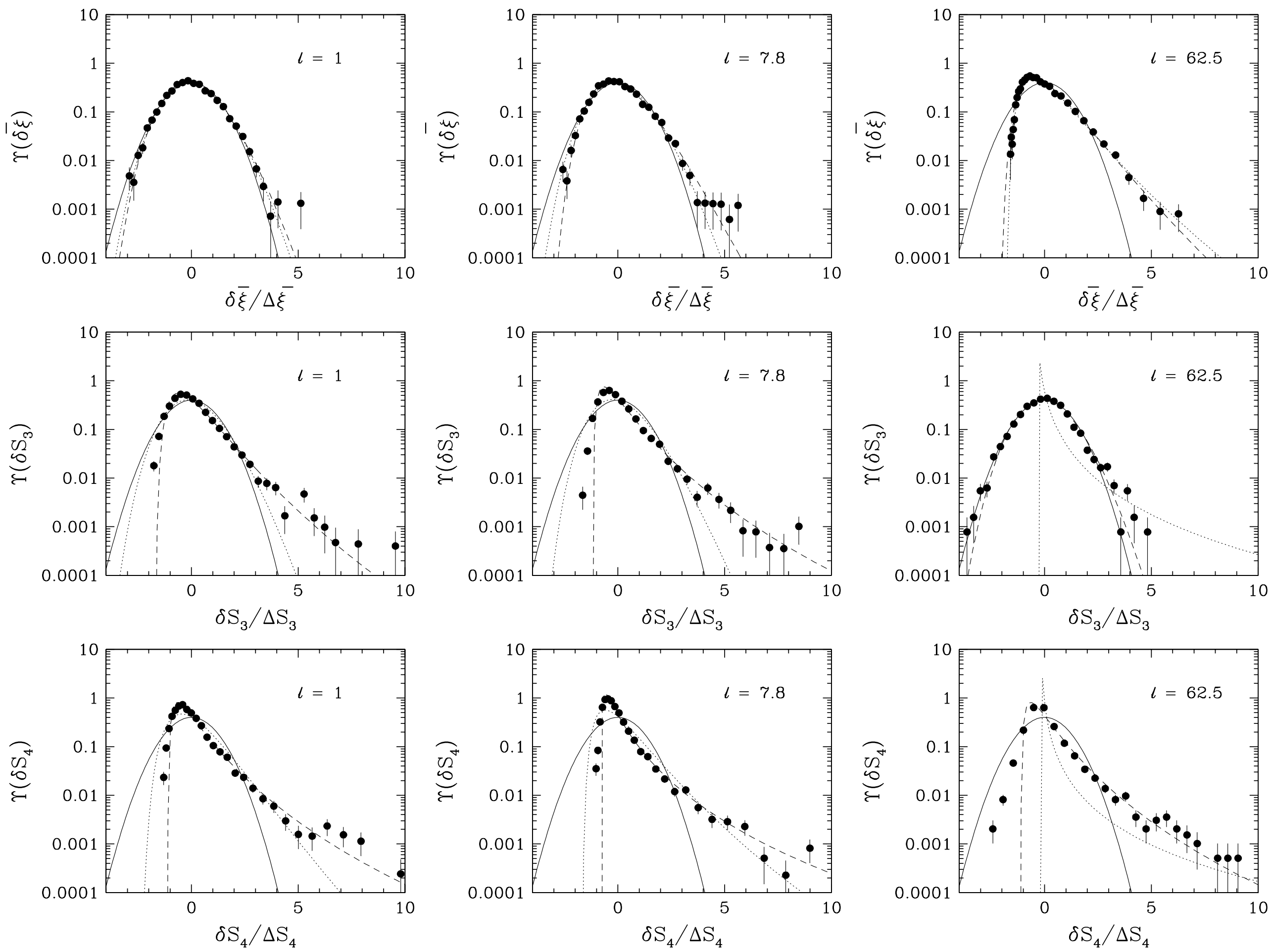

Figure 3. Analogous to Fig. 1 for $\Upsilon(\tilde{\bar{\xi}})$, $\Upsilon(\tilde{S}_3)$ and $\Upsilon(\tilde{S}_4)$.

The middle panels show an intermediate scale $\ell \simeq 7.8h^{-1}$ Mpc. On these scales (cf. paper I) both the CPDF and the cosmic error have a unimodal behaviour with an extremum (maximum for the CPDF and correspondingly minimum for the errors) for $N \sim N_{\rm max} = 26$. This explains why for the chosen values of $N = 5$, 50, and 500, function $\Upsilon(\tilde{P}_N)$ is skewed, approximately Gaussian, and skewed again respectively. For $N = 5$ lognormal is an excellent approximation, while the skewness for $N = 500$ is somewhat less than that of a lognormal.

Finally, the lower panels display the largest available scale $\ell = 62.5h^{-1}$ Mpc. The behaviour of P_N and $\Delta P_N/P_N$ is similar as previously with the extremum shifted to $N \sim N_{\rm max} \simeq 30000$. In this case, the cosmic error is always large, at least of order fifty percent (cf. paper I). All the curves are thus significantly skewed for the chosen values of $N = 25\,000$, $30\,000$ and $40\,000$. The agreement with the lognormal assumption is somewhat inaccurate, although the generalized lognormal improves the fit, especially for the left-hand panel. Note that the apparently abrupt limit for small values of $\delta\tilde{P}_N/\Delta P_N$ is due to the positivity constraint $\tilde{P}_N \geq 0$. This constraint becomes quite severe when the average value is much smaller than the errors. While there is still plenty of dynamic range for upscattering, there is a hard restriction for down scattering. This is only partly taken into account in our generalized lognormal model, and any modifications in this respect are left for future work. Finally, the practical criterion (11) is again valid for determining Gaussian approximation.

Note that the finite number $C = 512^3$ of sampling cells (see paper I), the CPDF is necessarily a multiple of $1/C$. This quantization could cause contamination of $\Upsilon(\tilde{P}_N)$ unless $P_N \gg 1/C \simeq 10^{-8.13}$. The condition $P_N \geq 10^{-6}$ adopted corresponds to at least ~ 100 cells per subsample in average with N particles. Despite that, a small amount of contamination might still persist for $\delta\tilde{P}_N \gtrsim -P_N$, i.e. at the left side of the plots on figure 5. The same effect might also alter the tail of the counts-in-cells measurements presented in paper I, although not significantly.

2.4 Void Probability and Scaling Function

According to the investigations in paper I, the cosmic error on P_0 and σ increases steadily with scale up to a sudden transition on scales $\ell \sim 5h^{-1}$ Mpc where it becomes large or infinite. This behavior was studied extensively by CBS where more of the details can be found. The most relevant consequence here is that in the available dynamic range the cosmic error is small, and $\Upsilon(\tilde{P}_0)$ and $\Upsilon(\tilde{\sigma})$ are nearly Gaus-

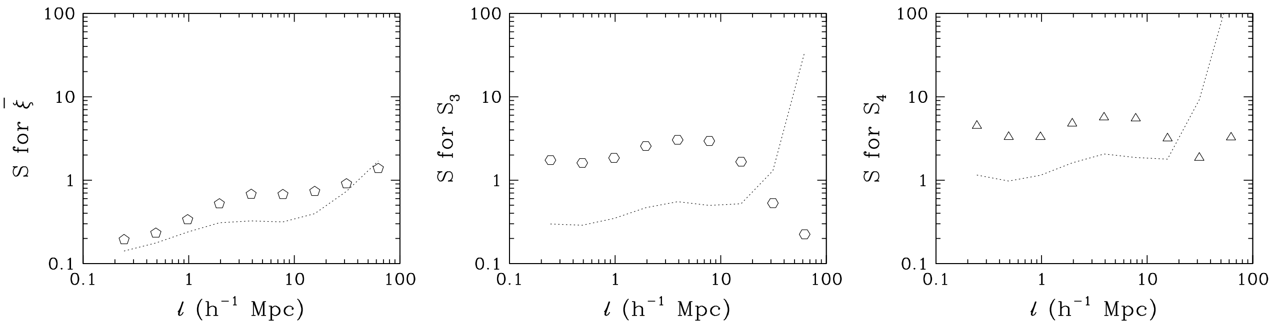

Figure 4. Same is in Fig. 2, but we consider here the skewness of $\tilde{\bar{\xi}}$ (left panel), $\tilde{S}_3$ (middle panel) and $\tilde{S}_4$ (right panel).

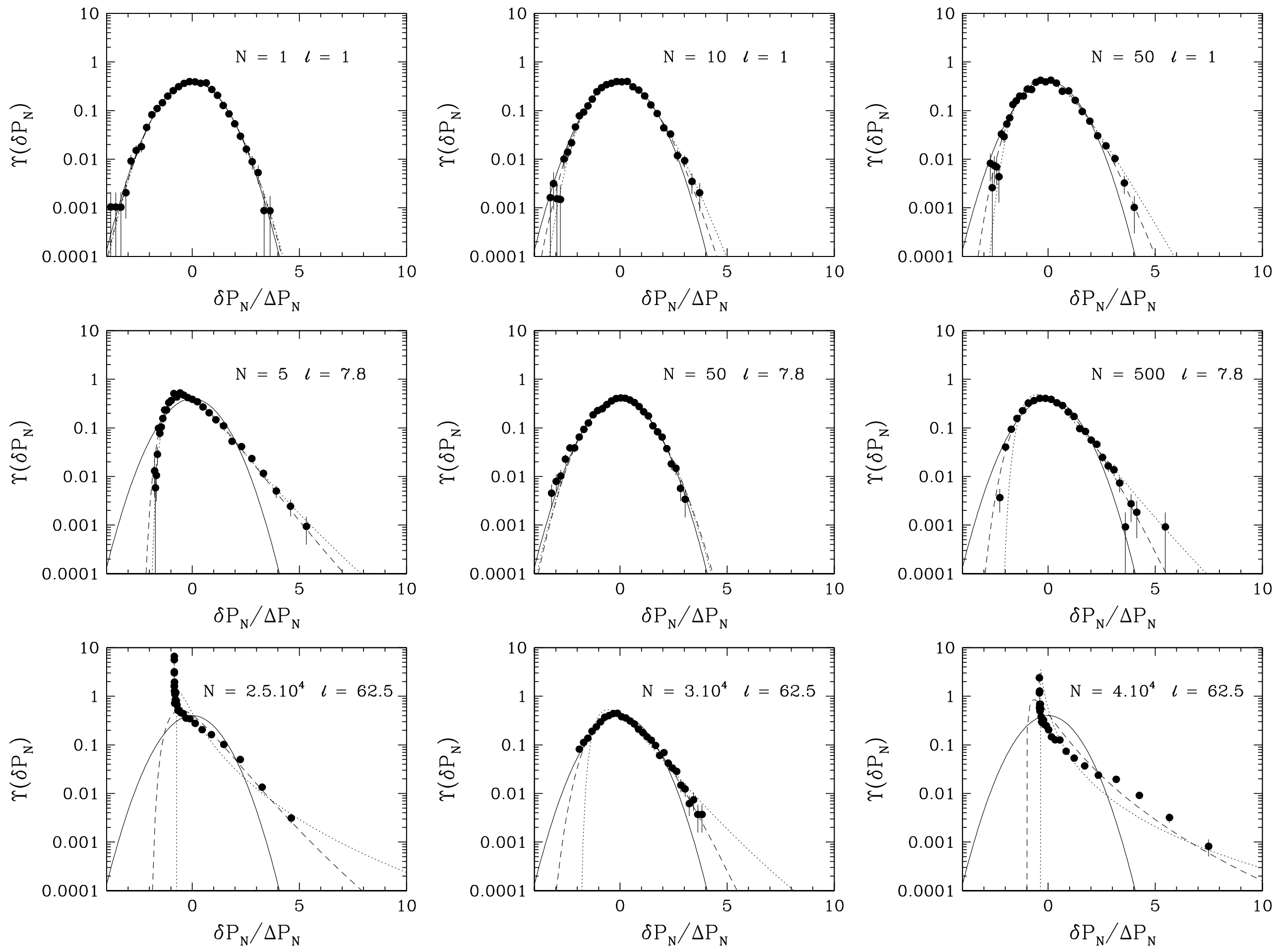

Figure 5. Same as in Fig. 1, but now, the distribution function of measurements $\Upsilon(\tilde{P}_N)$ is shown as a function of $\delta\tilde{P}_N/\Delta P_N$ for various scales and values of N as indicated on each panel.

sian. For this reason it would be superfluous to print the corresponding figures.

2.5 Cosmic Skewness and Cosmic Bias

According to Figs. 1–6, the degree of skewness of the cosmic distribution function increases with the order k and with $|N - N_{\max}|$, where $N_{\max}$ is the value for which P_N reaches its maximum. The cosmic skewness is already significant for third order statistics, F_3 and S_3. An important consequence of the large cosmic skewness is that the maximum $\Upsilon(\tilde{A})$, i.e. the most likely measurement, is shifted to the left from the ensemble average on Figs. 1, 3 and 5. Maximizing the Ansatz (9), which is always a good fit to the cosmic distribution function, yields

$$\flat_A = A_{\max}/A - 1 = \frac{\Delta A}{As}\left(\frac{1}{(1+s^2)^{3/2}} - 1\right), \tag{12}$$

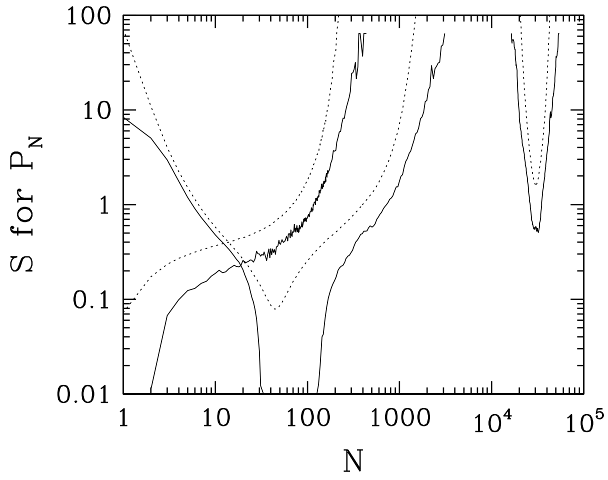

Figure 6. The skewness S of $\Upsilon(\tilde{P}_N)$ as a function of N for three scales, $\ell = 1h^{-1}$ Mpc (left curve), $\ell = 7.8h^{-1}$ Mpc (middle curve) and $\ell = 62.5h^{-1}$ Mpc (right curve). The dotted curves give the lognormal prediction, which is always larger than the measurement.

where $\flat_A$ is the *effective* cosmic bias. Since $s > 0$, it is negative, and its absolute value is smaller than the cosmic error,

$$|\flat_A| \lesssim 0.66 \frac{\Delta A}{A}. \qquad (13)$$

For a lognormal distribution, $s = \Delta A/A$,

$$|\flat_A| = 1 - \left[1 + (\Delta A/A)^2\right]^{-3/2} \leq 1. \qquad (14)$$

The effective cosmic bias becomes increasingly significant when the cosmic error is large. Similarly to the cosmic bias (SBC), $\flat_A \sim -(3/2)(\Delta A/A)^2$ from expanding eq. (14) in the small error regime.

The phenomenon of effective bias was already pointed out by SC (and preliminarily investigated by Colombi, Bouchet & Schaeffer, 1994). Since $A_{\max}$ is the most likely value of $\tilde{A}$, the only one available measurement in a catalog of the neighbouring Universe is likely to yield lower than average value. This is true even for an unbiased indicator such as $\tilde{F}_k$ or $\tilde{P}_N$. Unfortunately, this effect cannot be corrected for, but it can be taken into account in the framework of the maximum likelihood approach using the above results on the shape of $\Upsilon(\tilde{A})$.

3 BIVARIATE COSMIC DISTRIBUTION FUNCTION: A PRELIMINARY ANALYSIS

Figures 7 and 8 display contours of the joint cosmic distribution $\Upsilon(\tilde{A}, \tilde{B})$ (solid lines) for factorial moments and cumulants, respectively. For comparison the Gaussian limit is shown,

$$\Upsilon(\tilde{A}, \tilde{B}) = \frac{1}{2\pi \Delta A \Delta B \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\mathcal{Q}(\tilde{A}, \tilde{B})\right], \qquad (15)$$

$$\mathcal{Q}(\tilde{A}, \tilde{B}) = \frac{1}{1-\rho^2}\left[\tilde{x}_A^2 - 2\rho\tilde{x}_A\tilde{x}_B + \tilde{x}_B^2\right], \qquad (16)$$

where $\rho \equiv \langle \delta\tilde{x}_A \delta\tilde{x}_B \rangle$ is the cross-correlation coefficient. Dot-dashes display the above function with the measured ρ, $\Delta A/A$ and $\Delta B/B$, while long dashes represent the same function but with the parameters inferred from the theory of SCB with the E^{2}PT model (see paper I for details). The contours, correspond in the Gaussian limit to the 1σ (thin curves) level, $\mathcal{Q}(\tilde{A}, \tilde{B}) = 1$, and the 2σ (thick curves) level, $\mathcal{Q}(\tilde{A}, \tilde{B}) = 4$, are displayed in the coordinate system of the measured $\tilde{x}_A$ and $\tilde{x}_B$.

On $\ell = 7.1h^{-1}$ Mpc scales the theoretical predictions are expected to match the second order moments of Υ for factorial moments, and even the cross-correlations (see Paper I). This is illustrated by Fig. 7, where the long-dashed ellipses superpose well to the dot-dashed ones. For the cumulants the theory overestimates the errors slightly, which is reflected in the contours of Fig. 8, although cross-correlations are still reasonable, as indicated by the orientation of the ellipses.

The departure from the Gaussian limit is significant, except for the upper left panel on Figs. 7 and 8, and increases with order, in accord with the findings of the previous section. The contrast with Gaussianity increases with the cosmic error, and thus with the order considered. With the exception of $\overline{N}$, F_2, $\overline{\xi}$ and S_3, the measured cosmic error violates (11) at $\ell = 7.1h^{-1}$ Mpc (see paper I). Moreover, as shown previously, criterion (11) should be strengthened for cumulants S_k, $k \geq 3$. In conclusion, condition (11) distinguishes the Gaussian limit for $\Upsilon(\tilde{A}, \tilde{B})$ adequately when applied to both statistics $\tilde{A}$ and $\tilde{B}$.

Similarly to the monovariate distribution (§ 2), function $\Upsilon(\tilde{A}, \tilde{B})$ develops skewness and a significant tail for large values of $\tilde{x} = (\tilde{x}_A, \tilde{x}_B)$ when rule (11) is broken. There are three notable consequences:

(i) The effective cosmic bias (§ 2.5) is present again, i.e. the maximum of Υ is shifted from the average towards the lower left corner of the panels.

(ii) The contours tend to cover a smaller area than for the Gaussian limit.

(iii) As a result of the positivity constraint, there is a well defined lower vertical/horizontal bound in some panels, e.g., for $\tilde{x}_{F_4}$, $F_4 \geq 0$.

4 SUMMARY AND DISCUSSION

This paper has presented an experimental study of the cosmic distribution function of measurements $\Upsilon(\tilde{A})$, where $\tilde{A}$ is an indicator of a statistic related to counts-in-cells. The cosmic distribution was considered for the factorial moments F_k, cumulants $\overline{\xi}$ and S_N, the void probability P_0 with its scaling function, $\sigma \equiv -\ln(P_0)/F_1$, and finally the counts-in-cells P_N themselves. To analyse properties of the function $\Upsilon(\tilde{A})$, we used a state of the art τCDM simulation divided into 4096 sub-cubes large enough themselves to represent a full galaxy catalog. The statistics mentioned above were extracted from each subsample, and the resulting distribution of measurements was used to estimate $\Upsilon(\tilde{A})$.

While paper I concentrated on the first two moments of the cosmic distribution, the average and the errors, here the focus was shifted towards the general shape of function Υ itself, including its skewness, the cosmic skewness. The main results of this analysis are the followings:

(i) In contrast with popular belief, the cosmic distribu-

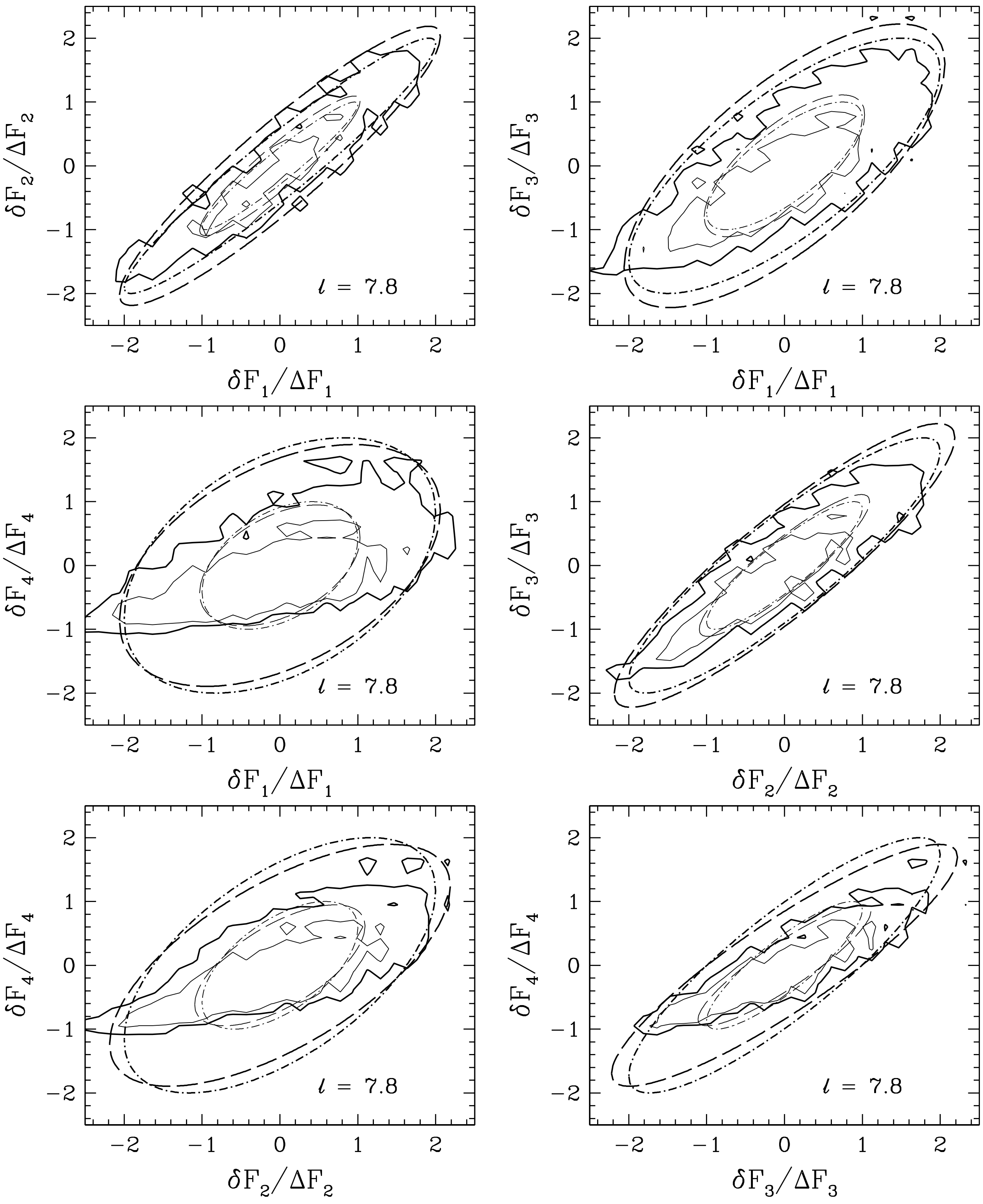

Figure 7. The joint cosmic distribution function for factorial moments, $\Upsilon(\tilde{F}_k, \tilde{F}_l)$. Thin and thick solid contours are displayed for two values of Υ which would correspond respectively to 1σ and 2σ contours in the Gaussian limit. The latter is shown as thin and thick dot-dashes. For comparison, the analytic prediction of SCB for E^2PT is also plotted with thin and thick long dashes corresponding to the Gaussian limit with theoretical cosmic errors and cross-correlation coefficient. The scale of the measurement is $\ell = 7.8\ h^{-1}$ Mpc as displayed on each panel. The image used to draw contour plots has 30^2 pixels. It was generated using bilinear interpolation from an other array with logarithmic binning in each coordinate in order to reduce the errors on the estimate of function $\Upsilon(\tilde{A}, \tilde{B})$ in each bin.

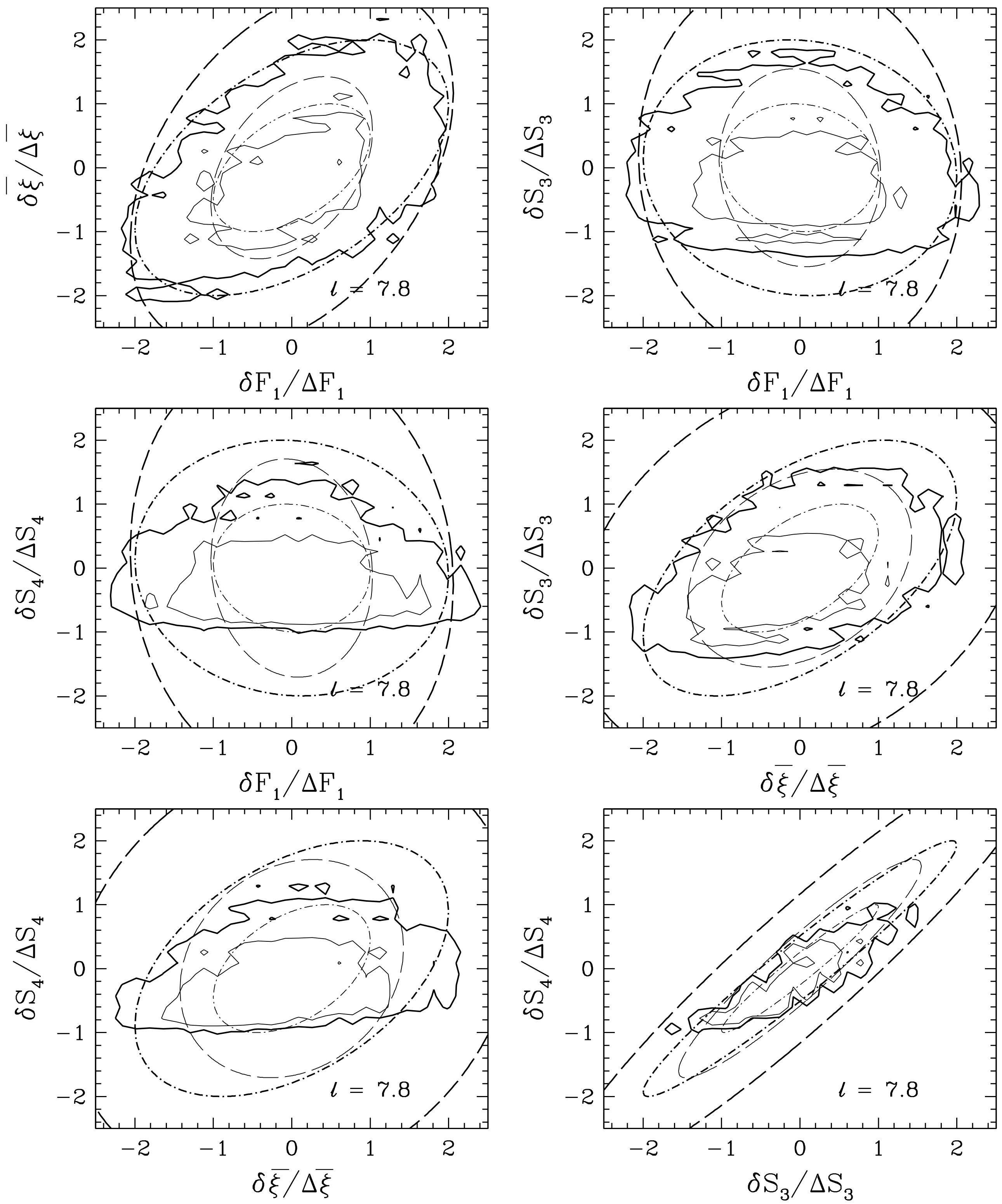

Figure 8. Same as in Fig. 7, but for the average count F_1 and the cumulants, $\overline{\xi}$, S_3 and S_4.

tion is *not Gaussian* in general. The most reassuring result is, however, that the Gaussian approximation appears to be valid whenever the cosmic errors are small, typically $\Delta A/A \lesssim 0.2$. This result is quite robust and it is insensitive to the particular statistic considered (except that a slightly more stringent condition might be chosen for cumulants S_k, $k \geq 3$). This means that for any quantity which can be reliably measured from a survey, a Gaussian error analysis should be valid.

When the relative cosmic error $\Delta A/A$ becomes significant,

Υ becomes increasingly skewed. Since $\Delta F_k/F_k$ and $\Delta S_k/S_k$ increase with k (SC, paper I), and $\Delta P_N/P_N$ with $|N - N_{\max}|$, where $N_{\max}$ is the maximum of the CPDF, so does the cosmic skewness, which eventually results in the break down of the Gaussian approximation. Functions $\Upsilon(\tilde{F}_k)$ and $\Upsilon(\tilde{\bar{\xi}})$ are well approximated by a lognormal law. Otherwise, a third order parametrisation matching the average, the variance and the skewness of the observed distribution is necessary, and in general sufficient. Such a generalization of lognormal distribution is proposed and found to be in agreement with the measurements in all regimes investigated. Note that there are other alternatives such as the Edgeworth expansion (e.g., Juszkiewicz et al. 1995) or the skewed lognormal approximation of Colombi (1994). This latter consists of applying Edgeworth expansion to $\log(\tilde{A})$. This method, when applicable, improves significantly the domain of validity of the Edgeworth expansion, normally only useful in the weakly non-Gaussian limit $\Delta A/A \lesssim 0.5$.

(ii) While paper I examined the cosmic bias resulting from the non-linear construction of certain estimators, here a new phenomenon was pointed out, which is similar in effect, but different in nature: the *effective cosmic bias*. It affects all estimators, including unbiased ones, and is a result of the cosmic skewness. Whenever the cosmic errors are large, the cosmic distribution function develops a skewness corresponding to a long tail. As a result, the most likely measurement will be smaller than the average. Such a phenomenon was pointed out earlier in SC, and here it has been found to be universal. As SCB and paper I found that the cosmic bias is usually insignificant compared to the cosmic errors, it is likely that the effective cosmic bias is responsible for some of the conspicuously low measurements from small galaxy catalogs. This is in contrast with the conjecture of Hui & Gaztañaga (1998, hereafter HG), who assumed that the cosmic bias resulting from the use of biased estimators could explain this phenomenon. The effective cosmic bias renders correction for the cosmic bias useless, in contrast with the proposition of HG. The effective cosmic bias (and the less significant cosmic bias if any) can be taken into account in the framework of a full maximum likelihood analysis, which relies on the shape of the cosmic distribution function approximated with sufficient accuracy.

(iii) A preliminary investigation of joint distribution $\Upsilon(\tilde{A}, \tilde{B})$ was performed for factorial moments and cumulants. It confirms the validity of the above points (i) and (ii) for cosmic bivariate distribution. In particular, a practical criterion for the validity of the Gaussian limit is that the cosmic error for both estimators be small enough, typically $\Delta A/A \lesssim 0.2$ and $\Delta B/B \lesssim 0.2$. This result can be safely generalized to N-variate distribution functions, thus providing the basis of full multivariate maximum likelihood analysis of data in the Gaussian limit.

We have not attempted to develop a more accurate multivariate approximation than (multivariate) Gaussian as this would go beyond the scope of this paper. However, we conjecture that an extension of our generalized lognormal distribution would be feasible (see the point of view of Sheth, 1995). An alternate approach, proposed by Amendola (1996), would employ a multivariate Edgeworth expansion. However, similarly with point (i) above for monovariate distributions, this approximation is only valid when the errors are small; but this is precisely the criterion for the Gaussian limit as we shown previously. A generalization of the lognormal distribution expanding the logarithm of the statistics via the multivariate Edgeworth technique provides a potential improvement of this method.

It is worth noting that the behaviour of the cosmic distribution function is expected to be extremely robust with respect to the particular model studied in this paper, τCDM. For example, SC, in their preliminary investigations, found essentially the same universal behaviour in Rayleigh-Levy fractals. Moreover, as discussed more extensively in Paper I, the results are sufficiently stable that the usual worries of galaxy biasing (not to be confused with cosmic and effective cosmic bias) and redshift distortions are unlikely to change them qualitatively. Indeed the shape of the cosmic distribution function is almost entirely determined by the magnitude of the cosmic error, and it is insensitive to which statistic is considered. The powerful universality found among entirely different statistics is likely to carry over when the two effects mentioned above, which are subtleties in comparison with the range of statistics investigated, are taken into account.

The results found in the present work and in paper I are encouraging for investigations in future large galaxy catalogs and for problems related to data compression (e.g. Bond 1995; Vogeley & Szalay 1996; Tegmark, Taylor & Heavens 1996; Bond, Jaffe & Knox 1998; Seljak 1998). For example, the cosmic error on factorial moments is expected to be small on a large dynamic range in the SDSS (see, e.g. CSS), implying according to the above findings that the cosmic distribution function should be nearly Gaussian in this regime. In that case, theory of the cosmic errors and cross-correlations, outlined in SC, CSS and SCB and thoroughly tested in paper I, will be sufficient for full multivariate maximum likelihood analyses. Preliminary investigations on current surveys are being undertaken by Szapudi, Colombi & Bernardeau (1999b) and Bouchet, Colombi & Szapudi (1999). Similarly the theoretical background is currently being developed for future weak lensing surveys (Berneardeau, Colombi, Szapudi, 1999), where statistical analyses will be conducted with indicators very close to counts-in-cells (see, e.g. Bernardeau, Van Waerbeke & Mellier 1997; Mellier 1998; Jain, Seljak & White 1999).

ACKNOWLEDGMENTS

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 dept 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://starwww.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.

REFERENCES

Amendola L., 1996, MNRAS, 283, 983
Bernardeau F., Colombi S., Szapudi I., 1999, in preparation
Bernardeau F., Van Waerbeke L., Mellier Y., 1997, A&A, 322
Bond J.R., 1995, Phys. Rev. Lett., 74, 4369
Bond J.R., Jaffe A.H., Knox L., 1998, preprint (astro-ph/9808264)
Bouchet F.R., Colombi S., Szapudi I., 1999, in preparation
Coles P., Jones B., 1991, MNRAS, 248, 1
Colombi S., 1994, ApJ, 435, 536
Colombi S., Bouchet F.R., Schaeffer R., 1994, A&A, 281, 301
Colombi S., Szapudi I., Jenkins A., Colberg J., 1999c, submitted to MNRAS (paper I)
Colombi S., Szapudi I., Szalay A.S., 1998b, MNRAS, 296, 253
Evrard A.E., et al., 1998, in preparation
Hui L., Gaztañaga E., 1998, preprint (astro-ph/9810194) (HG)
Jain B., Seljak U., White S.D.M., 1999, preprint (astro-ph/9901191)
Juszkiewicz R., Weinberg D.H., Amsterdamski P., Chodorowski M., Bouchet F.R., 1995, ApJ, 442, 39
Mellier Y., 1999, Ann. Rev. Astron. Astrophys., in press
Seljak U., 1998, ApJ, 503, 492
Sheth R.K., 1995, MNRAS, 277, 933
Szapudi I., Colombi S., Bernardeau F., 1999a, submitted to MNRAS (SBC)
Szapudi I., Colombi S., Bernardeau F., 1999b, in preparation
Szapudi I., Colombi S., 1996, ApJ, 470, 131 (SC)
Tegmark M., Taylor A.N., Heavens A.F., 1996, ApJ, 480, 22
Vogeley, M.S., Szalay A.S., 1996, ApJ, 465, 34

This paper has been produced using the Royal Astronomical Society/Blackwell Science LATEX style file.