Fermi National Accelerator Laboratory

Effects of Sampling on Statistics of Large Scale Structure

Stéphane Colombi, István Szapudi and Alexander S. Szalay

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

July 1997

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Distribution

Approved for public release; further dissemination unlimited.

Effects of Sampling on Statistics of Large Scale Structure

Stéphane Colombi[1,2], István Szapudi[3] and Alexander S. Szalay[4]

[1]CITA, 60 St George St., Toronto, ON, Canada, M5S 1A7

[2] Institut d'Astrophysique de Paris, CNRS, 98bis bd Arago, F-75014 Paris, France

[3]NASA/Fermilab Astrophysics Center, Fermi National Accelerator Laboratory, Batavia, IL 60510-0500

[4]Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218

Abstract

The effects of sampling are investigated on measurements of counts-in-cells in three-dimensional magnitude limited galaxy surveys, with emphasis on moments of the underlying smooth galaxy density field convolved with a spherical window. A new estimator is proposed for measuring the k-th order moment $\langle \rho^k \rangle$: the weighted factorial moment $\tilde{F}_k[\omega]$. Since these statistics are corrected for the effects of the varying selection function, they can extract the moments in one pass without the need of constructing a series of volume limited samples.

The cosmic error on the measurement of $\tilde{F}_k[\omega]$ is computed via the the formalism of Szapudi & Colombi (1996), which is generalized to include the effects of the selection function. The integral equation for finding the minimum variance weight is solved numerically, and an accurate and intuitive analytical approximation is derived $\omega_{\rm optimal}(r) \propto 1/\Delta(r)$, where $\Delta(r)$ is the cosmic error as a function of the distance from the observer. The resulting estimator is more accurate than the traditional method of counts-in-cells in volume limited samples, which discards useful information. As a practical example, it is demonstrated that, unless unforeseen systematics will prevent it, the proposed method will extract moments of the galaxy distribution in the future Sloan Digital Sky Survey (hereafter SDSS) with accuracy of order few percent for $k = 2, 3$ and better than 10% for $k = 4$ in the scale range of $1\ h^{-1}$ Mpc $\leq \ell \leq 50\ h^{-1}$ Mpc. In the particular case of the SDSS, a homogeneous (spatial) weight $\omega = 1$ is reasonably close to optimal.

Optimal sampling strategies for designing magnitude limited redshift surveys are investigated as well. The arguments of Kaiser (1986) are extended to higher order moments, and it is found that the optimal strategy depends greatly on the statistics and scales considered. A sampling rate $f \sim 1/3 - 1/10$ is appropriate to measure low-order moments with $k \leq 4$ in the scale range $1\ h^{-1} \lesssim \ell \lesssim 50\ h^{-1}$ Mpc. However, the optimal sampling rate increases with k, the order considered, and with $1/\ell$. Therefore count-in-cells statistics in general, such as the shape of the distribution function, high order moments, cluster selection, etc., require full sampling, especially at small, highly nonlinear scales $\ell \sim 1\ h^{-1}$ Mpc.

Another design issue is the optimal geometry of a catalog, when it covers only a small fraction of the sky. Similarly as Kaiser (1996), we find that a survey composed of several compact subsamples of angular size $\Omega_{\rm F}$ spread over the sky on a glass-like structure would do better, with regards to the cosmic error, than the compact or the traditional slice like configurations, at least at small scales. The required dynamic range of the measurements determines the characteristic size of the subsamples. It is however difficult to estimate, since an accurate and cumbersome calculation of edge effects would be required at scales comparable to the size of a subsample.

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

1. Introduction

The large scale structure of the Universe is generally admitted to be homogeneous at scales above ~ 150 Mpc. At smaller scales, observations of the galaxy distribution show a remarkable clustering as evidenced by voids, clusters, filaments, and superclusters. According to standard theories, these structures grew from small initial fluctuations under the influence of gravity. This, together with possible biasing, resulted in the random point process represented by galaxies. Thus, statistical methods can be applied efficiently to galaxy surveys to constrain models of large scale structure formation. Once the statistical tool is selected, two important questions influence its applicability. First, an optimal *sampling strategy* can be used to build a galaxy catalog (e.g. Kaiser 1986, hereafter K86), thus maximizing the information content with respect to the statistics used. Second, an optimal measurement method can be used to extract the maximum amount of information present in the catalog. The aim of this work is to address both of these questions in a quantitative way, focusing on low order moments of the probability distribution function (PDF) of the large scale galaxy density field. To achieve these goals, a number of plausible, nevertheless important assumptions were made, which are described next.

The galaxy distribution is assumed to be a discrete, locally Poissonian realization of an underlying smooth random field. To estimate the moments of the PDF of this random field, factorial moments (e.g. Szapudi & Szalay 1993a) of the count probability distribution function $P_N(\ell)$ (CPDF) are used. By definition, the CPDF represents the probability of finding N galaxies in a spherical (circular) cell of radius ℓ thrown at random in a three-dimensional (two-dimensional) galaxy catalog. The CPDF is easy to measure and widely used to test the *scaling behavior* of galaxy catalogs (see e.g. Alimi, Blanchard & Schaeffer 1990; Maurogordato, Schaeffer & da Costa 1992; Szapudi, Szalay & Bochàn 1992; Bouchet et al. 1993; Gaztañaga 1992, 1994; Szapudi, Meiksin & Nichol 1996, Szapudi & Szalay 1997a) and N-body simulations data sets (see e.g. Bouchet, Schaeffer & Davis 1991; Bouchet & Hernquist 1992; Baugh, Gaztañaga & Efstathiou 1995; Gaztañaga & Baugh 1995; Colombi, Bouchet & Hernquist 1996).

Throughout this work, we assume an ideal three-dimensional magnitude limited galaxy catalog $\mathcal{E}$, of depth $R_{\max}$, with magnitude limit $M_{\lim}$, covering a given volume V of the universe, and containing a number N_{obj} of spherical coordinates of galaxies, (z, θ, ϕ), z being the measured redshift.

A purely statistical approach is used: except for the effects of the selection function in a magnitude limited sample, all observational effects are ignored, such as extinction, confusion limit, and systematic errors due to the imperfection of the instruments.

Redshifts are considered as pure distances, i.e. effects of projection in redshift space are neglected. Such effects can significantly change the behavior of the CPDF, especially in the nonlinear regime (e.g. Kaiser 1987; Lahav et al. 1992; Matsubara & Suto 1995; Hivon et al. 1995).

Finally, it is assumed that the clustering of galaxies does not depend significantly on their luminosity, which is probably a crude approximation. Indeed, there are both theoretical (e.g. White et al. 1987; Mo & White 1996; Valls-Gabaud, Alimi & Blanchard 1989; Bernardeau & Schaeffer 1992, hereafter BeS) and observational arguments (e.g. Hamilton 1988; Davis et al. 1988; Dominguez-Tenreiro & Martinez 1989; Benoist et al. 1996) suggesting that the level of clustering of galaxies increases with their luminosity.

In realistic redshift surveys, the average number density of galaxies $n(r)$ changes with the distance r from the observer. The CPDF is traditionally defined in an *homogeneous* catalog, i.e. with constant $n(r)$. One way to bypass this problem is extraction of volume-limited subsamples $\mathcal{E}_{\mathrm{VL}}^i$ of

depth $R_i \leq R_{\mathrm{max}}$ from the main catalog $\mathcal{E}$ (see for example Maurogordato et al. 1992; Bouchet et al. 1993). In these subsamples, the objects are such that their apparent magnitude at distance $r = R_i$ would be larger than M_{lim}. Such a selection criterion renders the number density of galaxies in the catalog independent of distance, at the price of significant information loss. Although this can be partially recovered by cutting several volume-limited catalogs with various values of R_i, it would be preferable to extract all the available information from the catalog in a single measurement. This is possible by defining an inhomogeneous counts-in-cells measure which is corrected for the variation of $n(r)$ with the distance from the observer, similarly to, e.g., Efstathiou et al. 1990 and Szapudi & Szalay 1996. This, combined with a minimum variance weighting, results in unbiased, selection corrected estimators of the N-th factorial moment of the galaxy counts. To clarify the substantial gain from such an approach, we carefully calculate the errors on the measurements, and show how our method minimizes them.

Only statistical errors are considered, caused by the fact that only a finite part of the universe is accessible for observations. The corresponding *cosmic error* was calculated by Szapudi & Colombi (1996, hereafter SC) for count-in-cells measurements in an homogeneous catalog. The different contributions were classified as follows:

1. The finite volume error is due to fluctuations of the underlying random field at wavelengths larger than the size of the catalog. The finiteness of the sampled volume causes systematic effects on the measurement of the CPDF and its moments, even in N-body simulations (Colombi, Bouchet & Schaeffer 1994, 1995, hereafter CBSI and CBSII).

2. Edge effects are related to the geometry of the catalog: the galaxies near the edges of the catalog receive less statistical weight than those far from the boundaries. Edge effects can be theoretically corrected for, at least partially. A corrected estimator can be used for the two-point correlation function $\xi(r)$ (see, e.g. Ripley 1988; Landy & Szalay 1992) from which the variance of the PDF, $\overline{\xi}(\ell)$, can be obtained as a double integral of ξ over a cell of radius ℓ. A class of edge corrected estimators was actually recently introduced by Szapudi & Szalay (1997b) for the higher order moments as well. They, however, use a somewhat complicated procedure which is applicable only to the moments of fluctuations.

3. The shot noise error (or, equivalently, the error from discreteness effects) is due to the incomplete sampling of the underlying smooth field with a finite number of points. In particular, excessive undersampling causes degeneracy of the CPDF, rendering the confrontation of various models against observations difficult (e.g. Bouchet et al. 1993, CBSII). By definition, discreteness effects tend to zero as the average number density of objects in the catalog approaches infinity[1].

While unbiased estimators can be constructed in a number of ways, the above sources of errors can be reduced by giving an appropriate statistical weight depending on the region in the catalog. In a magnitude limited catalog for instance, density decreases far away from the observer, resulting in an increase in the shot noise. This alone would call for increasing statistical weight close to the observer, thus decreasing the contribution of distant portion of the survey. However, this reduces the effective sampled volume, and increases the finite volume and the edge contributions

---

[1] Note however that part of the shot noise belongs to the subclass of "edge-discreteness" effects, i.e. it is also an edge effect, and can be corrected for with appropriate techniques (Szapudi & Szalay 1997b).

to the error. There exists a "minimum variance" weighting which provides a compromise between various effects. By deriving and solving the integral equation for the minimum variance weighting scheme, we show that a significant gain in accuracy on the measurement can be achieved with our method, compared to the traditional volume-limited approach. Note that the word "optimal" is used synonymously with minimum variance in this paper, i.e. producing the smallest error in a class of unbiased estimators. Although this usage does not emphasize it, it is possible in principle to find even better estimators by extending the class in which the search is performed.

While the above procedure can optimize the way information is retrieved from an existing survey, calculation of cosmic errors helps in designing optimal surveys. In particular, an optimal sampling strategy can be found to build a three dimensional galaxy catalog for a given statistical indicator. Given the available telescope time, K86 found the optimal strategy for the measurement of the two-point correlation function of galaxies, using a simple model for the cosmic error. To reduce finite volume and edge effects, which are independent of the number of objects in the catalog, a sparse survey with large volume is needed. To reduce discreteness effects, on the other hand, a catalog as dense as possible is preferable. These competing effects determine the optimal sampling strategy with respect to a particular statistic. For the two-point function, K86 concluded in favor of sparse samples with approximately 1/10 to 1/20 of the candidates randomly selected for measuring their redshifts. Sparse sampling strategies were in fact used for several surveys, such as the Stromlo APM redshift survey (e.g. Loveday et al. 1992), the QDOT redshift survey (e.g. Moore et al. 1994), and the Durham/UKST redshift survey (e.g. Ratcliffe et al. 1997). However, the advent of multi-fiber spectroscopy makes large, complete galaxy surveys possible in a significantly shorter time than when K86 proposed his idea. Prime examples are the Sloan Digital Sky Survey (SDSS, e.g. Loveday 1996), and the 2DF Survey (e.g. Lahav 1996). While the problem of optimal sampling strategies might loose from its relevancy with these new developments, it is still worth to study it: either to facilitate preliminary investigations to prepare large surveys, or optimize the design of surveys aimed at statistical properties of rarer or harder to find objects at various wavelengths. Indeed, it will be shown how sparse sampling strategies are relevant for measuring low order moments in a reasonable scale range, while high order statistics, direct analysis of the CPDF shape, cluster selection, etc., are highly sensitive to discreteness effects, therefore proving that full sampling is more advantageous.

Another important design issue, when the catalog has poor sky coverage, is its geometry (e.g. Kaiser 1996, hereafter K96). This affects finite volume, edge, and shot noise effects in different ways. Although a general solution to this complicated problem will not be given, it will be discussed in broad terms, with suggestions for reasonable design principles. The solution for the conceptually simplest case will be given, which is relevant even for surveys which will eventually cover a continuous portion of the sky. The results presented here can facilitate the extraction of preliminary results before the catalog is finished.

This article is organized as follows. § 2 introduces the statistical indicators used along this paper, i.e. the weighted factorial moments corrected for selection effects. § 3 contains the calculation of the general expression for the cosmic error, extending the results of SC to a magnitude limited catalog with a spatial weight. The integral equation for the optimal weight is given. Some simple but still quite general assumptions on the underlying statistics are made to enable the numerical calculation of the optimal weight and the corresponding cosmic error. § 4 elaborates a practical example: an SDSS-like catalog. In particular, a useful approximation for the optimal weight is found. We emphasize the advantages of extracting the information from the data with the optimal weight, compared to the traditional volume limited approach. In § 5, sparse sampling strategies are

4

studied along with the suitable choice of geometry for the catalog. § 6 summarizes and discusses the results. In addition, three Appendices provide further information. § A contains mathematical details concerning the calculation of the cosmic error. § B explains and tests the numerical method used to estimate the optimal weight. § C contains mathematical formulae used to derive analytical expressions for the optimal sampling rate in some asymptotic regimes.

2. Weighted factorial moments corrected for selection effects

For a statistically homogeneous catalog, the CPDF P_N is the probability of finding N galaxies in a cell of size ℓ. The factorial moments of counts in cells are defined for $k \geq 0$ by

$$F_k(\ell) \equiv \langle (N)_k \rangle \equiv \sum_N (N)_k P_N, \tag{1}$$

where the falling factorials are defined as $(N)_k = N(N-1)\ldots(N-k+1)$, and $(N)_0 \equiv 1$, therefore $F_0 \equiv 1$. Under the assumption of infinitesimal Poisson sampling (Peebles 1980) the factorial moments correct directly for the discrete nature of galaxy catalogs. More precisely, their ensemble averages are equal to the moments of the (appropriately normalized) underlying *smooth* density field ρ ($\langle \rho \rangle = 1$),

$$F_k = \overline{N}^k \langle \rho^k \rangle, \tag{2}$$

where $\overline{N} \equiv \langle N \rangle = F_1$ is the average number of objects per cell (see Szapudi & Szalay 1993a).

Counts-in-cells statistics are estimated by a large number of sampling cells positioned at random in the catalog. For a homogeneous survey, the corresponding estimator of the factorial moment is expressed by the counts N_i in the ith cell, $1 \leq i \leq C$,

$$\tilde{F}_k^C \equiv \frac{1}{C} \sum_i^C (N_i)_k . \tag{3}$$

Equivalently, the CPDF can be estimated first (e.g. SC), which gives the above estimator through equation (1). The form (3) is meaningful if all the cells are equivalent, i.e. if they all have the same statistical weight $\omega = 1$. In an attempt to correct for finite volume and edge effects, a weight $\omega_{\ell,k}(r_i)$ (to be determined later) can be assigned to each cell i. This can depend on the position and size of the cell, and the statistic at hand:

$$\tilde{F}_k^C \equiv \frac{1}{C} \sum_i^C (N_i)_k \, \omega_{\ell,k}(r_i). \tag{4}$$

The weight $\omega_{\ell,k}$ is determined by minimizing the value of the cosmic error under the constraint of appropriate normalization.

A magnitude limited catalog $\mathcal{E}$, such as defined in introduction, is inhomogeneous since the selection function is not uniform. If n is the real number density of the underlying galaxy distribution (assuming that such number is well defined), the effective number density $n(r)$ of galaxies in a thin shell at distance r from the observer reads

$$n(r) = n\phi(r), \tag{5}$$

5

where ϕ is the selection function. In particular, the average number of galaxies in a cell of radius ℓ at distance r from the observer is given by

$$\overline{N}_r = n_\ell(r)v, \quad v \equiv \frac{4}{3}\pi\ell^3, \tag{6}$$

with

$$n_\ell(r) = n\phi_\ell(r), \tag{7}$$

and $\phi_\ell(r)$ is the average of the selection function over a cell. The approximation $\phi_\ell \simeq \phi$ is excellent for the relevant scales, valid within a few percents at worst. It will be used for practical calculations throughout.

To correct for selection effects we follow Szapudi & Szalay (1996) by changing equation (4) in

$$\tilde{F}_k^C \equiv \frac{1}{C}\sum_{i=1}^{C}\frac{(N_i)_k\,\omega_{\ell,k}(r_i)}{[\phi_\ell(r_i)]^k}. \tag{8}$$

This is the final form we propose for the estimator of the factorial moments. Note that *a priori* knowledge of the selection function $\phi(r)$ is assumed. Although implicitly always present, the k and ℓ dependence of the weights $\omega_{\ell,k}$ will be usually omitted. The normalization of the weights follows from the requirement that the estimator is unbiased. Taking the ensemble average of the above equation, and averaging all possible random realizations of C cells, gives

$$\frac{1}{\hat{V}(\ell)}\int_{\hat{V}(\ell)}d^3r\omega(r) = 1, \tag{9}$$

where $\hat{V}(\ell)$ is the *effective* sampled volume, i.e. the volume occupied by the center of all possible cells of radius ℓ *contained* in the catalog. While the proposed estimator can be used directly to estimate the factorial moments, it might be advantageous to introduce the inhomogeneous CPDF, $P_N(r)$. Once this is estimated, the inhomogeneous factorial moments $F_k(r)$ can be calculated. Using the scaling $\phi(r)^{-k}$ and summing over with the appropriate weights $\omega(r)$ is identical to the proposed estimator. This way, however, any high order moment can be calculated using the inhomogeneous CPDF without the need of rescanning the whole catalog.

3. Cosmic error and optimal weight

This section generalizes the formalism for computing the cosmic error presented by SC to the case of inhomogeneous selection function and weight. The equation for the optimal weight is derived in § 3.1. The locally Poissonian and hierarchical assumptions are used to simplify the calculations in § 3.2. A few comments follow on the interpretation of the results thus far (§ 3.3), and, finally, it is shown how this formalism can be applied to practical measurements (§ 3.4).

3.1. Formalism

The variance of $\tilde{F}_k^C$ is defined by

$$\left(\Delta\tilde{F}_k\right)^2 \equiv \left\langle\left\langle\tilde{F}_k^2\right\rangle\right\rangle_C - \left\langle\left\langle\tilde{F}_k\right\rangle\right\rangle_C^2, \tag{10}$$

where $\langle\ \rangle$, and $\langle\ \rangle_C$ are the ensemble average, and averaging over all possible sampling cells, respectively (we dropped the C dependence of $\tilde{F}_k$). For the proposed estimator

$$\left\langle\left\langle\tilde{F}_k^2\right\rangle\right\rangle_C = \left\langle\frac{1}{C^2}\sum_{i=1}^{C}\omega^2(r_i)\frac{\langle(N_i)_k^2\rangle}{[\phi(r_i)]^{2k}}\right\rangle_C + \left\langle\frac{1}{C^2}\sum_{i\neq j}^{C}\omega(r_i)\omega(r_j)\frac{\langle(N_i)_k(N_j)_k\rangle}{[\phi(r_i)\phi(r_j)]^k}\right\rangle_C. \tag{11}$$

Following SC, the evaluation of this expression in terms of the parameters of the distribution is facilitated by generating functions. Let us introduce the generating function, $P_r(x)$, of the probability $P_N(r)$ of finding N objects in a cell of size ℓ at position r, where the effective average number density is $n_\ell(r)$:

$$P_r(x) \equiv \sum_{N=0}^{\infty} x^N P_N(r). \tag{12}$$

Similarly, we define $P_{r_i,r_j}(x,y)$ as the generating function of the bivariate probability $P_{N,M}(r_i,r_j)$ of finding N and M galaxies in cells of radius ℓ respectively at positions r_i and r_j:

$$P_{r_i,r_j}(x,y) \equiv \sum_{N,M=0}^{\infty} x^N y^M P_{N,M}(r_i,r_j). \tag{13}$$

As SC, we write formally

$$\left(\Delta\tilde{F}_k\right)^2 = \left[\frac{\partial}{\partial x}\right]^k \left[\frac{\partial}{\partial y}\right]^k E^{C,V}(x+1,y+1)\Bigg|_{x=y=0}. \tag{14}$$

The generating function of the total error, $E^{C,V}(x,y)$, is asymptotically the sum of two generating functions

$$E^{C,V}(x,y) = \left(1-\frac{1}{C}\right)E^{\infty,V}(x,y) + E^{C,\infty}(x,y). \tag{15}$$

Function $E^{\infty,V}(x,y)$ generates the errors for hypothetical surveys with finite volume V and infinite number of sampling cells:

$$E^{\infty,V}(x,y) = \frac{1}{\hat{V}^2}\int_{\hat{V}} d^3r_1 d^3r_2 \omega(r_1)\omega(r_2)[\phi(r_1)\phi(r_2)]^{-k}\left\{P_{r_1,r_2}(x,y) - P_{r_1}(x)P_{r_2}(y)\right\}. \tag{16}$$

Function $E^{C,\infty}(x,y)$ generates the errors due the finite number of sampling cells used to do the measurement[2]:

$$E^{C,\infty}(x,y) = \frac{1}{C}\left\{\frac{1}{\hat{V}}\int_{\hat{V}} d^3r\,\omega^2(r)[\phi(r)]^{-2k}P_r(xy)\right.$$
$$\left. - \frac{1}{\hat{V}^2}\int_{\hat{V}} d^3r\,\omega(r)[\phi(r)]^{-k}P_r(x)\int_{\hat{V}} d^3r\,\omega(r)[\phi(r)]^{-k}P_r(y)\right\}. \tag{17}$$

Note that the generating function defined above can *only* be used for the error on the k-th order moment, i.e. for each k a slightly different generating function must be used. The reason for this is

---

[2]If the weight is homogeneous and if there are no selection effects ($\omega = \phi = 1$), as in SC, this function does not depend on the survey volume. In that case, it generates the errors for hypothetical surveys with infinite volume ($V = \infty$) and finite number C of sampling cells. This explains the formal notation "$E^{C,\infty}(x,y)$".

the implicit k dependence of the weight ω (the scaling with the selection function could be taken into account simply with the substitution $x \to \phi x - \phi + 1$). Although it would be possible to define a single but three variate generating function for all orders, it is simpler and more practical to use the above definition. Also, for the sake of completeness, we quoted both the cosmic error $E^{\infty,V}(x,y)$, and the measurement error $E^{C,\infty}(x,y)$. The former is an inherent property of the survey, while the latter is related to the finite number of sampling cells C used to do the measurement. Since this contribution can be rendered arbitrarily small with massive "oversampling", $C \to \infty$, or with an algorithm corresponding to $C = \infty$ (Szapudi 1997) it will not be considered further (see SC for a discussion on the "number of statistically independent cells", i.e. the number of cells needed to extract all the relevant information from the catalog). In what follows, the upper indices will be dropped from the cosmic error for simplicity, $E = E^{C,V} \simeq E^{\infty,V}$.

The above formalism provides the framework to compute the optimal sampling weight, which minimizes the cosmic error on the measurement of $\tilde{F}_k$. Taking into account the normalization of the weights leads to the following Lagrangian

$$\mathcal{L}[\omega,\lambda] \equiv \left[\frac{\partial}{\partial x}\right]^k \left[\frac{\partial}{\partial y}\right]^k E(x+1,y+1)\Bigg|_{x=y=0} + 2\lambda \left\{\frac{1}{\hat{V}}\int_{\hat{V}} d^3r\omega(r) - 1\right\}. \qquad (18)$$

The optimal weight ω is thus the solution of the following integral equation

$$\frac{1}{\hat{V}}\int_{\hat{V}} d^3r\omega(r)[\phi(r)\phi(u)]^{-k} \left[\frac{\partial}{\partial x}\right]^k \left[\frac{\partial}{\partial y}\right]^k \{P_{r,u}(x+1,y+1) - P_r(x+1)P_u(y+1)\}\Bigg|_{x=y=0} + \lambda = 0. \qquad (19)$$

The constant λ is determined by the normalization (9).

3.2. Approximations

To compute the optimal weight for measuring F_k and the corresponding cosmic error, generating functions $P_{r,u}(x,y)$ and $P_r(x)$ are needed. More precisely, after partial differentiation of order k at $x = y = 0$ in equations (14), the factors $F_{l,m}(\ell,r,u)$ are needed, which are defined as

$$F_{l,m}(\ell,r,u) = \langle (N)_l (M)_m \rangle \qquad (20)$$

up to order $l + m = 2k$. These quantities generalize the concept of factorial moments to bivariate distributions (Szapudi et al. 1995). In equation (20), the ensemble average is taken over two cells of size ℓ at positions r and u in the catalog, which can possibly overlap. This especially complicates the problem of finding the optimal weight for measuring F_k and the corresponding cosmic error.

Following SC, one can make reasonable assumptions about the underlying statistics to simplify the calculations considerably: the hierarchical tree model, and the locally Poissonian approximation. Then, as shown below, only the factorial moments F_l, $l \leq 2k$ and the two point-function $\xi(r)$ are required *a priori* for the computation of the optimal weight and the corresponding cosmic error on F_k. The calculations are detailed in Appendix A. We summarize here the important steps and hypotheses:

1. The integral (16) is split into two parts, according to whether or not the two cells overlap.

8

2. The calculation of the non-overlapping part requires the knowledge of the bivariate count-in-cells generating function for disjoint cells. Following SC, we simplify it by taking two particular but still quite general cases of the hierarchical model (e.g. Peebles 1980, Balian & Schaeffer 1989, hereafter BS) by SS and BeS (see Appendix A for more details). The hierarchical model is seen to be a good approximation for the higher order statistics in the observed galaxy distribution (e.g. Groth & Peebles 1977; Fry & Peebles 1978; Sharp, Bonometto & Lucchin 1984; Szapudi et al. 1992; Meiksin, Szapudi & Szalay 1992; Szapudi et al. 1995; Szapudi & Szalay 1997a) and in N-body simulations (e.g. Efstathiou et al. 1988; Bouchet et al. 1991; Bouchet & Hernquist 1992; Fry, Melott & Shandarin 1993; Bromley 1994; Lucchin et al. 1994; CBSI, CBSII; Colombi et al. 1996).

The function $Pr_1, r_2(x, y)$ is then Taylor-expanded to first order in $\xi(r_{12})/\overline{\xi}(\ell)$, where $\xi(r)$ is the two-point correlation function, $r_{12} \equiv |r_1 - r_2|$ and

$$\overline{\xi}(\ell) \equiv \frac{1}{v^2} \int_v d^3 r_1 d^3 r_2 \xi(r_{12}). \tag{21}$$

This approximation is becoming more accurate when the cells are far away from each other, however, it is still reasonable even when the cells touch each other (e.g. Bernardeau 1996).

3. To compute the overlapping contribution, the variations of the weight and of the selection function are neglected within the cells. We also assume local Poisson behavior, which considerably simplifies the writing of the bivariate generating function of counts for overlapping cells (see SC).

SC carried out *explicitly* the calculation to leading order in v/V for uniform survey and constant weight. They found that the cosmic error could be separated into three contributions

$$\Delta^2[\omega = 1, \phi = 1] \equiv \left(\frac{\Delta \tilde{F}_k}{F_k}\right)^2 = \Delta_{\rm F}^2 + \Delta_{\rm E}^2 + \Delta_{\rm D}^2. \tag{22}$$

The term $\Delta_{\rm F}^2$ is the *finite volume error* discussed in introduction, arising from the contribution of disjoint cells in equation (16). It is proportional to the integral of the two-point function over the sampled volume

$$\overline{\xi}(L) \equiv \frac{1}{V^2} \int_V d^3 r_1 d^3 r_2 \xi(r_{12}) \tag{23}$$

$(L \sim V^{1/3})$. While it depends on the clustering properties of the system, i.e.,

$$\Delta_{\rm F}^2 = \mathcal{F}_{\rm F} \left\{ F_l / \overline{N}^l, l \leq 2k \right\} \overline{\xi}(L), \tag{24}$$

it is independent of the average number density n [see eq. (2)]. Even if the number of objects in the catalog increases, the finite volume error remains *unchanged*.

Similarly, the *edge error* $\Delta_{\rm E}^2$ formally written as

$$\Delta_{\rm E}^2 = \mathcal{F}_{\rm E} \left\{ F_l / \overline{N}^l, l \leq 2k \right\} \overline{\xi} v / V, \tag{25}$$

is independent of average number density as well. Edge effects are caused by uneven weighting near the edges of the catalogue, thus increasing with v/V.

9

Finally, the *discreteness (shot noise) error* Δ_{D}^2 can formally be written as

$$\Delta_{\mathrm{D}}^2 = \mathcal{F}_{\mathrm{D}} \{ F_l, l \leq 2k \} \ v/V. \tag{26}$$

This is the only contribution to the cosmic error which depends on average number density n. To leading order in $\overline{N}$, it is proportional to $\overline{N}^{-k} v/V$ thus becomes important at small scales or when the number of objects N_{obj} in the catalog is small.

SC have tested successfully the validity of the above approximations by computing the cosmic error in artificial galaxy catalogs generated by Rayleigh-Lévy random walks, for which the clustering properties could be calculated exactly (e.g. CBSII).

When $\omega \neq 1$ and $\phi(r) \leq 1$, equation (22) generalizes to (see Appendix A)

$$\Delta_{\mathrm{cosmic}}^2[\omega, \phi] \simeq \Delta_{\mathrm{F}}^2[\omega] + \Delta_{\mathrm{E}}^2[\omega] + \Delta_{\mathrm{D}}^2[\omega, \phi], \tag{27}$$

with

$$\Delta_{\mathrm{F}}^2[\omega] = \frac{\Delta_{\mathrm{F}}^2}{\bar{\xi}(L)\hat{V}^2} \int_{\hat{V}} d^3 r_1 d^3 r_2 \omega(r_1)\omega(r_2)\xi(r_{12}), \tag{28}$$

$$\Delta_{\mathrm{E}}^2[\omega] = \frac{\Delta_{\mathrm{E}}^2}{\hat{V}} \int_{\hat{V}} d^3 r \omega^2(r), \tag{29}$$

$$\Delta_{\mathrm{D}}^2[\omega, \phi] = \frac{1}{\hat{V}} \int_{\hat{V}} d^3 r \omega^2(r)\Delta_{\mathrm{D}}^2(r). \tag{30}$$

The r dependence of Δ_{D}^2 in the integral (30) is caused by the selection function, as this type of error depends on the average count. On the other hand, the selection effects naturally canceled out in the finite volume and the edge errors, as expected.

The integral equation for the optimal weight becomes

$$\frac{\Delta_{\mathrm{F}}^2}{\bar{\xi}(L)\hat{V}} \int_{\hat{V}} d^3 u \omega(u)\xi(|r - u|) + \left\{ \Delta_{\mathrm{E}}^2 + \Delta_{\mathrm{D}}^2(r) \right\} \omega(r) + \lambda = 0, \tag{31}$$

and the constant λ is determined by the normalization (9). This standard equation can be solved numerically.

3.3. Interpretation

From equation (31), the following immediate conclusions can be drawn on how different contributions to the cosmic error influence the optimal weight:

1. If edge effects were dominant, the optimal weight would simply be uniform. This is contrary to intuition suggesting increasing weight at the edges to compensate for the lesser statistical weight carried by these regions of the catalog. This is how edge effects are corrected for the two-point function $\xi(r)$ (e.g. Ripley 1988). The finite extension of the cells, however, prevents us from correcting for edge effects with our indicator (8), which uses a simple multiplicative sampling weight. More involved additive correction for the moments of the *fluctuations* of counts in cells will be explained elsewhere using the formalism outlined in Szapudi & Szalay 1997b.

10

2. If discreteness effects were dominant, the correct weight approximately would be

$$\omega(r) \propto 1/\Delta_{\mathrm{D}}^2(r), \tag{32}$$

in agreement with intuition. Indeed, the effective number density $n(r)$ decreases with increasing distance from the observer, thus the shot noise error increases as well. This is compensated by ω decreasing with r.

3. Finite volume effects are difficult to predict without explicit calculations. The correlation function $\xi(r)$ is expected to follow approximately a power-law behavior from observations (Totsuji & Kihara 1969; Peebles 1974; Davis & Peebles 1983)

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma}, \quad r_0 \simeq 5 \ h^{-1} \ \mathrm{Mpc}, \quad \gamma \simeq 1.8, \tag{33}$$

for $0.1 \ h^{-1}$ Mpc $\lesssim r \lesssim 10 \ h^{-1}$ Mpc. At larger scales it decreases rapidly with scale, becoming negative at scales around $\sim 30 - 100 \ h^{-1}$ Mpc (e.g. Fisher et al. 1994; Tucker et al. 1996), although this turnaround value is presently uncertain. After that, it is expected to oscillate slowly around zero with very small amplitude. With such a behavior, it is not obvious to predict the optimal weight without explicit numerical calculations. The result will depend on the size and the geometry of the catalog. The only simplification is that if the finite volume error were dominant, the corresponding optimal weight would be independent of the order k.

3.4. Practical Measurements

As shown in § 3.2, even though it was considerably simplified with reasonable hypotheses, the calculation of the optimal weight and the cosmic error for F_k needs prior knowledge of F_l, $l \leq 2k$ (including F_k itself !) and of $\xi(r)$ (and of course of the selection function). As a result, we propose two procedures to perform a practical measurement in a galaxy catalog:

1. One possibility is to choose a model of large scale structure with given values of F_l, $l \leq 2k$ and $\xi(r)$. These values are used as input parameters in equation (31) to compute the optimal weight and the corresponding theoretical cosmic error [eq. (27)]. The value of F_k measured with this weight can be compared to the theoretical one, given the theoretical cosmic error. This procedure should be applied again to each competing model.

2. An alternative iterative approach starts with measuring directly the values of F_l, $l \leq 2k$ in the galaxy catalog, with a given weight, for example $\omega = 1$. With these values of F_l, one would solve equation (31) to find the optimal weight for measuring F_k, and thus perform a more accurate measurement. This can be repeated until convergence is achieved. We conjecture that a small number of iterations should be sufficient in practice. The main weakness of this model independent approach is that there is a cosmic error on the determination of the optimal weight itself. As discussed in SC (and earlier by CBSI and CBSII), the cosmic error is likely to be *systematic*, which implies that the optimal weight estimated this way might be *biased*. This bias, of course, would only increase the cosmic error on the measurement of F_k: the estimator for the moments is still unbiased by definition.

11

4. Example: the SDSS Catalog

The future SDSS will be a likely proving ground of the methods proposed in this paper. Therefore a "SDSS-like" catalog, $\mathcal{E}$, is used to illustrate the applicability of the theory outlined so far. This section is organized as follows. § 4.1 presents the properties of the hypothetical survey considered, i.e., its luminosity function, geometry, and the underlying statistics (function $\xi(r)$, factorial moments F_k). Then, in § 4.2, the optimal weight is computed with the corresponding cosmic error. A simple approximation for $\omega(r)$ is found which practically minimizes the cosmic error and avoids solving numerically integral equation (31). Finally, in § 4.3, we show the advantages of our optimal strategy of using the full catalog compared to the alternative of extracting volume limited subsamples. Details on the method used to solve numerically equation (31) are given in Appendix B.

4.1. Properties of the Survey

The luminosity function of the catalog $\mathcal{E}$ is assumed to be of the Schechter form (Schechter 1976)

$$\varphi(L/L_*) = \phi_*(L/L_*)^\alpha \exp(-L/L_*), \tag{34}$$

with parameters taken from Efstathiou, Ellis & Peterson (1988, see also Efstathiou 1996)

$$\alpha = -1.07, \quad \phi_* = 0.0156h^3 \text{ Mpc}^{-3}, \tag{35}$$

where h represents the uncertainty of a factor two on the Hubble constant: $H_0 = 100\ h$ km/s/Mpc. In what follows, we shall use $h = 0.5$, i.e.

$$H_0 = 50 \text{ km/s/Mpc}. \tag{36}$$

However, the results derived hereafter should not depend significantly on the value of H_0. In fact, the Hubble constant influences the shape of the power-spectrum of initial fluctuations through the parameter , $= \Omega h^2$, where Ω is the density parameter of the universe (e.g. Efstathiou, Bond & White 1992).

The average number density in a thin shell at distance r from the observer is

$$n(r) = \phi_*, \ [\alpha, L_{\lim}(r)/L_*], \tag{37}$$

where $L_{\lim}(r)$ is the minimum required luminosity for a galaxy at distance r from the observer to be included in the catalog, and , is the incomplete gamma function. If K-correction is neglected,

$$L_{\lim}(r)/L_* = 10^{0.4(M_* - M_{\lim} + 25 + 5\log_{10} r)}. \tag{38}$$

In the above equation, r is expressed in Mpc. Our choice of M_* and the magnitude limit $M_{\lim}$ is

$$M_* = -19.68 + 5\log(H_0/100), \quad M_{\lim} = 18.3. \tag{39}$$

Note that, with the above value of α, $n(r)$ diverges at $r = 0$, formally implying that the average number density of the real, total galaxy distribution is infinite. This, however, does not affect the following calculations.

For the geometry, we assume that the catalog covers a cone of depth

$$R_{\max} = 1200 \text{ Mpc} \tag{40}$$

with elliptic basis defined as follows in Cartesian coordinates:

$$z^2 = \left(\frac{x}{\tan \theta_{\mathrm{M}}} \right)^2 + \left(\frac{y}{\tan \theta_{\mathrm{m}}} \right)^2, \tag{41}$$

$$\theta_{\mathrm{M}} = 65°, \quad \theta_{\mathrm{m}} = 55°. \tag{42}$$

With the above choice of the parameters the catalog would typically contain

$$N_{\mathrm{obj}} \simeq 830000 \tag{43}$$

objects.

The underlying statistics is chosen as follows:

1. The two-point correlation function is an estimate of the nonlinear matter autocorrelation function for the standard Cold Dark Matter (CDM) model, which actually contains the H_0 dependence of the subsequent calculations. It is computed by Fourier transform of the power-spectrum obtained from the nonlinear ansatz of Peacock & Dodds (1994). The normalization is chosen such that the variance in a sphere of radius 8 h^{-1} Mpc is unity. The choice of the two-point function fixes its average over a cell $\overline{\xi}$ and therefore the factorial moment of order 2.

2. As the optimal weights for the measurement of F_k, $k \leq 4$ will be discussed, prior knowledge of the statistics up to $k = 8$ is needed. Factorial moments F_k, $3 \leq k \leq 8$ are derived from the measurements of Gaztañaga (1994) of the coefficients Q_N:

$$Q_3 = 1.35, \quad Q_4 = 2.33, \quad Q_5 = 4.02, \quad Q_6 = 6.7, \quad Q_7 = 10, \quad Q_8 = 12. \tag{44}$$

By definition,

$$Q_N \equiv \frac{1}{{}_{,N} v^N} \int_v d^3 r_1 \ldots d^3 r_N \xi_N(r_1, \ldots, r_N), \tag{45}$$

where ξ_N is the N-point correlation function (see e.g. Peebles 1980; BS) and

$${}_{,N} \equiv \frac{N^{N-2} \overline{N}^N \overline{\xi}^{N-1}}{N!}. \tag{46}$$

In general, Q_N can depend on scale. We assume that the hierarchical model applies, which implies that Q_N = constant independent of scale. However, the cosmic errors should be fairly robust against small variations of the cumulants.

The generating function of P_N [eq. (12)] can be expressed as

$$P(x) = \exp\left[\sum_{N=1}^{\infty} (x-1)^N {}_{,N} Q_N \right]. \tag{47}$$

This is a completely general equation, true outside of the hierarchical model framework as well (BS; Szapudi & Szalay 1993a).

The factorial moments F_k can be computed from the coefficients Q_N, $\overline{N}$ and $\overline{\xi}$ through $F_k = [\partial/\partial x]^k P(x)|_{x=1}$ [see eq. (1)].

The volume limited subsamples $\mathcal{E}^i_{\mathrm{VL}}$ extracted from our virtual catalog $\mathcal{E}$ are of depths $R_i = 200$, 400, 600, 800, 1000 and 1200 Mpc. Typically, they are expected to contain respectively $N_{\mathrm{obj},i} \simeq$ 40100, 138 500, 208 800, 211 800, 164 900 and 104 200 objects.

13

4.2. The optimal weight and the corresponding cosmic error

The assumed catalog has a non-spherical geometry similar to the future SDSS. Therefore the weight should depend both on the angles and the distance from the observer r, except when the finite volume error contribution is negligible. For simplicity, however, only the radial direction will be considered, i.e. the cosmic error will be minimized in the subspace of functions $\omega(r)$. After integration over the angles, the cosmic error becomes

$$
\begin{aligned}
\Delta_{\mathrm{cosmic}}^2[\omega, \phi] \quad \simeq \quad & \frac{\Delta_{\mathrm{F}}^2}{\overline{\xi}(L)\hat{V}^2} \int_{\hat{R}_{\mathrm{min}}}^{\hat{R}_{\mathrm{max}}} \int_{\hat{R}_{\mathrm{min}}}^{\hat{R}_{\mathrm{max}}} r_1^2 dr_1 r_2^2 dr_2 \hat{\Omega}(r_1)\omega(r_1)\hat{\Omega}(r_2)\omega(r_2)\tilde{\xi}(r_1, r_2) \\
& + \frac{\Delta_{\mathrm{E}}^2}{\hat{V}} \int_{\hat{R}_{\mathrm{min}}}^{\hat{R}_{\mathrm{max}}} r^2 dr \hat{\Omega}(r)\omega^2(r) \\
& + \frac{1}{\hat{V}} \int_{\hat{R}_{\mathrm{min}}}^{\hat{R}_{\mathrm{max}}} r^2 dr \hat{\Omega}(r)\omega^2(r)\Delta_{\mathrm{D}}^2(r),
\end{aligned} \tag{48}
$$

where $\hat{\Omega}(r)$ is the solid angle covered by cells at distance r from the observer, $\hat{R}_{\mathrm{min}}$, $\hat{R}_{\mathrm{max}}$ denote the distance of the closest and furthest cell to the observer, and

$$
\begin{aligned}
\tilde{\xi}(r_1, r_2) \quad \equiv \quad & \frac{1}{\hat{\Omega}(r_1)\hat{\Omega}(r_2)} \int_{\hat{\Omega}(r_1)} \int_{\hat{\Omega}(r_2)} \sin\theta_1 d\theta_1 d\varphi_1 \sin\theta_2 d\theta_2 d\varphi_2 \\
& \times \xi\left([r_1^2 + r_2^2 - 2r_1 r_2(\cos(\varphi_1 - \varphi_2)\sin\theta_1 \sin\theta_2 + \cos\theta_1 \cos\theta_2)]^{1/2} \right).
\end{aligned} \tag{49}
$$

In equation (48), one can identify the three lines respectively as the finite volume, the edge effect and the shot noise contribution. The optimal radial weight is the solution of the following integral equation (provided that a solution exists),

$$
\frac{\Delta_{\mathrm{F}}^2}{\overline{\xi}(L)\hat{V}} \int_{\hat{R}_{\mathrm{min}}}^{\hat{R}_{\mathrm{max}}} u^2 du \hat{\Omega}(u)\omega(u)\tilde{\xi}(r, u) + \left\{ \Delta_{\mathrm{E}}^2 + \Delta_{\mathrm{D}}^2(r) \right\} \omega(r) + \lambda = 0. \tag{50}
$$

Figure 1 shows the optimal weight given by the numerical solution of equation (50) in various situations (see Appendix B for details of the numerical method). The finite volume error contribution is assumed to be dominant for this plot. In this case, the optimal weight is independent of the statistical object under study, i.e., for F_k, of the order k. The long dashes correspond to a catalog with similar characteristics as our SDSS-like catalog, but covering the full sky, in order to have the exact solution $\omega(r)$ only depending on r. The dots correspond to the same situation, but the two-point correlation function is assumed to be $\overline{\xi}(\ell) = (\ell/16)^{-1.8}$ over all the available dynamic range, with ℓ expressed in Mpc. The four continuous curves correspond to four volume limited catalogs $\mathcal{E}_{\mathrm{VL}}^i$ of our SDSS-like survey, of depths $R_i = 200, 600,$ and 1000 and 1200 Mpc. The case $R_i = 1200$ Mpc is valid as well for the parent sample $\mathcal{E}$. The discontinuities on the extremities of the curves are boundary effects related to the way we discretize the integral equation (50). However, except for the left extremity of the dotted curve, they are likely to express the fact numerically that the optimal weight from finite volume effects is singular at the edges of the survey, at least for the right extremity of the dotted curve (see Appendix B.2). There are also some small irregularities on the solid curves, but these are random fluctuations due to the finite number of steps in the Monte-Carlo simulation used for computing the angular average (49) (see Appendixes B.1, B.2). Apart from these details, we see that the optimal weight is a smooth function, close to unity. Actually, taking
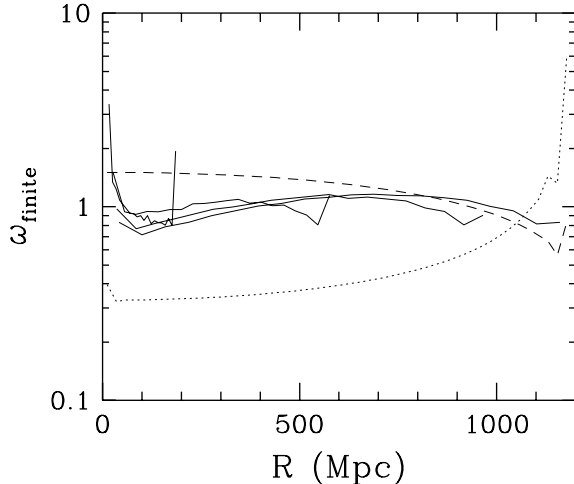
14

Fig. 1.— The optimal weight ω for measuring the factorial moment $F_k(\ell)$ ($k \geq 1$) is plotted as a function of distance R from the observer in the case the finite volume error is dominant. The results displayed here correspond to $\ell = 10$ Mpc; other values of ℓ are similar. The solid curves correspond to various volume limited subsamples $\mathcal{E}_{\mathrm{VL}}^i$ extracted from our SDSS like catalog $\mathcal{E}$ (see § 4.1). The depth R_i of the subsamples increases with the x-coordinate of the right end point of the curves: respectively $R_i = 200, 600, 1000$ and 1200 Mpc. The latter case is valid for the parent sample $\mathcal{E}$ as well. The dashed curve corresponds to a catalog exactly the same as $\mathcal{E}$ but covering the full sky. The dotted curve is the same, but the two-point function is assumed to be a power-law $\overline{\xi}(\ell) = (\ell/18)^{-1.8}$ over all the available dynamic range.

$\omega = 1$ gives almost the same value as the optimal weight for the finite volume error, at least for all the examples considered here. Thus a homogeneous weight $\omega = 1$ approximately minimizes the finite volume error in the space of radial weights $\omega(r)$. Rigorously, this result is not necessarily true for a catalog with a complicated geometry. However, it should be valid if the catalog is compact enough, which will be assumed in the following.

As a result, we propose the following approximation for the optimal weight, in the general case:

$$\omega(r) \propto 1 \left/ \left[\Delta_{\mathrm{F}}^2 + \Delta_{\mathrm{E}}^2 + \Delta_{\mathrm{D}}^2(r) \right] \right. . \tag{51}$$

The coefficient of proportionality is determined by the normalization (9). This approximation is quite natural, because it properly takes into account the relative weight of each contribution to the cosmic error. If one of them is dominant, then ω converges to the correct solution of the integral equation (31), although only approximately if the finite volume error is dominant.

Figure 2 shows the optimal (radial) weight obtained from the numerical solution of the integral equation (50) (solid curve on each panel). Each line of panels corresponds to a given value of k, which increases from top to bottom. Each column of panels corresponds to a fixed value of the scale ℓ (from left to right, $\ell = 0.1$ Mpc, 1 Mpc and 10 Mpc). The dashes and dots display the case when the finite volume error is negligible, and dominant, respectively. The long dashes show approximation (51). They overlap surprisingly well with the solid curves.

Note that at large scales, the optimal weight tends to unity, because discreteness effects become negligible, and edge effects dominant (e.g. SC). (The case of $\ell = 100$ Mpc is not shown, since it is
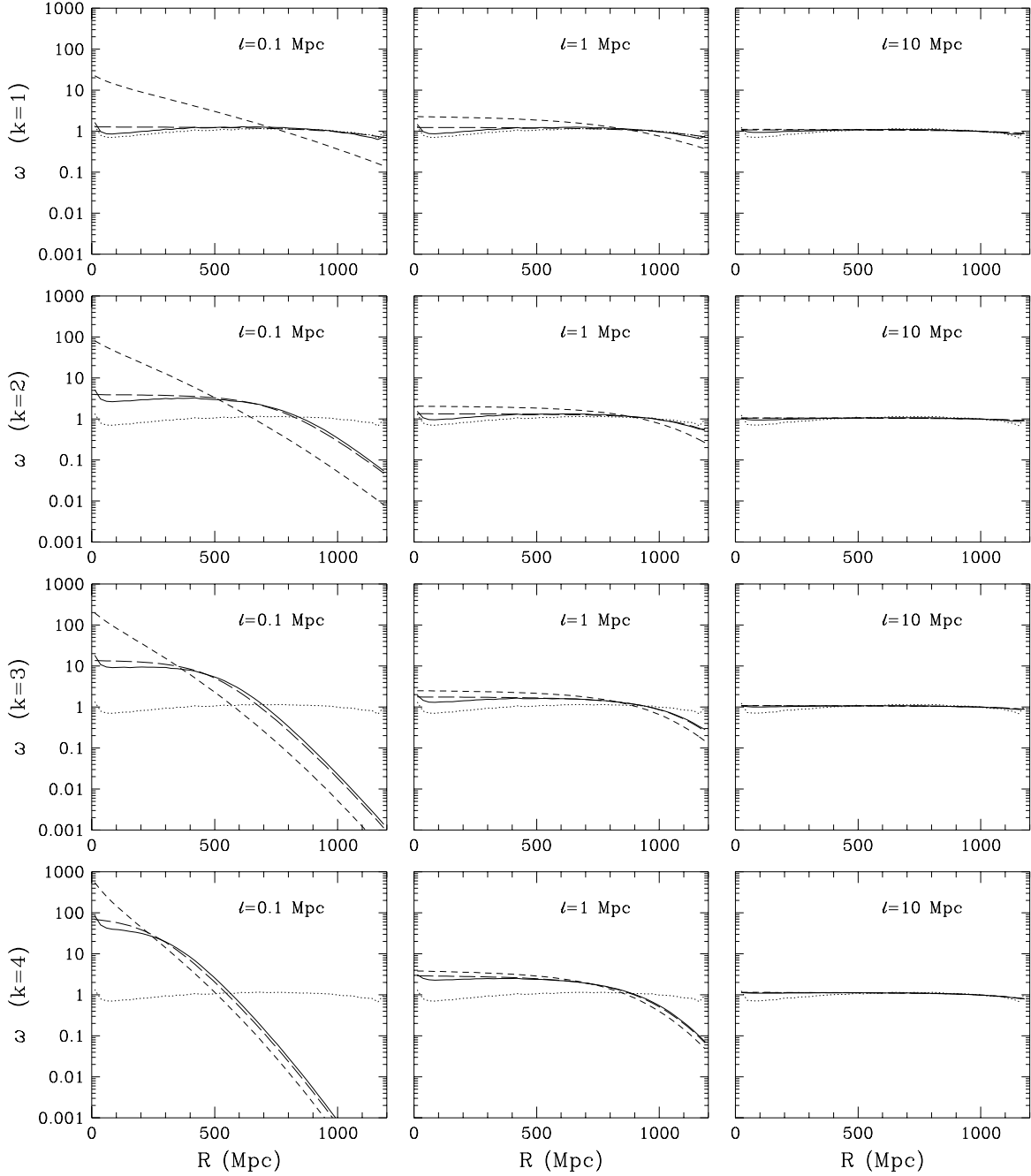
15

Fig. 2.— The optimal weight for measuring F_k in our virtual SDSS like catalog (see § 4.1) is plotted as a function of the distance R from the observer (solid curves). Each panel corresponds to a choice of (k, ℓ). The order k increases from top to bottom and the scale from left to right. The dashed curves are in the assumption that the finite volume error is negligible, while the reverse is true for the dotted curves. The long-dashed curves correspond to our proposed approximation (51) for the optimal weight.

quite similar to $\ell = 10$ Mpc). The departure of ω from unity on the other hand increases at higher order k, and smaller scales. Then discreteness effects tend to dominate the cosmic error (e.g. SC), and, as discussed in § 3.3, the weight strongly (exponentially) decreases with r. These arguments are partly illustrated by Figure 3. For each value of k, various contributions to the cosmic error calculated from equation (48) using the optimal weight, are plotted as functions of scale. The solid, dotted-dashed, and long dashed-short dashed lines correspond respectively to the cosmic error, the finite volume error, and the edge effect plus shot noise contribution.

In figure 4, the cosmic error is displayed as a function of scale for all values of k. The solid lines correspond to the result given by the optimal weight and the dots to $\omega = 1$. There are also dotted-dashed lines almost perfectly matching the solid ones: they correspond to approximation (51). The degree of matching suggests that this is indeed an excellent approximation. Triangles, squares, hexagons and circles respectively correspond to $k = 1, 2, 3$ and 4: the cosmic error increases with the order k. According to the figure, $\omega = 1$ provides a satisfactory weighting scheme for our mock SDSS catalog on scales larger than ~ 1 Mpc.

4.3.   Volume limited subsamples versus full catalog

According to the above analysis, the choice $\omega = 1$ approximately minimizes the finite volume error. As there is no selection effect in an homogeneous catalog, the discreteness error is minimized as well with $\omega = 1$; finally, so is the edge effect contribution (§ 3.3). This confirms the common wisdom, that the optimal weight in an homogeneous catalog is $\omega \simeq 1$, be it volume limited or a two-dimensional galaxy catalog.

In figure 5, the cosmic error on the factorial moments is displayed as a function of scale for our catalog $\mathcal{E}$ and its volume limited subsamples $\mathcal{E}_{VL}^i$ defined in end of § 4.1. From top to bottom, we have $k = 1, 2, 3$ and 4. The solid lines correspond to $\mathcal{E}$ with optimal weight. The dots, dashes, long dashes, dot-dashes, dot-long dashes and long dashes-short dashes correspond respectively to $R_i = 200, 400, 600, 800, 1000$ and 1200 Mpc.

The figure illustrates clearly that a measurement with optimal weights using the full catalog yields smaller variance than any volume limited subsample (hereafter VLS). Small VLSs (with small depth) are denser than large VLSs (with large depth), so the shot noise error is more significant on the latter than on the former. The opposite is true for the finite volume error and the edge effect error. Large VLS are thus suited only for probing large scales, while small VLS probe small scales, especially at high order k. According to this argument, it is possible to construct a VLS which is fine tuned for a particular scale. At *this* scale, the cosmic error is almost (but not quite) as small as the one obtained from the optimal weights from the full survey. However, this is not true for other scales, i.e. the dynamic range is quite narrow. Therefore, a series of VLSs has to be constructed, each optimized for a different scale. As a result, the collection of VLSs can achieve almost as small errorbars as the optimal measurement on the full catalog only at the expense of a lot more work. In summary, a single optimal measurement yields smaller errorbars more efficiently than a strategy based on a series of VLSs.

5.   Sparse sampling strategies

So far we dealt with the problem of extracting information from existing catalogs in an optimal way. Another degree of freedom arises, during of the design of a survey. Next we will be concerned
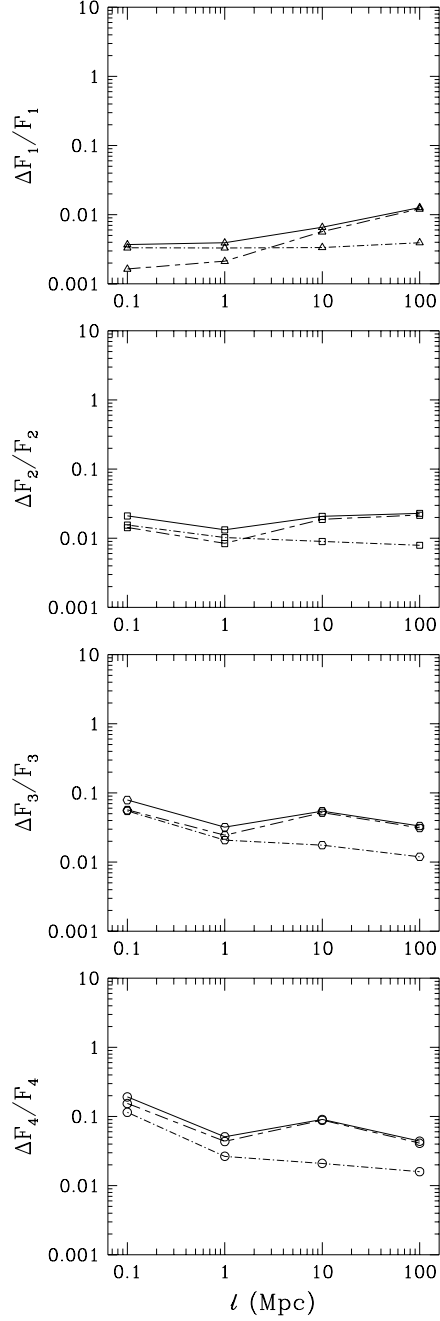
Fig. 3.— The cosmic error is shown as a function of scale, when the factorial moments F_k are measured with the optimal weight in our virtual SDSS like catalog $\mathcal{E}$ (see § 4.1). Each panel corresponds to a value of the order k. The solid, dotted-dashed and long-dashed short-dashed curves correspond respectively to the total error, the finite volume, and the edge plus discreteness effect contribution.
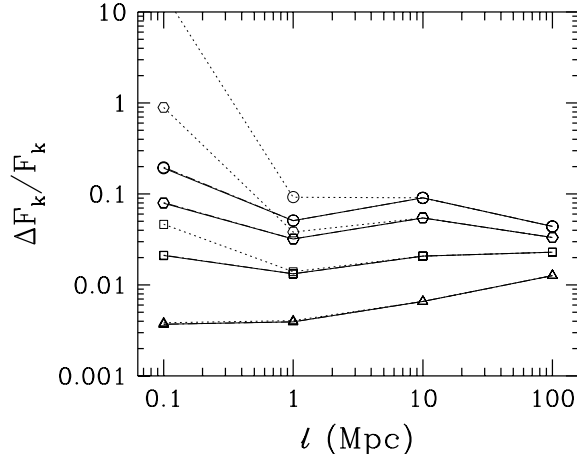
18

Fig. 4.— The cosmic error for our virtual SDSS like catalog $\mathcal{E}$ (see § 4.1) is shown as a function of scale, when the factorial moments F_k are measured with the optimal weight (solid curves), approximation (51) (dotted-dashed curves almost perfectly matching the continuous lines), and uniform weights (dots). The triangles, squares, hexagons and circles respectively correspond to $k = 1, 2, 3$ and 4.

with the optimal design, especially with the optimal use of the available telescope time. This in turn inevitably leads to the issues of sparse sampling and optimal survey geometry. We discuss them as follows:

1. In the spirit of K86 (see also the recent work of Heavens & Taylor 1997), we explore the question: given a fixed amount of telescope time, how to build a statistically optimal three-dimensional magnitude limited catalog, if one has the freedom to sample randomly a fraction $f \leq 1$ of the visible galaxies? As mentioned in the introduction, a small sampling rate f allows the construction of a deep but sparse survey. This results in small finite volume and edge errors, but large discreteness effects. The reverse is true when f is large. The best compromise between these requirements yields the optimal sampling rate, which depends on the scale considered and on the statistic. Here, we extend the calculations of K86 using a more accurate estimate of the cosmic error. While originally only the two-point function was considered, the optimal sampling rate will be calculated for higher order factorial moments, F_k, $k \leq 4$. In § 5.1, to simplify the analysis, the survey is assumed to have full sky coverage. The conclusions, however, do not depend significantly on this assumption. We also suppose that the redshifts are collected individually. In § 5.2, the changes brought by multi-fiber spectroscopy are discussed, as today this is the most widespread method for collecting redshifts.

2. Following K96, § 5.3 considers the question of optimal survey geometry. A survey is assumed to cover a fraction of the sky with redshifts collected by multifiber spectrographs. Thus the catalog can be naturally decomposed into small patches corresponding to the field of view of the telescope. The design goal is the optimal arrangement of these patches. Some of the choices are compact, elongated (or VLA like), or quasi randomly spread over the sky. This issue is immensely complicated by several non-linear factors and details of the actual
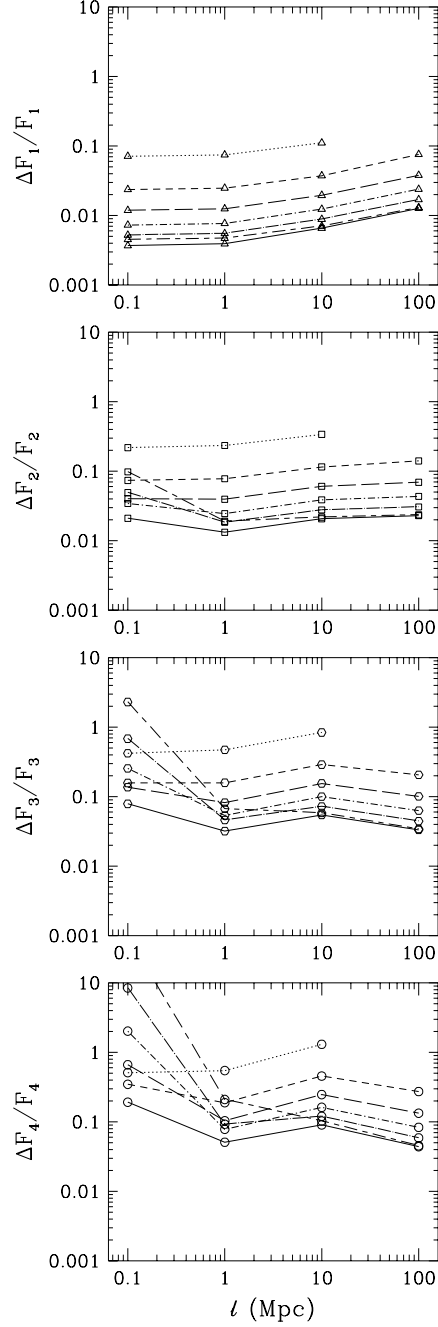
19

Fig. 5.— The cosmic error is displayed as a function of scale for our virtual SDSS like catalog $\mathcal{E}$ and various volume-limited subsamples $\mathcal{E}^i_{\rm VL}$ (see § 4.1), as expected for factorial moments F_k measured with the optimal weight. Each panel corresponds to a value of k, increasing from top to bottom. The solid curves correspond to the parent sample $\mathcal{E}$. The dots, short dashes, long dashes, dot-dashes, dots-long dashes, short dashes-long dashes correspond respectively to the subsamples of depth $R_i = 200, 400, 600, 800, 1000$ and 1200 Mpc. Note, that for $R_i = 200$ Mpc, there is a point missing for $\ell = 100$ Mpc, as a sphere of such radius is too large to be included in the subsample.

parameters of the proposed survey. Accordingly, we only attempt to illustrate the problem, and give an approximate solution under generic circumstances.

5.1. Full Sky Survey

Given the luminosity function of equation (34), and assuming, as in K86, that the time required for measuring the redshift of a galaxy is inversely proportional to the luminosity L, the total telescope time required for constructing a magnitude limited catalog is

$$T_{\text{total}} \propto \int_0^{R_{\max}} r^4 dr \phi_*, \; [\alpha - 1, L_{\lim}(r)/L_*], \tag{52}$$

where $R_{\max}$ is the depth of the survey, and the redshifts are assumed to be collected individually. The magnitude limit $M_{\lim}$ of the catalog is related to its depth $R_{\max}$ through

$$M_{\lim} = 5 \log_{10} \left(\frac{R_{\max}}{R_{\text{ref}}} \right) + M_{\text{ref}}, \tag{53}$$

where R_{ref} and M_{ref} are constants. This implies that

$$T_{\text{total}} \propto \phi_* R_{\max}^5, \tag{54}$$

in agreement with K86. Thus sampling a fraction f of the galaxies for a fixed telescope time results in

$$R_{\max} = R_{\text{ref}} f^{-1/5}. \tag{55}$$

Given a scale ℓ and a value of the order k, the optimal sampling rate f by definition minimizes the cosmic error of the factorial moment $F_k(\ell)$. Before any analytical estimates of the optimal sampling rate, let us consider an example: a full-sky survey $\mathcal{S}(f)$, with luminosity function and statistics identical to § 4.1. We choose

$$M_{\text{ref}} = 15.5, \quad R_{\text{ref}} = 391 \text{ Mpc}. \tag{56}$$

For $f = 1$ this is roughly a full sky CfA2 catalog, although actually denser (see, e.g. de Lapparent, Geller & Huchra 1989). On average it contains

$$N_{\text{obj}}(f = 1) \equiv N_{\text{ref}} \simeq 72400 \tag{57}$$

galaxies.

For this hypothetical survey, the optimal sampling rate was found numerically by calculating the cosmic error as explained in detail in Appendix B.3. Figure 6 shows the results for F_k, $1 \leq k \leq 4$, as a function of scale. The symbols (respectively triangles, squares, pentagons and hexagons for $k = 1, 2, 3$ and 4) take into account effects of the selection function by using the approximation (51) for the optimal weight ω. The curves (respectively dots, short dashes, long dashes, dot-dashes for $k = 1, 2, 3$ and 4) have an uniform selection function: they correspond to a homogeneous sample of same size $R_{\max}(f)$ and involving the same number of objects $N_{\text{obj}}(f)$ as $\mathcal{S}(f)$. Apart from the small shift for $k = 1$, the curves superpose quite well to the symbols, showing that such an approximation is valid. This reduces significantly the complexity of the calculation of the optimal sampling rate. Note, however, that our choice of R_{ref} [eq. (56)] was not arbitrary: R_{ref} was chosen to be twice the
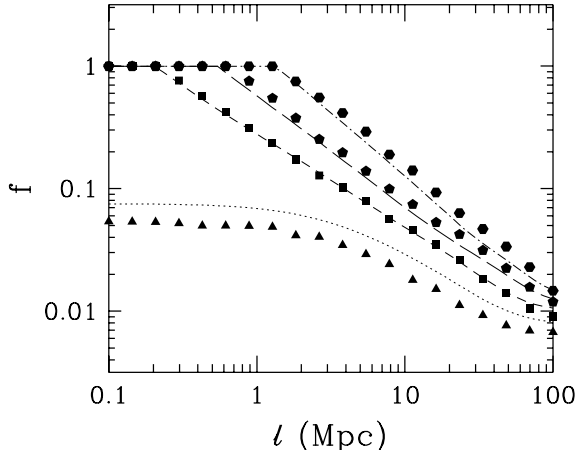
Fig. 6.— The optimal sampling rate f for measuring the factorial moment of order k is shown as a function of scale for our virtual survey $\mathcal{S}(f)$ (see text). The symbols use a realistic selection function, whereas the curves assume uniform selection. The triangles (dots), squares (short dashes), pentagons (long dashes) and hexagons (dot-dashes) correspond to $k = 1, 2, 3$, and 4, respectively.

radius of the volume-limited subsample which contains the largest number of objects (for $f = 1$). The following scaling is thus true

$$R_{\rm ref} 10^{-0.2 M_{\rm ref}} = {\rm constant} \equiv C_{\rm ref}. \tag{58}$$

With this choice of $R_{\rm ref}$, the sphere of radius $R_{\rm max}(f)$ includes most of the detectable galaxies. With a larger (smaller) value of $R_{\rm ref}$ than given by equation (56), the curves on figure 6 would be shifted upwards (downwards).

Figure 7 shows the cosmic error on the measured factorial moments as a function of the sampling rate f for various scales $\ell = 0.1, 1, 10$ and 100 Mpc. The symbols are the same as in figure 6. Although it is an excellent approximation for determining the optimal sampling rate, a uniform selection is inaccurate for estimating the cosmic error in general, except to some extent for $k \leq 2$, and for large scales otherwise.

From figure 6, the optimal sampling rate f exhibits a remarkable power-law behavior up to the saturation to unity, except for $k = 1$, when it rapidly converges to a value smaller than unity at small scales. Moreover, f increases with the order k, corresponding to the increasing relative contribution of the discreteness error with k. These features can be further explained by analytical calculations as follows.

According to the previous findings, calculations will be simplified by assuming a uniform selection. This appears to be good approximation for determining f. The analytical formulae of SC yield the relative cosmic error on the measurement of F_k, $k \leq 3$. This can be used to find the optimal sampling rate in the weakly and highly non-linear regimes. The details of the calculations can be found in appendix C.

In the highly nonlinear regime, $\overline{\xi} \gg 1$, i.e. at small scales, edge effects are expected to be negligible compared to finite volume effects (e.g. SC). The optimal sampling rate yields the best compromise between discreteness and finite volume effects. If $R_{\rm ref}$ is sufficiently large, $\overline{\xi}(R_{\rm max})$ is expected to
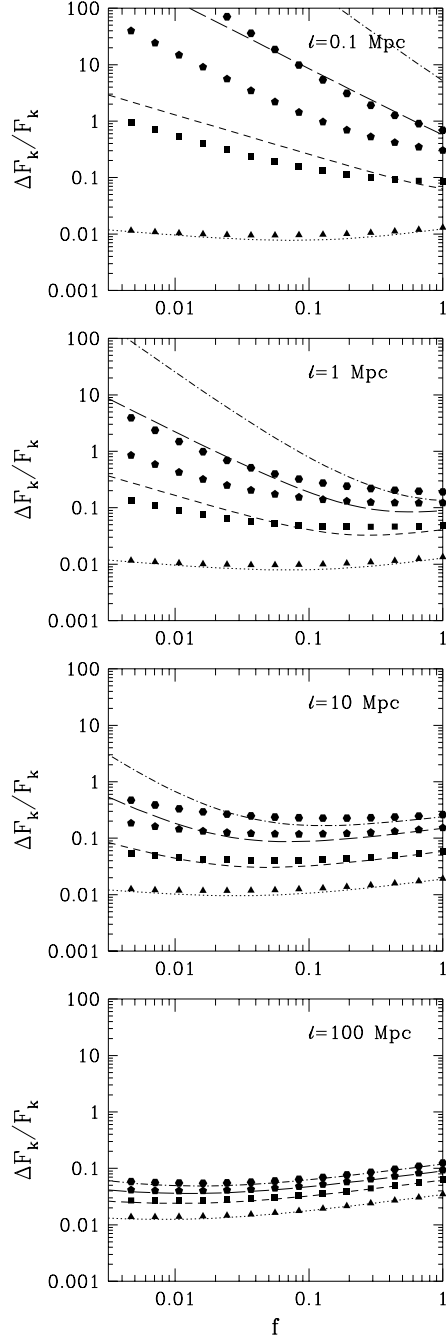
22

Fig. 7.— The cosmic error on the measured factorial moment of order k is displayed, as a function of the sampling rate f for our virtual survey $\mathcal{S}(f)$. Each panel corresponds to a given choice of scale $\ell = 0.1$, 1, 10 and 100 Mpc from top to bottom. As in figure 6, the symbols use a realistic selection function whereas the curves correspond to uniform selection.

23

exhibit an approximate power-law behavior. For a Harrison-Zeldovich power-spectrum $\langle|\delta_k|^2\rangle \propto k$,

$$\overline{\xi}(R_{\max}) \simeq (R_{\max}/L_0)^{-\gamma_L}, \quad \gamma_L = 4. \tag{59}$$

The power-law behavior of the average correlation function in the nonlinear regime is a good approximation [eq. (33)], and, although not absolutely necessary, simplifies the computations:

$$\overline{\xi}(\ell) = (\ell/\ell_0)^{-\gamma}, \quad \ell \lesssim \ell_0. \tag{60}$$

With the above hypotheses, the optimal weight, when not saturated to unity, is

$$f \simeq \left[\frac{5k-3}{\gamma_L} \frac{\alpha_k^{\mathrm{D}}}{\alpha_k^{\mathrm{F}}} \frac{R_{\mathrm{ref}}^{\gamma_L+3(k-1)}}{L_0^{\gamma_L}} \frac{N_{\mathrm{ref}}^{-k}}{\ell_0^{\gamma(k-1)}} \right]^{\frac{5}{5k+\gamma_L-3}} \ell^{-\frac{5(3-\gamma)(k-1)}{5k+\gamma_L-3}}, \quad \overline{\xi} \gg 1. \tag{61}$$

In this expression, the quantities α_k^{D} and α_k^{F} are numbers depending on the order k and on Q_l, $l \leq 2k$. Their ratio writes, for $k \leq 3$,

$$\frac{\alpha_1^{\mathrm{D}}}{\alpha_1^{\mathrm{F}}} = 1, \quad \frac{\alpha_2^{\mathrm{D}}}{\alpha_2^{\mathrm{F}}} \simeq \frac{1}{8Q_4}, \quad \frac{\alpha_3^{\mathrm{D}}}{\alpha_3^{\mathrm{F}}} \simeq \frac{Q_3}{47.4 Q_6}. \tag{62}$$

Equation (61) is not valid for the Gaussian case, except for $k = 1$. Then f is independent of scale when $\overline{\xi} \gg 1$, explaining the result obtained for the triangles and the dots in figure 6.

In the weakly nonlinear regime, $\overline{\xi} \ll 1$, i.e. at large scales, edge effects typically dominate over finite volume effects (e.g. SC). Thus the optimal sampling rate f results from a competition between edge effects and discreteness effects (see Appendix C for the details). At large scales, $\overline{\xi}$ is not a power-law, (except for *very* large scales, $\ell \gtrsim 100$ Mpc), neither is the optimal sampling rate

$$f = \frac{2k^2}{3\alpha_k^{\mathrm{E}}} \frac{R_{\mathrm{ref}}^3}{N_{\mathrm{ref}}} \frac{1}{\overline{\xi}\ell^3}, \quad \overline{\xi} \ll 1, \tag{63}$$

where $k^2/\alpha_k^{\mathrm{E}}$ is slowly increasing with k:

$$\frac{1}{\alpha_1^{\mathrm{E}}} \simeq 0.18, \quad \frac{4}{\alpha_2^{\mathrm{E}}} \simeq 0.23, \quad \frac{9}{\alpha_3^{\mathrm{E}}} \simeq 0.26. \tag{64}$$

For $\gamma < 3$ the optimal weight is a decreasing function of scale. It depends on the order k considered, and, for nonlinear scales, on the details of the higher order statistics through the ratio $\alpha_k^{\mathrm{D}}/\alpha_k^{\mathrm{F}}$. However, it has a weak dependence on the total telescope time T_{total} according to the following argument: $T_{\mathrm{total}} \propto R_{\mathrm{ref}}^5$, $N_{\mathrm{ref}} \propto R_{\mathrm{ref}}^3$, therefore in the highly nonlinear regime [eq. (61)],

$$f \propto T_{\mathrm{total}}^{\frac{\gamma_L-3}{5k+\gamma_L-3}}, \quad \overline{\xi} \gg 1, \tag{65}$$

constituting a weak dependence of the optimal sampling rate on the total observing time, at least for $k \geq 2$. In the weakly nonlinear regime, $\overline{\xi} \ll 1$, f does not depend at all on T_{total}. Consequently, the results displayed in figure 6 are more general than initially suspected: they should be roughly valid for any full sky survey of depth larger than a few hundred Mpc (so that $\overline{\xi}(R_{\max})$ is approximately a power-law), to the extent that the assumptions we have made for the underlying statistical properties (see § 4.1) are realistic.

To give orders of magnitude, let us compute numerically what would be typically the optimal sampling rate from formulae (61) and (63). Assuming $H_0 = 50$ km/s/Mpc, a reasonable choice for the correlation length in equation (59) is $L_0 \simeq 50$ Mpc. The same values are taken as in equations (56) and (57) for R_{ref} and N_{ref}. With $\gamma = 1.8$, $\ell_0 = 16$ Mpc and the values of Q_N quoted in equation (44), from equation (61) the optimal weight f_k is obtained corresponding to each value of k in the highly nonlinear regime:

$$f_1 \sim 0.05, \tag{66}$$

$$f_2 \sim \min(1, 0.2\ell^{-0.55}), \tag{67}$$

$$f_3 \sim \min(1, 0.3\ell^{-0.75}). \tag{68}$$

Setting $\ell = 10$ Mpc in equation (67) gives $f_2 \sim 1/18$, in rough agreement with figure 6. This result is similar to the findings of K86. This is not surprising, although his calculation was done for rather larger scales $\ell \sim 30$ Mpc. In the weakly nonlinear regime, equation (63) must be used. For a CDM spectrum normalized to COBE ($\overline{\xi}(16$ Mpc$) \simeq 1.22^2$, see, e.g. Bunn & White (1997)) and $\ell = 100$ Mpc, $f_2 \sim 1/135$ would be optimal: a quite small sampling rate, in rough agreement with figure 6.

More importantly, however, the sampling rate has to be optimized for range of scales and statistics, not only for fixed values of k and ℓ. The optimal sampling rate for $(k, \ell) = (2, 100)$, $f \sim 1/130$, is not optimal for other scales and orders: according to figure 7 (symbols), it dramatically penalizes small scales, especially when k is large: for $k = 4$ and $\ell = 1$ Mpc, the error is about one order of magnitude larger than for $f = 1$. Conversely, $f \simeq 1$ is a good choice for small scales, however it would increase the errors on large scales by a factor of $2 \div 3$. A good compromise seems to be f of order $1/10$, in approximate agreement with the initial findings $f = 1/20$ of K86. This choice unfortunately only optimizes for low-order statistics, $k \leq 4$. The higher is k, the higher should be f. In particular, if one wants to analyze the probability distribution function of the density field in terms of shape (e.g., Bouchet et al. 1993, CBSII), perform cluster selection (Szapudi & Szalay 1993a, Szapudi & Szalay 1996), or any other investigation which depends on the full hierarchy of the factorial moments, a sparse sampling strategy seems inappropriate (as already stated in CB-SII). In summary, although for low order moments a sampling rate $f \leq 1$ can be found, which globally optimizes the measurement, higher order moments require full sampling. This conclusion is further supported by the arguments of the next section, where it is shown that, with the advent of multifiber spectroscopy, there is more to be lost by undersampling than any possible gain.

Previously, we assumed that the telescope time t_{obs} required to measure the redshift of a galaxy of luminosity L was proportional to $1/L$. This assumption probably breaks down for faint objects when the background noise dominates. Then the scaling $t_{\text{obs}} \propto 1/L^2$, as used by Heavens & Taylor (1997), would probably be more appropriate. Using this scaling does not significantly change the previous results and conclusions. The greatest difference is at large scales, where the optimal sampling rate becomes at most twice as small compared to figure 6, where $t_{\text{obs}} \propto 1/L$ was assumed[3]. Note also that when $t_{\text{obs}} \propto 1/L^2$, assuming uniform selection is not seen to be a good approximation anymore, thus analytical estimates similar to equations (61) and (63) are quite inaccurate.

5.2. Multifiber spectroscopy

---

[3]This comparison supposes that the optimal sampling rate does not depend on total telescope time.

The above calculation was assuming that the redshifts are collected individually. In today's astronomy, the problem is complicated by the fact that multifiber spectrographs are used to collect redshifts. With this technique it is possible to measure simultaneously N_S redshifts in a patch of the sky of fixed angular size Ω_S. This limits the validity of the previous calculations to the case when the typical number of candidates in such a field, $N_{can} \propto f^{2/5}$, is larger than N_S, as already discussed in K86[4].

Even if the available telescope time is too short to allow $N_{can} \gtrsim N_S$, a sparse sampling strategy still can make sense. It is possible to spend more time per field in order to go deeper, and thus allow a smaller value of f. There is a price to pay for such strategy: if T_{total} is fixed, the total solid angle Ω covered by the survey will be reduced. More details of this argument can be found in K86. Here let us only remind the scaling of the relevant quantities with the sampling rate, such as the depth R_{max} of the catalog, its solid angle Ω and its volume $V = \Omega R_{max}^3/3$:

$$ R_{max} = R_{ref}f^{-1/3}, \quad \Omega = \Omega_{ref}f^{2/3}, \quad V = V_{ref}f^{-1/3}. \tag{69} $$

Similarly to K86, we are going to see how these new relations affect the optimal sampling rate, compared to § 5.1. Just as before, uniform selection will be assumed, because it was shown that the details of the selection function do not influence the results. Unfortunately, the varying shape of the catalog renders the calculation of the finite volume error costly: for each value of f, Ω changes, and a new calculation of the two-point correlation matrix $\tilde{\xi}(r_1, r_2)$ is required [eq. (49)]. This would go beyond the scope of this paper. Instead the finite volume error is simply calculated for a spherical catalog of same volume V. This is a reasonable approximation if the catalog is compact enough, although it would obviously fail for the extreme case of a pencil beam survey. We shall partly come back to that problem in § 5.3.

Similarly to § 5.2, and after the calculations detailed in Appendix C, the result in the highly nonlinear regime is

$$ f \simeq \left[\frac{9k - 3}{\gamma_L} \frac{\alpha_k^D}{\alpha_k^F} \frac{L_{ref}^{\gamma_L + 3(k-1)}}{L_0^{\gamma_L}} \frac{N_{ref}^{-k}}{\ell_0^{\gamma(k-1)}} \right]^{\frac{9}{9k + \gamma_L - 3}} \ell^{-\frac{9(3-\gamma)(k-1)}{9k + \gamma_L - 3}}, \quad \overline{\xi} \gg 1, \tag{70} $$

where the typical size L_{ref} of the catalog is defined as

$$ V_{ref} \equiv \frac{4\pi}{3} L_{ref}^3. \tag{71} $$

In the weakly nonlinear regime, $\overline{\xi} \ll 1$, the edge effects are likely to dominate over finite volume effects. To leading order in v/V the solution is insensitive to the shape of the catalog, although as shown in § 5.3, this is not true in extreme cases. The optimal sampling rate is

$$ f = \frac{2k^2}{\alpha_k^E} \frac{L_{ref}^3}{N_{ref}} \frac{1}{\overline{\xi}\ell^3}, \quad \overline{\xi} \ll 1. \tag{72} $$

Equations (70) and (72) are quite similar to eqs. (61) and (63) in § 5.1. For example, in the weakly nonlinear regime, the optimal sampling rate is roughly three times larger for multifiber

---

[4]By taking into account the galaxy density fluctuations from patch to patch over the sky, a more realistic constraint would be $N_{can}[1 - \sqrt{1/N_{can} + w(\Omega_S)}] \gtrsim N_S$, where $w(\Omega_S)$ is the average of the angular correlation function over the patch. This condition accounts for the expected Poisson fluctuations and intrinsic correlations of the density field, up to second order.

spectroscopy, than in § 5.1 (instead of the factor two of K86). In the highly nonlinear regime, the difference is apparently smaller, although this is probably partly due to our way of estimating the finite volume error, and changes slowly with scale. For multifiber spectroscopy, f is typically twice than for individual collection of redshifts for $k = 1$, 30% more for $k = 2$ and 20% more for $k = 3$. In the weakly nonlinear regime, our estimate of the cosmic error is accurate enough to compute the gain on the errors for optimal compared to full sampling, even though uniform selection was used (see lower-panel of Fig. 7). For a CDM spectrum normalized to COBE at $\ell = 100$ Mpc with the corresponding optimal sampling rate $f_2 \sim 1/45$, the gain is merely 1.5 in reduction of $\Delta F_2 / F_2$. This is to be compared with 2.8 for individual redshifts. These results qualitatively agree with K86: the optimal sampling rate is increased with multiple collection of redshifts compared to individual collection, and the corresponding gain in the cosmic error is smaller. This strengthens the conclusions of § 5.1: even if a sparse sampling strategy could be optimal for special applications, such as measuring low-order statistics, the resulting gain is too small compared to the corresponding increase of errors for high order statistics which are more sensitive to sampling. Another important point to note is that sparse sampling strategy consists of measuring deeper redshifts, thus less controllable systematic errors are also likely to increase, which are not included in the previous discussions.

5.3. Survey geometry

So far we neglected the dependence of the cosmic error and the corresponding optimal sampling rate f on catalog geometry: this amounts to replacing the geometry with a sphere of same volume. For a deep catalog covering only a small part of the sky, such as a pencil beam survey, this is clearly not a good approximation. The finite volume error depends significantly on the catalog geometry. So does the shot noise error[5], and particularly the edge effect error, especially when the cell size becomes comparable to the size of the largest cell contained in the catalog. This did not show up in our calculations to leading order in v/V: in that case, the edge effect and the shot noise errors do not depend on the catalog geometry.

Since geometry is another degree of freedom in the design of surveys, it can be used to achieve predetermined statistical goals. For typical redshift surveys, the geometry has an "atomic" building block, Ω_S, corresponding to the field of view of the telescope. Within this, a certain number of spectra can be collected. The design problem which will be discussed in this section is (K96): what is the optimal distribution of fields of size Ω_S on the sky? Note that this problem is relevant even to redshift surveys, which, as the SDSS, plan to uniformly cover a large portion of the sky. Since such a survey takes a long time to carry out, it is important that in the initial phases the individual fields should be placed in such a manner that as much preliminary information as possible could be extracted. Such strategy can not only provide preliminary results before the full completion of the survey, but it gives an excellent early check, whether the envisioned goals of the survey can be achieved when completed. Note that in such a case, the design problem acquires another dimension, i.e. time, since analysis can be performed at several stages before full completion. Generalization of the following arguments for this is trivial. Some of the possibilities for the geometry are a compact structure, an elongated one (perhaps VLA-like), or a (quasi) random distribution, etc. Judging

---

[5]through "edge-discreteness" effects as mentioned in the introduction.

all possibilities is extremely difficult, not to mention that for realistic surveys other "worldly" factors play important roles, such as weather, dark time, season, etc., therefore the aim is here only to illustrate the problem and present some practical suggestions by solving it under simplified circumstances.

Each field corresponds to an elementary galaxy catalog of volume $V_S = \Omega_S R_{max}^3/3$ at position $\hat{n}_i$ in the sky. The centrally important quantity for the finite volume error, the integral of the two-point function over the whole survey can be decomposed as

$$\overline{\xi}(L) = \frac{1}{M_S^2} \sum_{i \neq j} w(\theta_{ij}) + \frac{1}{M_S} \overline{\xi}(L_S). \tag{73}$$

In this equation, M_S is the total number of fields. The angular correlation function $w(\theta)$ is defined by

$$w(\theta) \equiv \frac{1}{V_S^2} \int d^3 r_1 \int d^3 r_2 \xi(r_{12}), \tag{74}$$

where each integral is performed on disjoint elementary catalogs with positions $\hat{n}_1$ and $\hat{n}_2$ on the sky such that $\theta = (\hat{n}_1, \hat{n}_2)$. The quantity $\overline{\xi}(L_S)$ is the average of ξ over an elementary catalog:

$$\overline{\xi}(L_S) \equiv \frac{1}{V_S} \int_{V_S} d^3 r_1 d^3 r_2 \xi(r_{12}). \tag{75}$$

According to equation (73), values of θ_{ij} could be chosen, which minimize $\overline{\xi}(L)$ and therefore the finite volume error. This choice would not affect the edge effect error and the shot noise error significantly: they do not depend on the geometry of the catalog to leading order in v/V, i.e. at small scales. Although the solution depends on the large scale behavior of the two-point function, we conjecture the best configuration to be a "glass" spread over the largest possible area of the sky. This configuration would maximize the distance between the fields, and, in agreement with intuition, it would decrease the coherence of the fields as much as possible. Clearly, the "compact" configuration is the worst, followed by the "line" configuration. These conclusions are very similar to K96, who discussed optimal survey strategies for measuring weak lensing related two-point statistics in Fourier space.

However, the above arguments are relevant only if the considered scale ℓ is small compared to the size ℓ_{max} of the largest cell included in the catalog. When ℓ becomes comparable to ℓ_{max}, our calculations of the full cosmic error are not valid anymore, although the conclusion remain approximately valid for the finite volume error. The exact calculation of the cosmic error in the regime where $\ell \lesssim \ell_{max}$ is quite tedious (see appendix B of SC for an example). Qualitative description, however, can be given. The shot noise error and the edge effect error increase with $v/\hat{V}$, where $\hat{V}$ is the volume occupied by positions of cells *included* in the catalog. This ratio increases toward larger scales both because v is increasing and $\hat{V}$ decreasing. This latter effect is obviously more prominent for "glass" , or "line" configurations. The edge effect error (and the shot noise error) is therefore more important at large scales for these configurations, than for the compact geometry.

As before, the choice of the optimal geometry of the catalog results from the competition between various contributions to the cosmic error, namely the finite volume error against the edge effect and the shot noise errors. The details of the sampling strategy depend on the value of Ω_S, on total telescope time, the needed dynamic range in scales, the statistical aim, and the clustering properties of the universe. A good compromise could be a survey composed of several compact

28

subsamples of size $\Omega_{\rm F} \geq \Omega_{\rm S}$, spread over a large fraction of the sky on a glass like structure. The optimal value of $\Omega_{\rm F}$ is a complicated function of the various parameters of the survey. While the details of this difficult issue are left for subsequent research, it is clear that the above arguments give an approximate solution once the survey parameters and the scale range are fixed.

6. Discussion

In this article, we examined extensively the measurement of low-order moments of the count probability distribution function (CPDF) in three-dimensional magnitude limited galaxy catalogs, with special emphasis on issues related to the effects of the selection function. A new estimator was proposed: the weighted factorial moment of count-in-cells, corrected for selection effects [eq. (8)]. The following questions were studied in detail:

(i) Given a catalog, what is the optimal way of measuring factorial moments?

(ii) What is the optimal sampling strategy for constructing a catalog to measure factorial moments?

Both of these question are intimately related to the variance of the proposed unbiased estimator. Thus the cosmic error was computed, by extending the calculations of Szapudi & Colombi (1996, SC) for the new estimator which includes a local statistical weight, and for a general selection function. Similarly to SC, the hierarchical model and local Poisson behavior was used to simplify the calculations. The local (but not the global) variations of the selection function were neglected as well. This allowed for the first time the accurate estimation of the optimal weight, i.e. the one which minimizes the cosmic error [eqs. (27) to (31)].

To illustrate numerically the first question (i), a virtual SDSS-like catalog was considered. For this, we demonstrated the advantages of our new estimator, which extracts all the relevant information from the catalog at once. This method not only yields higher accuracy in a wider dynamic range than the more traditional volume limited method, but it is significantly more efficient as well. As an added benefit, our calculation of the cosmic error finds the best volume limited strategy if other reasons necessitate its use. A remarkably simple expression for the optimal weight ω was found, which provides an excellent approximation to the solution of the corresponding integral equation

$$\omega(r) \propto 1 \left/ \left[\Delta_{\rm F}^2 + \Delta_{\rm E}^2 + \Delta_{\rm D}^2(r) \right] \right., \tag{76}$$

where $\Delta_{\rm F}^2$, $\Delta_{\rm E}^2$ and $\Delta_{\rm D}^2(r)$ are respectively the finite volume, the edge effect and the discreteness errors. As a result, the optimal weight for an homogeneous sample, i.e. with uniform selection function, is very well approximated by $\omega = 1$. Unlike N-point correlation functions, it is impossible to correct for edge effects with our estimator, due to the finite extension of the smoothing kernel. A different approach, however, is presented elsewhere (Szapudi & Szalay 1997b).

Interestingly, it appears that $\omega = 1$ is a good approximation for the optimal weight in the SDSS catalog, at scales larger than ~ 1 Mpc. As illustrated by figure 4, the corresponding cosmic error on the estimates of $\langle \rho^k \rangle$ should be rather small, of order $1 - 2\%$ for $k = 2$, $3 - 5\%$ for $k = 3$ and $5 - 10\%$ for $k = 4$. These results, however, depend on the details of the model we used for clustering properties of the universe.

Note the similarity of our results with those of Feldman, Kaiser & Peacock (1994, hereafter FKP), who did similar calculations for finding the optimal weight to measure the power-spectrum

29

$P(k) \equiv \langle |\delta_k|^2 \rangle$, where δ_k is the Fourier transform of the density contrast $\delta = \rho - 1$. They assume that the modes are normally distributed in Fourier space and consider modes corresponding to scales small compared to the size of the catalog. Their result is

$$\omega(r) \propto \frac{1}{1/n(r) + P(k)}, \tag{77}$$

in excellent agreement with equation (76). Indeed, from equation (77), the optimal weight presents a plateau at small r and decreases exponentially at large r. The same conclusions apply as well if we compare our optimal weight to the one obtained for measuring optimally the two-point correlation function $\xi(r)$ (e.g., Efstathiou 1996; Hamilton 1997a).

To address question (ii), we considered the sparse sampling strategy proposed by Kaiser (1986, K86, see also Heavens & Taylor 1997). This consist of randomly sampling a fraction f of the visible candidates to build a three dimensional catalog. The optimal sampling rate f minimizes the cosmic error, when the total available telescope time is fixed. A small sampling rate f results in a sparse but deep catalog, therefore decreasing the finite volume and the edge effects. Then discreteness error becomes the limiting factor. The reverse is true for high sampling rate. To measure factorial moments up to order four in a reasonable range of scales such as 1 Mpc $\lesssim \ell \lesssim 100$ Mpc, $f = 1/10 - 1/3$ yields a good compromise between the above effects, depending whether the redshifts are collected individually, or, as usual today, collectively. This is in qualitative agreement with K86, although slightly larger for the following reasons: our calculation of the cosmic error is more realistic; we consider higher order (4 vs. 2) statistics than K86; we emphasize dynamic range from small scales to large scales, the latter being solely considered by K86. Note that the optimal sampling rate increases with the order k of the statistic considered and decreases with scale[6] Therefore, when properties sensitive to higher than fourth order are considered, such as the CPDF shape in the nonlinear regime, cluster selection, higher, most likely full sampling will be optimal (Colombi et al. 1995, Szapudi & Szalay 1996), all the more since we have shown that a sparse sampling strategy yields only a marginal gain compared to full sampling even at the largest scales.

Finally, the design of optimal geometry for a galaxy survey covering some fraction of the sky (e.g. Kaiser 1996, K96) was considered. This is an important problem even for surveys eventually covering a large, continuous portion of the sky, since early results can be obtained by the appropriate (evolving) geometry. Again, the design is governed by the competition between edge and discreteness effects, requiring compactness at large scales, and finite volume effects, requiring large sky coverage. For a reasonably deep survey, where the finite volume error is expected to be quite small, the compact geometry is probably the best choice, because it allows the largest possible scale range for the measurement of the moments. A glass like configuration, where the survey is spread out over the sky in pencils beam subsamples, would reduce the finite volume error. It would be, however, suboptimal for edge effects at large scales, thus constraining the dynamic range of the measurement. A possible compromise is to increase the size of each subsample, for compactness, and still spread them over the sky in a quasi random glass like structure, for reducing the finite volume effects. The details of such a construction, however, depend so much on the goals of the survey in terms of scale range, desired statistics, and predetermined conditions as well, such as the

---

[6]Except for very large scale $\ell \gtrsim$ a few hundred Mpc. In this regime, one expects $\overline{\xi} \propto \ell^{-4}$ [eq. (59)] implying $f \propto \ell$ [eqs. (63), (72)], in agreement with the calculations of Heavens & Taylor (1997) concerning the power-spectrum.

field of view of the telescope, the number of fibers of the multifiber spectrograph, and finally on the statistics of clustering in the universe, that we only attempted to illustrate the problem by solving it for the conceptually simplest cases.

While the calculation presented in this work provides sufficient details for most practical applications, there are several points where generalization or extension could provide more accuracy, if needed:

1. We did not investigate the dependence of the results on the details of the clustering in the universe, since a particular model was assumed to illustrate our method (see § 4.1). As shown by SC, the cosmic error on statistics of order k depends on statistics of order $l \leq 2k$. The dependence becomes stronger at small scales, possibly altering the optimal weight ω. At very large scales, where the non-Gaussianity is less important, the results are only weakly sensitive to higher order statistics. At the level of approximation used in this paper, however, our qualitative results both on optimal weight and sampling rate are expected to be valid for a broad spectrum of realistic statistics. Extended perturbation theory (Colombi et al. 1997) can be used to estimate the sensitivity of the results within reasonable limits. An overall amplification for the optimal sampling rate of a factor 4 is predicted for the variance at nonlinear scales when one passes from an effective spectral index $n_{\text{eff}} = -9$ to $n_{\text{eff}} = -1$ (corresponding respectively to scale-free initial conditions $\langle |\delta_k^2| \rangle \propto k_{\text{linear}}^n$ with $n_{\text{linear}} = -2$ and $n_{\text{linear}} = 0$). The calculations in this paper correspond to an intermediate value $n_{\text{eff}} \sim -3$ (corresponding to $n_{\text{linear}} = -1$), which is most supported by observations.

2. So far, we employed the hierarchical tree model as an approximation to simplify the calculations of the cosmic error. Another possibility is to use tree-level perturbation theory predictions (see, e.g. Bernardeau 1994, 1996), which are seen from N-body simulations to be valid in the regime $\bar{\xi} \lesssim 1$ (e.g., Juszkiewicz, Bouchet & Colombi 1993; Bernardeau 1994; Baugh et al. 1995; Gaztañaga & Baugh 1995; Colombi et al. 1996, 1997). The use of perturbation theory predictions for computing the cosmic error will be presented elsewhere (Szapudi, Bernardeau & Colombi, 1997).

3. As mentioned in the introduction, redshift distortion was completely disregarded, although it affects the measured statistics in two different ways, depending on the scaling regime considered. At large scales, coherent flows enhance the density contrast along the line of sight, therefore increasing the amplitude of the N-point correlation functions. At small scales, on the contrary, the "finger of god" effect by the high velocity dispersion of large clusters tends to smear out clustering and reduce the amplitude of the N-point correlation functions. The net effect on the parameters Q_N is a decrease at small scales and no change at large scales (e.g. Matsubara & Suto 1994; Hivon et al. 1995). One consequence is that, at variance with what is expected from real space measurements in N-body simulations, function $Q_N(\ell)$ is fairly flat in redshift space. As a result, the hierarchical model used along this paper appears to be an even better approximation for estimating the cosmic error in redshift surveys. However, a correction for redshift distortions would be needed to transform the measured moments into configuration space, which is left for future work.

4. Possible dependence of the clustering on luminosity, and morphology was neglected as well. If clustering of galaxies increased with their luminosity (as some observations might indicate), magnitude limited surveys would have inhomogeneous clustering properties. Indeed, only

31

bright objects are seen deep in the catalog, thus clustering would change with depth implying a systematically increasing bias with distance from the observer. In that case, our estimator [eq. (8)] would have to be corrected for this bias, otherwise the volume limited approach is the only alternative. In fact, if the bias is unknown a priori, the volume limited approach provides a way to estimate it. Using an already existing formalism (e.g. Bernardeau & Schaeffer 1992), one can probably directly include the effects of biasing in the calculation of the cosmic error, but this is left for future work. A similar effect is caused by the change of clustering with z in a very deep catalog (Suto 1997).

5. It is worth to emphasize again, that our purely statistical approach cannot account for systematic observational errors, introduced by the instruments, uneven sky, seeing, emission from our own galaxy, inappropriate K-corrections, problems with star galaxy separation, inaccuracy of redshifts, etc. It is only hope, that in a well controlled sample, these effects do not constitute the dominant source of error, or, turning it around, this condition yields well defined accuracy goals for observations.

6. There are known alternatives over constructing an estimator $\tilde{F}_k$ from the CPDF. For instance, for many applications, the connected moments, or cumulants S_N (or, equivalently $Q_N = S_N/N^{N-2}$) are desirable. These numbers are ratios of combinations of factorial moments (for example $S_3 = \langle\delta^3\rangle/\langle\delta^2\rangle^2$). Kim & Strauss (1997) proposed an alternate method to measure S_N in galaxy catalogs, by fitting the Edgeworth expansion (e.g. Juszkiewicz et al. 1995) convolved with a Poissonian to the measured CPDF. They claim that this method, less sensitive to the tails of the CPDF, is more robust than the traditional moment method which was refined in this paper. However, their error estimation is quite ad-hoc, even if normalized with N-body simulations. Moreover, their method is valid only in the weakly nonlinear regime $\bar{\xi} \lesssim 1$, and, in its present form, can be applied only to volume limited catalogs, in contrast with our indicator $\tilde{F}_k$.

7. The CPDF depends on the full hierarchy of factorial moments. In a finite galaxy catalog, however, only a limited amount of information is available, due to the cosmic error and other sources of noise. Thus data compression must be possible without loosing significant information. For example, information content in the tails of the measured CPDF is small because of large fluctuations associated to rare events (e.g., Colombi et al. 1994). Data compression is already used for applying maximum likelihood tests to the galaxy distribution and the cosmic microwave background (e.g., Bond 1995; Vogeley & Szalay 1996; Tegmark, Taylor & Heavens 1996; Jaffe, Knox & Bond 1997). For counts-in-cells statistics, such a compression is complicated by the non-Gaussanity of the likelihood function at small and intermediate scales (e.g. SC).

8. Furthermore, the estimators $\tilde{F}_k$ are not statistically independent (i.e. $\langle\tilde{F}_k\tilde{F}_{k'}\rangle \neq F_k F_{k'}$ if $k \neq k'$, $\langle\tilde{F}_k\tilde{F}_{k'}\tilde{F}_{k''}\rangle \neq F_k F_{k'} F_{k''}$, $k \neq k' \neq k''$, etc.), therefore they do not provide independent tests of theoretical models. An ensemble of statistically independent estimators, but as close as possible to the objects one wants to study to facilitate the physical interpretation of the measurements (in this paper, the moments of the galaxy distribution), would provide much stronger constraints on the models. Such approach is used for measuring the power-spectrum of the galaxy distribution and of the cosmic microwave background (e.g., Hamilton 1997b; Knox, Bond & Jaffe 1997; Tegmark & Hamilton 1997). As before, non-Gaussanity makes the construction of such a hierarchy of estimators difficult.

This paper constitutes the second step in a major investigation on the theoretical errors on counts in cells. SC laid the groundwork for all subsequent calculations, and here, their formalism was extended and applied for magnitude limited redshift surveys with realistic selection function. We proposed a new set of estimators for the factorial moments, which includes compensation for the effects of the selection function and a minimum variance weighting. The integral equation for this weight was solved, an excellent approximation found, as well as the corresponding errors were calculated. Optimal sparse sampling strategies were considered as well, and it was shown that in most cases the decrease in variance does not outweigh the disadvantages of the corresponding information loss. Therefore full sampling is advocated for most applications requiring high order statistics. Finally, the question of optimal survey geometry was addressed, and we found that quasi random distribution of fields is a reasonable choice, when a small fraction of the sky is covered by the survey.

Appendix

A.   Calculation of the Cosmic Error

This section generalizes the calculation of the cosmic error by SC to the case of a non-uniform selection function, and for our estimator which includes local (spatial) statistical weight.

Following CBSII and SC, equation (16) is separated into two parts, according to whether the cells overlap or not

$$E = E_{\text{overlap}} + E_{\text{disjoint}}, \tag{A1}$$

with

$$E_{\text{overlap}} \equiv \frac{1}{\hat{V}} \int_{r_{12} \leq 2\ell} d^3 r_1 d^3 r_2 \ldots, \tag{A2}$$

$$E_{\text{disjoint}} \equiv \frac{1}{\hat{V}} \int_{r_{12} \geq 2\ell} d^3 r_1 d^3 r_2 \ldots, \tag{A3}$$

and $r_{12} \equiv |r_1 - r_2|$. Except when specified, we assume that the cell size is small compared to the survey size. Our calculations will thus be valid at leading order in v/V.

A.1.   Contribution from Disjoint Cells

The contribution to the error from disjoint cells corresponds to *finite volume effects*. It requires the knowledge of the generating function $P_{r_1,r_2}(x,y)$ of the bivariate counts in disjoint cells at positions r_1 and r_2. The problem is that the average number density of galaxies depends here on the distance r from the observer. According to the the reasoning of Szapudi & Szalay (1993a) in terms of Poisson processes [see, e.g., their eq. (5.2)],

$$P_{r_1,r_2}(x,y) = \exp\left\{\sum_{N,M} \frac{(x-1)^N (y-1)^M \overline{N}_{r_1}^N \overline{N}_{r_2}^M}{N! M! v^{N+M}}\right.$$

$$\int_{v_1} d^3u_1 \ldots d^3u_N \int_{v_2} d^3u_{N+1} \ldots d^3u_{N+M} \xi_{N+M}(u_1, \ldots, u_{N+M}) \Bigg\}, \quad \text{(A4)}$$

where $\overline{N}_r$ is the average number of objects per cell expected at position r [eq. (6)]. There are two approximations in SC which are used for the bivariate generating function; We refer to SC for details about the underlying hypotheses in these approximations, which are particular cases of the *hierarchical model* (e.g. Peebles 1980, BS), and for those it is trivial to take into account the above supplementary dependence of the average densities $\overline{N}_{r_1}$ and $\overline{N}_{r_2}$ on distance from the observer. The first approximation, hereafter SS, was derived by Szapudi & Szalay (1993a, 1993b). It assumes that the integral in equation (A4) can be well approximated as $\overline{N}_{r_1}^N \overline{N}_{r_2}^M Q_{N+M,\ N}(r_1),\ M(r_2)NM\xi$ up to linear order in $\xi/\overline{\xi}$. The r dependence (through $\overline{N}$) of quantity $,\ N$ [defined by eq. (46)] is now explicitly written, thus

$$P_{r_1,r_2}(x,y) \simeq P_{r_1}(x)P_{r_2}(y)[1 + R_{r_1,r_2}(x,y)] + \mathcal{O}(\xi^2/\overline{\xi}^2), \quad \text{(A5)}$$

$$R_{r_1,r_2}(x,y) = \xi \sum_{M,\ N=1}^{\infty} (x-1)^N (y-1)^M Q_{N+M,\ M}(r_1),\ _N(r_2)NM, \quad \text{(A6)}$$

where $\xi = \xi(r_{12})$. The second approximation, hereafter BeS, was proposed by Bernardeau & Schaeffer (1992). It is

$$R_{r_1,r_2}(x,y) = \tau\left\{(1-x)\overline{N}(r_1)\overline{\xi}\right\} \tau\left\{(1-y)\overline{N}(r_2)\overline{\xi}\right\} \xi/\overline{\xi}^2, \quad \text{(A7)}$$

where

$$\tau(s) = s\sqrt{2\sum_{N\geq 2}(-s)^{N-2}Q_N \frac{N^{N-2}(N-1)}{N!}}. \quad \text{(A8)}$$

It was noted by SC that the two approximations SS and BeS, although quite different formally, give practically identical results for the cosmic error on the factorial moments in realistic cases. Also, as shown by SC, in the case of a sample without selection effects and with homogeneous weighting ($\phi = \omega = 1$), the relative finite volume error on F_k does not depend on $\overline{N}$, neither on the average density. This means, as shown later, that in the more general case $\phi(r) \leq 1$, all ϕ-dependent terms disappear in the final expression for the finite volume error on F_k.

A.2. Contribution from Overlapping Cells

The overlapping contribution is discussed under the assumption that the variations of $\phi(r)$ and $\omega(r)$ are small across the length 2ℓ. This is a reasonable approximation when the cell size ℓ is small enough. With the additional assumption of local Poisson behavior (SC)

$$E_{\text{overlap}}(x,y) \simeq \frac{1}{\hat{V}} \int_{\hat{V}} d^3r \, \omega^2(r)[\phi(r)]^{-2k} \, E_{\text{overlap}}^{\phi=\omega=1}(x,y)\Big|_{\overline{N}=\overline{N}_r}, \quad \text{(A9)}$$

where the quantity $E_{\text{overlap}}^{\phi=\omega=1}(x,y)$ is the overlapping contribution to the generating function of the cosmic error computed by SC when there are no selection effects and the weighting function is unity. There is an r dependence accounting explicitly for the fact that the average number density $n\phi(r)$ depends on the distance from the observer: $E_{\text{overlap}}^{\phi=\omega=1}(x,y)$ is calculated at $\overline{N} = \overline{N}_r$. The overlapping contribution of the error can be thus inferred from the calculations of SC with a supplementary weighted integral.

34

A.3. Cosmic Error on Factorial Moments

Let us now concentrate on the error on the factorial moments, obtained from equation (14). The finite volume contribution from disjoint cells is

$$\left(\Delta^{\mathrm{F}}\tilde{F}_k\right)^2 = \frac{1}{\hat{V}^2}\int_{r_{12}\geq 2\ell} d^3r_1 d^3r_2 \omega(r_1)\omega(r_2)\eta(r_1,r_2) \tag{A10}$$

with

$$\eta(r_1,r_2) = [\phi(r_1)\phi(r_2)]^{-2k}\left[\frac{\partial}{\partial x}\right]^k\left[\frac{\partial}{\partial y}\right]^k P_{r_1}(x+1)P_{r_2}(y+1)R_{r_1,r_2}(x+1,y+1)\Bigg|_{x=y=0}. \tag{A11}$$

From equations (47) and (A6) or (A7), function η depends only on r_{12}, as follows

$$\eta(r_1,r_2) \propto \xi(r_{12}). \tag{A12}$$

Thus the selection effects *cancel out* in the finite volume contribution to the cosmic error on the factorial moments. The results of SC (see their § 4.3) can therefore be directly used here, just by replacing their $\overline{\xi}(L) \equiv V^{-2}\int_V d^3r_1 d^3r_2 \xi(r_{12})$ by the weighted average

$$\overline{\xi}_{\mathrm{eff}}(\hat{L}) = \frac{1}{\hat{V}^2}\int_{r_1\in\hat{V},r_2\in\hat{V},r_{12}\geq 2\ell} d^3r_1 d^3r_2 \omega(r_1)\omega(r_2)\xi(r_{12}). \tag{A13}$$

The quantity $\hat{L}$ stands for the effective catalogue size $\hat{L} = \hat{V}^{1/3}$. At leading order to v/V the integral (A13) reduces to

$$\overline{\xi}_{\mathrm{eff}}(\hat{L}) \simeq \hat{V}^{-2}\int_{\hat{V}} d^3r_1 d^3r_2 \omega(r_1)\omega(r_2)\xi(r_{12}). \tag{A14}$$

For $\omega = 1$, naturally $\overline{\xi}_{\mathrm{eff}}(\hat{L}) = \overline{\xi}(\hat{L})$. Thus the finite volume error with the notation of § 3.2 is,

$$\Delta_{\mathrm{F}}^2[\omega] \equiv \left(\frac{\Delta^{\mathrm{F}}\tilde{F}_k}{F_k}\right)^2 \simeq \left(\frac{\Delta_{\phi=\omega=1}^{\mathrm{F}}\tilde{F}_k}{F_k}\right)^2\frac{\overline{\xi}_{\mathrm{eff}}(\hat{L})}{\overline{\xi}(L)} = \Delta_{\mathrm{F}}^2\frac{\overline{\xi}_{\mathrm{eff}}(\hat{L})}{\overline{\xi}(L)}. \tag{A15}$$

The calculation of the contribution $(\Delta^{\mathrm{overlap}}\tilde{F}_k)^2$ from overlapping cells, the sum of the edge, and the discreteness errors (see § 3.2), is easier. Using the fact that $\overline{N}_r \propto \phi(r)$, $F_k \propto \overline{N}^k$ and that the relative edge effect error Δ_{E}^2 does not depend on $\overline{N}$ in the case $\phi = \omega = 1$, it is straightforward to obtain equations (29) and (30).

B. Numerical Calculations

B.1. Cosmic Error and Minimum Variance Weight

In this section the numerical solution of integral equation (50) is considered. Let us define the step Δ as

$$\Delta \equiv (\hat{R}_{\mathrm{max}} - \hat{R}_{\mathrm{min}})/N_{\mathrm{bin}}, \tag{B16}$$

and the numbers

$$r_i = \hat{R}_{\min} + (i - 1/2)\Delta, \quad i = 1, \ldots, N_{\text{bin}}. \tag{B17}$$

The solution is obtained in the space of step functions ω verifying

$$\omega(r) = \omega_i, \quad r \in [r_i - \Delta/2, r_i + \Delta/2[. \tag{B18}$$

Equation (50) in its discretized form can be written as

$$\sum_{j=1}^{N_{\text{bin}}} \breve{\xi}_{i,j}\omega_j + \breve{\alpha}_i\omega_i + \breve{\lambda}_i = 0, \tag{B19}$$

where

$$\breve{\xi}_{i,j} \equiv \frac{\Delta_{\text{F}}^2}{\overline{\xi}(L)\hat{V}} \int_{r_i - \Delta/2}^{r_i + \Delta/2} \int_{r_j - \Delta/2}^{r_j + \Delta/2} u^2\hat{\Omega}(u)v^2\hat{\Omega}(v)\tilde{\xi}(u,v)dudv, \tag{B20}$$

$$\breve{\alpha}_i = \int_{r_i - \Delta/2}^{r_i + \Delta/2} \left(\Delta_{\text{E}}^2 + \Delta_{\text{D}}^2(u)\right) u^2\hat{\Omega}(u)du, \tag{B21}$$

$$\breve{\lambda}_i = \lambda \int_{r_i - \Delta/2}^{r_i + \Delta/2} u^2\hat{\Omega}(u)du. \tag{B22}$$

Note that the cosmic error is simply

$$\Delta_{\text{cosmic}}^2[\omega] = \sum_{i,j=1}^{N_{\text{bin}}} \breve{\xi}_{i,j}\omega_i\omega_j + \sum_{i=1}^{N_{\text{bin}}} \breve{\alpha}_i\omega_i^2. \tag{B23}$$

The most difficult part in determining the optimal sampling vector $(\omega_i)_{i=1,\ldots,N_{\text{bin}}}$ is the calculation of the correlation matrix $\breve{\xi}_{i,j}$. A new binning is necessary for computing the double integral (B20). We proceed as follows:

$$\breve{\xi}_{i,j} \simeq \frac{\Delta_{\text{F}}^2}{2\overline{\xi}(L)\hat{V}} \sum_{k,l=1}^{N_{\text{subbin}}} \left[u_k^2\overline{v}_l^2\hat{\Omega}(u_k)\hat{\Omega}(\overline{v}_l)\tilde{\xi}(u_k,\overline{v}_l)\Delta u_k\Delta\overline{v}_l + \overline{u}_k^2 v_l^2\hat{\Omega}(\overline{u}_k)\hat{\Omega}(v_l)\tilde{\xi}(\overline{u}_k,v_l)\Delta\overline{u}_k\Delta v_l\right], \tag{B24}$$

where

$$\Delta u_k = \Delta v_k = \begin{cases} \Delta/(N_{\text{subbin}} - 1), & 2 \leq k \leq N_{\text{subbin}} - 1 \\ \frac{1}{2}\Delta/(N_{\text{subbin}} - 1), & k = 1 \text{ or } N_{\text{subbin}}, \end{cases} \tag{B25}$$

$$\Delta\overline{u}_k = \Delta\overline{v}_k = \Delta/N_{\text{subbin}}, \tag{B26}$$

$$u_k = r_i + (k - 1)\Delta/(N_{\text{subbin}} - 1), \quad v_l = r_j + (l - 1)\Delta/(N_{\text{subbin}} - 1), \tag{B27}$$

$$\overline{u}_k = r_i + (k - 1/2)\Delta\overline{u}_k, \quad \overline{v}_l = r_j + (l - 1/2)\Delta\overline{v}_l. \tag{B28}$$

The double summation in equation (B24) avoids calculation of function $\tilde{\xi}(u,v)$ for $u = v$, where it might diverge (for example, if $\xi(r) \propto r^{-\gamma}$ at small scales, then function $\tilde{\xi}(u,u)$ diverges when $\gamma \geq 2$). It remains to compute function $\tilde{\xi}(u,v)$, $u \neq v$. In equation (49), one can define θ as the angle between the directions $\hat{r}_1$ and $\hat{r}_2$:

$$\mu \equiv \cos(\theta) = \cos(\varphi_1 - \varphi_2)\sin\theta_1\sin\theta_2 + \cos\theta_1\cos\theta_2. \tag{B29}$$

Equation (49) can then be rewritten as

$$\tilde{\xi}(u,v) = \int_{-1}^{1} P(\mu|u,v)d\mu\,\xi\left([u^2 + v^2 - 2uv\mu]^{1/2}\right),$$ (B30)

where $P(\mu|u,v)$ is the probability distribution function of μ, given u and v. (For a spherical catalog $P(\mu|u,v) = 1/2$). Typically, if u and v are large enough compared to $\hat{R}_{\mathrm{min}}$, the function $P(\mu|u,v)$ does not depend sensitively on the values of u and v. It increases with μ, from zero for $-1 \leq \mu \leq \mu_{\mathrm{min}}$, where μ_{min} corresponds to the largest effective angular size of the catalog, to some maximum at $\mu = 1$, P_{max}, depending on the angular size of the catalog. For our SDSS-like catalog, P_{max} is of order 2. The calculation of $P(\mu|u,v)$ is done numerically for a discrete set of values of (μ_i, r_j, r_k), $1 \leq i \leq N_\mu$, $0 \leq j,k \leq N_{\mathrm{bin}} + 1$ and $r_0 \equiv \hat{R}_{\mathrm{min}}$, $r_{N_{\mathrm{bin}}+1} \equiv \hat{R}_{\mathrm{max}}$. For each value of (j,k), a Monte-Carlo simulation is done, i.e., directions $\hat{r}_j$ and $\hat{r}_k$ are randomly chosen such that the corresponding cells are included in the catalog. A reasonably accurate calculation, with a few percent absolute errors requires $N_{\mathrm{iter}} \sim 100,000$ iterations for $N_\mu = 100$. For a given value of (μ, u, v), the estimate $P_{\mathrm{interpol}}(\mu, u, v)$ of $P(\mu, u, v)$ is obtained from a bilinear interpolation between $P(\mu_i, r_j, r_k)$, $P(\mu_i, r_{j+1}, r_k)$, $P(\mu_i, r_j, r_{k+1})$, $P(\mu_i, r_{j+1}, r_{k+1})$ where μ_i is as close as possible to μ, and $r_j \leq u \leq r_{j+1}$, $r_k \leq v \leq r_{k+1}$. By definition, we take $P(\mu_i, r_j, r_k) \equiv 0$ if $j = 0$ or $k = 0$.

We use the same Monte-Carlo simulation as above (except that we use only one direction $\hat{r}_j$) to compute the effective solid angle on an array $\hat{\Omega}(r_j)$, and then proceed with linear interpolations to compute estimates $\hat{\Omega}_{\mathrm{interpol}}(u)$ of $\hat{\Omega}(u)$ at $u \neq r_j$.

To compute the angular average (B30) the following variable is defined

$$z(\mu) = \frac{1}{(\gamma_z - 2)uv}\left(u^2 + v^2 - 2uv\mu\right)^{1-\gamma_z/2},$$ (B31)

where $\gamma_z \simeq \gamma$ is the expected power-law index of the two-point correlation function at small scale, or some value close to it. With the definitions $\mu(z) \equiv z^{-1}(z(\mu))$, $z_{\mathrm{m}} = z(\mu = -1)$, $z_{\mathrm{M}} = z(\mu = 1)$,

$$\Delta z_i = \begin{cases} (z_{\mathrm{M}} - z_{\mathrm{m}})/(N_z - 1), & 2 \leq i \leq N_z - 1, \\ \frac{1}{2}(z_{\mathrm{M}} - z_{\mathrm{m}})/(N_z - 1), & i = 1 \text{ or } N_z, \end{cases}$$ (B32)

$$z_i = z_m + (i-1)(z_{\mathrm{M}} - z_{\mathrm{m}})/(N_z - 1),$$ (B33)

the correlation matrix can be written as

$$\tilde{\xi}(u,v) \simeq \sum_{i=1}^{N_z} [uv(\gamma_z - 2)z_i]^{\gamma_z/(2-\gamma_z)}\, P_{\mathrm{interpol}}(\mu(z_i), u, v)\xi\left([uv(\gamma_z - 2)z_i]^{1/(2-\gamma_z)}\right)\Delta z_i.$$ (B34)

The calculation of $\breve{\alpha}_i$ follows naturally from

$$\breve{\alpha}_i = \sum_{k=1, N_{\mathrm{subbin}}} \left(\Delta_{\mathrm{edge}}^2 + \Delta_{\mathrm{D}}^2(u_k)\right) u_k^2 \hat{\Omega}_{\mathrm{interpol}}(u_k)\Delta u_k,$$ (B35)

and one proceed similarly for $\breve{\lambda}_i$:

$$\breve{\lambda}_i = \lambda \sum_{k=1, N_{\mathrm{subbin}}} u_k^2 \hat{\Omega}_{\mathrm{interpol}}(u_k)\Delta u_k.$$ (B36)

The numerical estimate of the effective sample volume reads

$$\hat{V} \simeq \sum_{i=1}^{N_{\rm bin}} \frac{1}{r_{i+1} - r_i} \left[\left(\hat{\Omega}(r_{i+1}) - \hat{\Omega}(r_i) \right) \left(r_{i+1}^4 - r_i^4 \right) / 4 + \left(\hat{\Omega}(r_i) r_{i+1} - \hat{\Omega}(r_{i+1}) r_i \right) \left(r_{i+1}^3 - r_i^3 \right) / 3 \right] . \tag{B37}$$

The same integration scheme is used to normalize the weight [eq. (9)]. To compute the sample volume V of our SDSS like catalog, a more accurate estimator is used, with exact calculation of $\Omega(r)$ and an integral on variable r using the trapezoidal method with $100,000$ points.

For the calculation of the terms $\Delta_{\rm F}^2$, $\Delta_{\rm E}^2$ and $\Delta_{\rm D}^2$, we use the results of SC corresponding to the SS approximation and to $\gamma = 1.8$ where γ is the assumed logarithmic slope of function $\xi(r)$ for computing the numerical coefficients in the different terms contributing to the cosmic error [see equations (53) to (68) in SC]. Note that, as discussed in SC, these numerical coefficients are quite insensitive to γ if it stays reasonably close to 1.8. Therefore the same expression is applied for the cosmic error as a function of Q_N, $\overline{N}$ and $\overline{\xi}$ even if $\gamma \neq 1.8$ or if function $\overline{\xi}(\ell)$ is not exactly a power-law, like in the CDM case.

B.2.   Accuracy Tests

The difficulty in obtaining accurate results for the optimal weight relies mostly in the estimation of the correlation matrix $\breve{\xi}_{i,j}$. Here, we assume that the finite volume error dominates over the edge effect and the shot noise errors, although we performed extensive accuracy tests including all sources of error.

In what follows, we study the optimal weight $\omega(r)$, which depends only on the shape of the two-point correlation function (and thus not on higher order statistics). We consider a spherical sample, where there is no need in principle to compute numerically $P(\mu|u,v)$ as straightforwardly $P(\mu|u,v) = 1/2$. Furthermore, the two-point correlation function is assumed to be a power law of index $-\gamma$ [eq. (33)], which allows the analytical computation of the correlation matrix $\breve{\xi}_{i,j}$:

$$\begin{aligned} \breve{\xi}_{i,j} &= \frac{8\pi^2 \Delta_{\rm F}^2}{\overline{\xi}(L)\hat{V}} \left[F(r_i + \Delta/2, r_j + \Delta/2) + F(r_i - \Delta/2, r_j - \Delta/2) \right. \\ &\left. - F(r_i + \Delta/2, r_j - \Delta/2) - F(r_i - \Delta/2, r_j + \Delta/2) \right] \end{aligned} \tag{B38}$$

with

$$\begin{aligned} F(u,v) &\equiv \frac{r_0^\gamma}{(2-\gamma)(3-\gamma)(4-\gamma)} \left[uv \left\{ |u-v|^{4-\gamma} + (u+v)^{4-\gamma} \right\} \right. \\ &\left. + \left\{ |u-v|^{6-\gamma} - (u+v)^{6-\gamma} \right\} / (6-\gamma) \right], \end{aligned} \tag{B39}$$

and

$$\hat{V} = \frac{4}{3} \pi \hat{R}_{\rm max}^3. \tag{B40}$$

Therefore

$$\breve{\lambda}_i = 4\pi r_i^2 \Delta \left(1 + \frac{\Delta^2}{12 r_i^2} \right). \tag{B41}$$

In figure 8, the minimum variance weight is shown as a function of distance from the observer, obtained from three different calculations, each with $N_{\rm bin} = 50$. The solid curve corresponds to
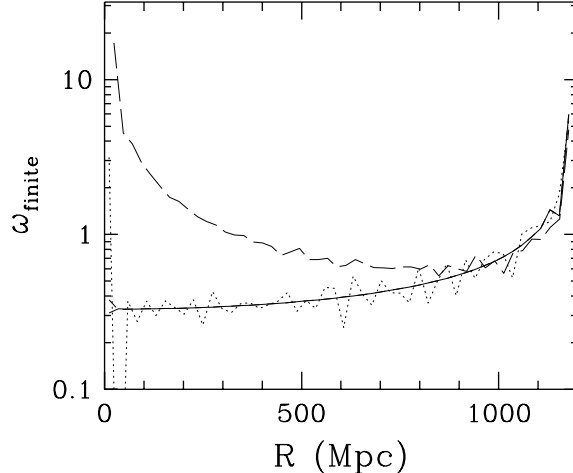
38

Fig. 8.— The minimum variance weight ω is shown as a function of distance R from the observer in the case the finite volume error is dominant (the results displayed here correspond to $\ell = 10$ Mpc, but would not change significantly for other values of ℓ). For the finite volume error, ω does not depend on the order k. The two point function is assumed to be a power-law $\xi(r) \propto r^{-\gamma}$. The calculations were performed assuming spherical geometry, except for the long dashes which show the result for a SDSS-like geometry. The solid curve corresponds to the full analytical calculation of the correlation matrix $\breve{\xi}_{i,j}$. A dashed curve is almost perfectly superposing to the solid one. In that case, the numerical scheme given by equations (B24), (B36) and (B34) (with $P_{\text{interpol}}(\mu|u,v) \equiv 0.5$) has been used. The dotted line shows the result as calculated via a Monte-Carlo simulation plus bilinear interpolation to compute $P_{\text{interpol}}(\mu|u,v)$. The long dashes, corresponding to a SDSS geometry, were obtained the same way as the dots.

the analytical result [eqs. (B38) and (B41)]. There is a dashed curve almost perfectly superposing to the solid one, except from the extreme left part of it. In that case, equations (B24), (B36) and (B34) with $P(\mu|u,v)_{\text{interpol}} = 0.5$ were used to compute $\breve{\xi}_{i,j}$ and $\breve{\lambda}_i$. We have taken $N_{\text{subbin}} = 30$ and $N_z = 50$ (as for all the figures of the main text, when it is relevant). For both the solid and the dashed curves, there is an irregularity at each end. The left irregularity, corresponding to the weight given (nearly) at the center of the catalog, is due to the limitations of the discrete approach. In the continuous limit, the function $\omega(r)$ is expected to behave smoothly in the neighborhood of $r = 0$. The right irregularity is more difficult to understand, and will be discussed later. The fluctuating dotted curve is the same as the dashed one, but we computed the probability distribution function $P(\mu|u,v)$ by Monte-Carlo simulation as explained in § B.1 (with $N_{\text{iter}} = 100,000$ iterations for $N_\mu = 100$). The left irregularity for this is quite dramatic, as there is one value of the estimated weight which is negative. Note, however, that the calculation of the finite volume error is fairly robust with respect to the weight: the weights given by the solid, the (invisible) short-dashed or the dotted curve give exactly the same result. The long-dashed curve is the same as the dotted one, expect that it corresponds to an SDSS like geometry covering thus one quarter of the sky. The irregularities are less pronounced in that case, due to the anisotropy of the distribution $P(\mu|u,v)$ of values of μ (see also the solid curves of Fig. 1).

Figure 9 shows the effect of changing N_{bin} when calculating the optimal weight. There are six curves, corresponding respectively to $N_{\text{bin}} = 6, 12, 25, 50, 100$ and 200. The method used for the
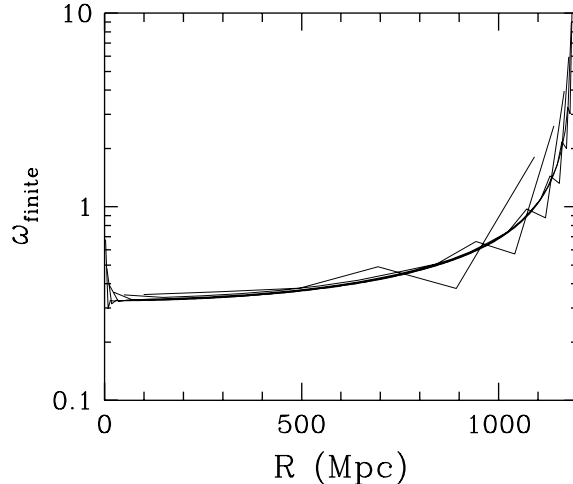
Fig. 9.— Same as figure 8 except that the sensitivity of the results to the value of $N_{\rm bin}$ are tested for $N_{\rm bin} = 6, 12, 25, 50, 100$ and 200. The numerical scheme given by equations (B24), (B36) and (B34) (with $P_{\rm interpol}(\mu|u,v) \equiv 0.5$) was used.

calculation is the same as the (almost invisible) dashed line in figure 8. Except for the right and the left end-points of the curves, resolution is not critical for determining $\omega(r)$. The left irregularity on each curve was shown to be unphysical and due to the numerical limitations of our calculation. The right one is more of a concern, since the estimator of $\log_{10}[\omega(\hat{R}_{\rm max})]$ increases linearly with $N_{\rm bin}$, suggesting that the actual optimal weight is singular at $\hat{R}_{\rm max}$. As a consequence, $N_{\rm bin}$ has to be large enough to resolve this singularity sufficiently. The cosmic error can be sensitive to the singularity, especially if the order k is large and if the finite volume error is significant. In realistic cases, however, the two-point correlation does not behave like a power-law of index $-\gamma = -1.8$ up to arbitrarily large scales, thus the singularity is expected to be less pronounced or to disappear. For example, in the CDM case, the effect is less significant, although there is still a slight instability at $\hat{R}_{\rm max}$ (see Fig. 1).

B.3. The optimal sampling rate

The methods for generating figures 6 and 7 will be explained in this section. To simplify the calculations and since it was found in § 4.2 to be a good approximation, we take for the optimal weight ω the expression given by equation (51). To estimate the cosmic error from equation (B23), the matrix $\breve{\xi}_{i,j}$, and the vector $\breve{\alpha}_i$ have to be computed for rather large values of $R_{\rm max}$ in order to scan a large enough range of the sampling rate f [see eq. (55)]. Solving the implicit equation for minimizing the cosmic error as a function of the sampling rate by recalculating the matrix $\breve{\xi}_{i,j}$ for each value of f would be too costly from the computational point of view. Instead, we calculated $\breve{\xi}_{i,j}$ for $R_{\rm max} \equiv 5000$ Mpc and with $N_{\rm bin} = 200$. A given value of f corresponds to some $\hat{R}_{\rm max}(f)$ and thus roughly to some value of r_i. This is taken to be the as close as possible to $\hat{R}_{\rm max}(f)$. To compute the cosmic error corresponding to some value of f, quadratic interpolation is performed between the values obtained for $\hat{R}_{\rm max}(f) = r_{i-1}$, $\hat{R}_{\rm max}(f) = r_i$ and $\hat{R}_{\rm max}(f) = r_{i+1}$. Even this smoothing procedure cannot guarantee smoothness for f as a function of scale, although

40

the calculation is valid within a few percents. The implicit equation for finding the minimum of function $\Delta^2_{\mathrm{cosmic}}(f)$ is easy to solve by bisection as this function is convex (see figure 7).

C.   Asymptotic Regimes

Using the results of SC (valid at leading order in v/V), this section presents an analytic estimate of the optimal sampling rate f for a three-dimensional galaxy catalog (§ 5.1 and 5.2). This is possible in the highly and weakly nonlinear regimes, where $\overline{\xi} \gg 1$ and $\overline{\xi} \ll 1$, respectively. According to the results of § 5.1, uniform selection $\phi = 1$ can be assumed.

From the calculations of SC, the cosmic error on the factorial moments of order $k \leq 3$ in the highly nonlinear regime is (using the SS approximation, see § A.1)

$$\left(\frac{\Delta^2_{\mathrm{cosmic}}\tilde{F}_1}{F_1}\right)^2 \simeq \overline{\xi}(L) + \overline{\xi}\frac{v}{V}\left[\frac{1}{N_{\mathrm{c}}} + 5.51\right], \tag{C42}$$

$$\left(\frac{\Delta^2_{\mathrm{cosmic}}\tilde{F}_2}{F_2}\right)^2 \simeq 4Q_4\overline{\xi}(L) + \overline{\xi}\frac{v}{V}\left[\frac{0.5}{N_{\mathrm{c}}^2} + \frac{6.6Q_3}{N_{\mathrm{c}}} + 42.2Q_4\right], \tag{C43}$$

$$\left(\frac{\Delta^2_{\mathrm{cosmic}}\tilde{F}_3}{F_3}\right)^2 \simeq \frac{9Q_6}{Q_3^2}\overline{\xi}(L) + \overline{\xi}\frac{v}{V}\left[\frac{0.19}{Q_3}\frac{1}{N_{\mathrm{c}}^3} + \frac{4.93Q_4}{Q_3^2}\frac{1}{N_{\mathrm{c}}^2} + \frac{38.8}{Q_3^2}\frac{1}{N_{\mathrm{c}}} + \frac{185Q_6}{Q_3^2}\right], \tag{C44}$$

where

$$N_{\mathrm{c}} \equiv \overline{N}\,\overline{\xi} \tag{C45}$$

is the typical number of galaxies in a cell located in an overdense region (e.g., BS). At small enough scales and if $\gamma < 3$, it is possible that $N_{\mathrm{c}} \ll 1$ irrespectively of f. In that case, count-in-cells statistics is dominated by discreteness effects (BS). The edge effect error, corresponding to the term proportional to $\overline{\xi}v/V$ and independent of N_{c}, is negligible compared to the shot noise error. Equations (61) or (70) follow straightforwardly.

In the weakly nonlinear regime, where $\overline{\xi} \ll 1$, the Gaussian limit at leading order in $\overline{\xi}$ is a good approximation. The expression (C42) for the error on F_1 is still valid. At higher order, however,

$$\left(\frac{\Delta^2_{\mathrm{cosmic}}\tilde{F}_2}{F_2}\right)^2 \simeq 4\overline{\xi}(L) + 17.05\overline{\xi}\frac{v}{V} + (0.65 + 4\overline{N})\frac{1}{\overline{N}^2}\frac{v}{V}, \tag{C46}$$

$$\left(\frac{\Delta^2_{\mathrm{cosmic}}\tilde{F}_3}{F_3}\right)^2 \simeq 9\overline{\xi}(L) + 34.6\overline{\xi}\frac{v}{V} + \left(0.88 + 5.83\overline{N} + 9\overline{N}^2\right)\frac{1}{\overline{N}^3}\frac{v}{V}. \tag{C47}$$

If the scale is large enough, edge effects are expected to dominate over the finite volume errors, but this property depends on the details of the underlying statistics and on the geometry of the catalog. More rigorous estimates of the cosmic error, taking fully into account the geometry of the catalog would be needed to prove it, a level of accuracy outside of the scope of this paper. Exact calculations of the cosmic error when the size of the cell becomes comparable to the size of the catalog are indeed quite tedious (see for example Appendix B of SC). Although our approach is valid only to leading order in v/V, setting $v/V = 1$ in equations (C42), (C46), and (C47) yields at least a lower limit for the relative contribution of edge effects compared to finite volume effects.

The result is that the edge error is a few times larger than the finite volume error (for a spherical catalog). Assuming that edge effects dominate, equations (63) and (72) naturally follow.

Note that if finite volume effects dominate edge effects in the regime $\overline{\xi} \ll 1$, and $\overline{N} \gg 1$, the optimal sampling rate, with the notations of § 5, becomes

$$f = \left[\frac{1 - 3\zeta}{\gamma_L \zeta N_{\text{ref}}} \left(\frac{L_{\text{ref}}}{L_0} \right)^{\gamma_L} \right]^{\frac{1}{1+(\gamma_L - 3)\zeta}} , \qquad (C48)$$

where $\zeta = 1/5$ or $1/9$ according to whether the redshifts are collected individually (§ 5.1) or collectively (§ 5.2). For the same choice of the parameters as in § 5, we find $f \sim 1/20$ for individual collection of redshifts and $f \sim 1/10$ for simultaneous collection independently of scale and of the order k. As a result, the general conclusions of our analysis in § 5.1 and 5.2 are not changed, even if finite volume effects dominate over edge effects.

REFERENCES

Alimi J.-M., Blanchard A., Schaeffer R., 1990, ApJ, 349, L5

Balian R., Schaeffer R., 1989, A&A, 220, 1 (BS)

Baugh C.M., Gaztañaga E., Efstathiou G., 1995, MNRAS, 274, 1049

Benoist C., Maurogordato S., da Costa L.N., Cappi A., Schaeffer R., 1996, ApJ, in press (astro-ph/9605117)

Bernardeau F., 1994, A&A, 291, 697

Bernardeau F., 1996, A&A, 312, 11

Bernardeau F., Schaeffer R., 1992, A&A, 255, 1 (BeS)

Bond J.R., 1995, Phys. Rev. Lett., 74, 4369

Bouchet F.R., Schaeffer R., Davis M., 1991, ApJ, 383, 19

Bouchet F.R., Hernquist L., 1992, ApJ, 400, 25

Bouchet F.R., Strauss M.A., Davis M., Fisher K.B., Yahil A., Huchra J.P., 1993, ApJ, 417, 36

Bromley B.C., 1994, ApJ, 437, 541

Bunn E.F., White M. 1997, ApJ, 480, 6

Colombi S., Bernardeau F., Bouchet F.R., Hernquist L., 1997, MNRAS, 287, 241

Colombi S., Bouchet F.R., Hernquist L., 1996, ApJ, 465, 14

Colombi S., Bouchet F.R., Schaeffer R., 1994, A&A, 281, 301 (CBSI)

Colombi S., Bouchet F.R., Schaeffer R., 1995, ApJS, 96, 401 (CBSII)

Davis M., Meiksin A., Strauss M.A., da Costa N.L., Yahil A., 1988, ApJ, 333, L9

Davis M., Peebles P.J.E., 1983, ApJ, 267, 465

Dominguez-Tenreiro R., Martinez V.J., 1989, ApJ, 339, L9

Efstathiou G., 1996, in Les Houches Lectures, Session LX, Cosmology and Large Scale Structure (Elsevier Science, Amsterdam), eds. R. Schaeffer, J. Silk, M. Spiro & J. Zinn-Justin, p. 133

Efstathiou G., Bond J.R., White S.D.M., MNRAS, 258, 1p

Efstathiou G., Ellis R.S., Peterson B.A., 1988, MNRAS, 232, 431

Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R.S., Frenk C. S., 1990, MNRAS, 247, 10P

Efstathiou G., Frenk C.S., White S.D.M., Davis M., 1988, MNRAS, 235, 715

Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23

Fisher K.B., Davis M., Strauss M.A., Yahil A., Huchra J., 1994, MNRAS, 266, 50

Fry J.N., Peebles P.J.E., 1978, ApJ, 221, 19

Fry J.N., Melott A.L., Shandarin S.F., 1993, ApJ, 412, 504

Gaztañaga E., 1992, ApJ, 398, L17

Gaztañaga E., 1994, MNRAS, 268, 913

Gaztañaga E., Baugh C.M., 1995, MNRAS, 273, L1

Groth E.J., Peebles P.J.E., 1977, ApJ, 217, 385

Hamilton A.J.S., 1988, ApJ, 331, L59

Hamilton A.J.S., 1997a, MNRAS, in press (astro-ph/9701008)

Hamilton A.J.S., 1997b, MNRAS, in press (astro-ph/9701009)

Heavens A.F., Taylor A.N., 1997, MNRAS, in press (astro-ph/9705215)

Hivon E., Bouchet F.R., Colombi S., Juszkiewicz R., 1995, A&A, 298, 643

Jaffe A.H., Knox L., Bond J.R., 1997, to appear in the Proceedings of the 18th Texas Symposium
    on Relativistic Astrophysics (astro-ph/9702109)

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

Kaiser N., 1986, MNRAS, 219, 785 (K86)

Kaiser N., 1987, MNRAS, 227, 1

Kaiser N., 1996, submitted to ApJ (astro-ph/9610120) (K96)

Kim R.S., Strauss M.A., 1997, submitted to ApJ (astro-ph/9702144)

Knox L., Bond J.R., Jaffe A.H., 1997, to appear in the Proceedings of the 18th Texas Symposium
    on Relativistic Astrophysics (astro-ph/9702110)

Lahav O., 1996, in Mapping, Measuring and Modelling the Universe (Valencia, Spain), eds. P.
    Coles & V. Martinez (astro-ph/9601054)

Lahav O., Itoh M., Inagaki S., Suto Y., 1993, ApJ, 402, 387

Landy S.D., Szalay A., 1993, ApJ, 412, 64

de Lapparent V., Geller M.J., Huchra J.P., 1989, ApJ, 343, 1

Loveday J., 1996, in Dark Matter in Cosmology, Quantum Measurements, Experimental Gravita-
    tion, Proceedings of the XXXIst Rencontres de Moriond (Editions Frontières, Gif-sur-Yvette,
    France), p. 215

Loveday J., Peterson B.A., Efstathiou G., Maddox S.J., 1992, ApJ, 390, 338

Lucchin F., Materese S., Melott A.L., Moscardini L., 1994, ApJ, 422, 430

Matsubara T., Suto Y., 1994, ApJ, 420, 497

Maurogordato S., Schaeffer R., da Costa L.N., 1992, ApJ, 390, 17

Meiksin A., Szapudi I., Szalay A., 1992, ApJ, 394, 87

Mo H.J., White S.D.M., 1996, MNRAS, 282, 347

Moore B., Frenk C.S., Efstathiou G., Saunders W., 1994, MNRAS, 269, 742

Peacock J.A., Dodds S.J., 1994, MNRAS, 267, 1020

Peebles P.J.E., 1974, A&A, 32, 197

Peebles P.J.E., 1980, The Large Scale Structure of the Universe (Princeton University Press, Prince-
    ton, N.Y., U.S.A.)

Ratcliffe A., Shanks T., Parker Q.A., Fong R., 1997, preprint (astro-ph/9702227)

Ripley B.D., 1988, Statistical Inference for Spatial Processes (Cambridge: Cambridge Univ. Press)

Schechter P., 1976, ApJ, 203, 297

Sharp N.A., Bonometto S.A., Lucchin F., 1984, A&A, 130, 79

Szapudi I., 1997, submitted to ApJ

Szapudi I., Bernardeau F., Colombi S., 1997, in preparation

Szapudi I., Colombi S., 1996, ApJ, 470, 131 (SC)

Szapudi I., Dalton G., Efstathiou G.P., Szalay A., 1995, ApJ, 444, 520

Szapudi I., Meiksin A., Nichol R., 1996, ApJ, 473, 15

Szapudi I., Szalay A., 1993a, ApJ, 408, 43

Szapudi I., Szalay A., 1993b, ApJ, 414, 493

Szapudi I., Szalay A., 1996, ApJ, 459, 504

Szapudi I., Szalay A., 1997a, ApJ, 481, L1

Szapudi I., Szalay A., 1997b, preprint FERMILAB-Pub-97/108-A (astro-ph/9704241), submitted
    to ApJLett

Szapudi I., Szalay A., Boschán P., 1992, ApJ, 390, 350

Suto Y., 1997, private communication

Tegmark M., Hamilton A.J.S., 1997, to appear in the Proceedings of the 18th Texas Symposium
    on Relativistic Astrophysics (astro-ph/9702019)

Tegmark M., Taylor A.N., Heavens A.F., 1996, preprint (astro-ph/9603021)

Totsuji H., Kihara T., 1969, PASJ, 21, 221

Tucker D.L., Oemler A. Jr., Kirshner R.P., Lin H., Shectman S.A., Landy S.D., Schechter P.L.,
    Müller V., Gottlöber S., Einasto J., 1996, preprint

Valls-Gabaud D., Alimi J.M., Blanchard A., 1989, Nature, 341, 215

Vogeley, M.S., Szalay A.S., 1996, ApJ, 465, 34

White S.D.M, Davis M., Efstathiou G., Frenk C.S., 1987, Nature, 330, 451