ON RECOVERING THE NONLINEAR BIAS FUNCTION FROM COUNTS-IN-CELLS MEASUREMENTS

ISTVÁN SZAPUDI AND JUN PAN
Institute for Astronomy, University of Hawaii, Honolulu, HI 96822
Received 2003 August 28; accepted 2003 October 20

ABSTRACT

We present a simple and accurate method to constrain galaxy bias based on the distribution of counts in cells. The unique feature of our technique is that it is applicable to nonlinear scales, where both dark matter statistics and the nature of galaxy bias are fairly complex. First, we estimate the underlying continuous distribution function from precise counts-in-cells measurements, assuming local Poisson sampling. Then a robust, non-parametric inversion of the bias function is recovered from the comparison of the cumulative distributions in simulated dark matter and galaxy catalogs. Obtaining continuous statistics from the discrete counts is the most delicate and novel part of our recipe, which corresponds to a deconvolution of a (Poisson) kernel. For this we present two alternatives: a model-independent algorithm based on Richardson-Lucy iteration, and a solution using a parametric skewed lognormal model. We find that the latter is an excellent approximation for the dark matter distribution, but the model-independent iterative procedure is more suitable for galaxies. Tests based on high-resolution dark matter simulations and corresponding mock galaxy catalogs show that we can reconstruct the nonlinear bias function down to highly nonlinear scales with high precision in the range of \(-1 \leq \delta \leq 5\). As far as the stochasticity of the bias, we have found a remarkably simple and accurate formula based on Poisson noise, which provides an excellent approximation for the scatter around the mean nonlinear bias function. In addition, we have found that redshift distortions have a negligible effect on our bias reconstruction; therefore, our recipe can be safely applied to redshift surveys.

Subject headings: cosmology: observations — dark matter — galaxies: statistics — large-scale structure of universe — methods: numerical

1. INTRODUCTION

The principal aim of statistical analysis of galaxy catalogs is to extract information about the initial fluctuations in the early universe, their subsequent gravitational growth, and processes of galaxy formation. To decipher the available data, the distribution of the underlying dark matter has to be inferred from the distribution of galaxies. The two distributions in principle can be assumed to be quite different, i.e., galaxies are biased tracers of the underlying dark matter statistics (Kaiser 1984; Bardeen et al. 1986; Lahav & Saslaw 1992; Dekel & Lahav 1999). In a class of phenomenological models, galaxy fluctuations \(\delta_g\) are assumed to be a function of the matter fluctuation field, \(f(\delta_m)\), which includes the simplest case of linear bias \(\delta_g = b \delta_m\) (Fry & Gaztañaga 1993; Szapudi 1995; Matsubara 1995, 1999). Accurate knowledge of this function is needed to interpret galaxy statistics in light of theories of structure formation.

The bias function can be extracted from measurements of large-scale structure statistics and the cosmic microwave background (CMB). One possibility is to use second-order statistics from several sources, especially CMB maps, together with galaxy catalogs to constrain bias (e.g., Efstathiou et al. 2002; Lahav et al. 2002; Verde et al. 2003). The underlying idea is that fluctuations in the CMB have a direct relationship with density fluctuations in the early universe. The comparison with present-day galaxy statistics reveals information on both gravitational amplification and galaxy formation. The application of these methods is limited to fairly large scales, since nonlinear growth and bias for galaxies, as well as secondary anisotropies for the CMB, become more and more complex on smaller scales.

Another class of methods uses higher order statistics (Fry & Gaztañaga 1993; Gaztañaga & Frieman 1994; Fry 1994; Szapudi 1998a; Feldman et al. 2001; Verde et al. 2002) or velocity information (e.g., Branchini et al. 2000; Zaroubi et al. 2002) from galaxy catalogs to derive the bias function internally. These methods again are limited to linear to weakly nonlinear scales, where both gravitational instability and bias theory are well understood. Since the most reliable large-scale structure data are available on smaller scales, it is natural to seek methods for constraining bias that are applicable on small, nonlinear scales. The principal aim of this work is to introduce such a technique based on a direct comparison of counts in cells in simulations and data.

We generalize a simple and elegant idea by Sigad, Branchini, & Dekel (2000), based on the relation between the (continuous) cumulative probability distribution functions of the galaxy and matter density fluctuation fields, \(C_g(\delta_g)\) and \(C_m(\delta_m)\):

\[
C_m(\delta_m) = C_g[f(\delta_m)],
\]

where \(C(\delta) = \int_{-\delta}^{\delta} p(\delta') d\delta'\). The above relation allows in principle the recovery of the bias function if both cumulative distributions are known:

\[
\delta_g = f(\delta_m) = C_g^{-1}[C_m(\delta_m)].
\]

Sigad et al. (2000) used an approximate form of this relation by simply replacing the cumulative probability distribution with the cumulative distribution of counts in cells, \(C_{N,N_{\text{max}}}(N) = \sum_{N < N_{\text{max}}} P_N\), and postulating that the dark matter cumulative distribution is well described by a lognormal function. Both of these approximations render the original form of this method fairly approximate. The dark matter distribution does not exactly follow a lognormal distribution (as we show later),
and discreteness effects (the difference between the continuous and discrete distribution) are important only for the densest galaxy catalogs. We improve the original idea on the local Poissonian assumption appears to be correct, and we assume it for the rest of this work.

As long as equation (3) can be inverted, equation (2) can be used to calculate the bias. The inversion, however, is a fairly delicate process. Most of the technical challenges of our aim were met by devising a pair of new methods for deconvolving this somewhat unstable equation in a robust way.

2. METHODS

The technical challenge of the idea outlined in § 1 consists of estimating the continuous probability distribution function from counts-in-cells measurements (CIC) in a robust way. Note that there are several proven and fast methods to estimate CIC distributions from galaxy catalogs or in simulations (Szapudi 1998b; Szapudi et al. 1999; S. Colombi & I. Szapudi 2004, in preparation). In this work, we use the former method for all measurements.

The count probability distribution function (CPDF) $P_N$ is the probability that a certain cell has $N$ objects (galaxies). It is directly related to the continuous function under the locally Poissonian approximation

$$ P_N = \int_{-\infty}^{1} p(\delta) \frac{[\langle N \rangle (1 + \delta)]^N e^{-(\langle N \rangle + \delta)}}{N!} d\delta, \quad (3) $$

where $\langle N \rangle$ is the mean CIC. Some models of galaxy formation predict sub-Poissonian scatter for very small halos (e.g., Somerville et al. 2001; Casas-Miranda et al. 2002; Berlind & Weinberg 2002). As long as a specific theoretical model written in a convolution form similar to the above is available, our technique can be easily adapted. For most scales, however, the local Poissonian assumption appears to be correct, and we assume it for the rest of this work.

To invert equation (3) in a model-independent way, we use the Richardson-Lucy (RL) method (see Appendix C in Binney & Merrifield 1998 and references therein). This iterative method is based on Bayes’s theorem. The kernel in equation (3) is Poissonian,

$$ K(N, \delta) = \frac{[\langle N \rangle (1 + \delta)]^N e^{-(\langle N \rangle + \delta)}}{N!}. \quad (4) $$

Since $N$ is within $[0, N_{\text{max}}]$ in practice, the $K(N, \delta)$ has to be normalized to unity with respect to $N$. In the probabilistic spirit of this method, the functions need rescaling: $\hat{K} = K / \sum_N K$ and $\hat{p}_1 = p \sum_N K$. Starting from an initial guess of $\hat{p}_1$, we can calculate $P_{N,1}$ via equation (3). From this, a better approximation of $\hat{p}$ is obtained,

$$ \hat{p}_2 = \sum_{N=0}^{N_{\text{max}}} \frac{P_N}{P_{N,1}} \hat{K}(N, \delta). \quad (5) $$

This is used in turn as the input for the next iteration. The improvement of the fit $P_{N,i}$ after the ith iteration is quantified by the cost function $\chi^2 = \sum_N (P_N / P_{N,i} - 1)^2$, where $P_N$ is the measured CIC distribution. One caveat is the phenomenon of “overlearning,” when the recovered probability distribution

![Fig. 1.—Richardson-Lucy inversion of an artificial PDF for a number of different iterations. The solid line corresponds to a lognormal input model from which the CPDF was generated according to eq. (3). Noise was added to the CPDF at the 5% level. The dotted line is the result after 5 iterations, the dash line corresponds to 12 iterations, while the dash-dotted line displays 100 iterations. After 100 iterations, overlearning problems appear (see Fig. 2, as well).](image1)

![Fig. 2.—Cost function $\log(\chi^2)$ of the Richardson-Lucy method (see text) as a function of iterations for the example in Fig. 1. Note that the change of the cost function becomes very slow after about 10–15 iterations, which would be a sensible stopping range for the iterations.](image2)
starts to fit small fluctuations in the measured \( P_N \). This can inject artificial features into the results after (too) many iterations. Numerical experiments indicate, that, after about 10 iterations, \( \log(\chi^2) \) becomes fairly small and changes little. The process is illustrated in Figure 1, for which a lognormal model for the PDF was used to generate a CPDF curve with noise generated at the 5% level. The inverted PDF in the figure is fairly accurate after 12 iterations, while the results of 100 iterations clearly show the signs of overlearning. Figure 2 displays \( \log(\chi^2) \) as a function of iteration. Indeed, beyond 10–12 iterations, it hardly changes; this is the point of diminishing returns, after which overlearning kicks in. The displayed cost function is fairly typical, and we have found that an inspection of \( \log(\chi^2) \) always allows a simple determination of a sensible stopping point. Because of the slow change in the cost function, the inversion is only logarithmically sensitive to finding the optimal number of iterations; therefore, this prescription is fairly robust.

Numerical experiments show (see § 3) that this method converges fairly fast and arrives at a robust result if \( \langle N \rangle \gtrsim 0.1 \). For smaller average counts, convergence slows down such that a few hundred iterations are needed, as Poisson noise starts to dominate. We recommend that our method be used down to this value for robust results.

Another difficulty of the RL inversion is purely computational. Typically, \( N_{\text{max}} \sim 10^4 \) for a CPDF of a dark matter dataset from an N-body simulation, and the kernel is a broad function. For each pair of \( N \) and \( P_N \), we need to sample at least one point at the peak of the kernel \( N/\langle N \rangle - 1 \), and a few hundred points on each side to keep the integration accurate. The last step thus costs about \( 10^{10} \) calculations of the kernel. Storing the kernel would require an array of floats of dimension \( 10^4 \times 10^6 \). This is unfeasible on a typical small workstation. Similar computational problems arise with other direct inversion methods, such as singular value decomposition, etc. This motivates us to define an alternative, model-dependent method, which requires more modest computational resources and can extend toward even smaller scales, i.e., sparser sampling of the distribution.

### 2.2. Skewed Lognormal Model Fit

An obvious alternative is to use physically motivated parametrized PDF models. This idea was explored by Kim & Strauss (1998), who adopted Gaussian Edgeworth expansions to the third order as a model for \( p(\delta) \) in deconvolving equation (3) and thereby fit for skewness and kurtosis. According to their investigations, such a method is severely limited to the weakly nonlinear regime, for which the Gaussian Edgeworth expansion is a good approximation. This is especially important in our application, in which not only the first few moments are interesting, but also very high order moments which contain information on the full behavior of the PDF. Therefore, we conclude that the Gaussian Edgeworth expansion is not a viable model for our purposes.

A more promising empirical model can be built based on the lognormal distribution (e.g., Coles & Jones 1991). This model is physically motivated, arising naturally from perturbation theory of the logarithm of the dark matter density field (Szapudi & Kaiser 2003). Indeed, the skewed lognormal distribution (SLN3) appears to be an excellent approximation for the PDF of the dark matter field on a wide range of scales (Colombi 1994; Ueda & Yokoyama 1996). It is plausible then that SLN3 might be a good approximation for galaxy distributions as well, as long as the bias is moderate; the tests of § 3 indeed confirm this idea. Next, we outline how one can proceed to invert equation (3) under the assumption of SLN3.

With the notation \( \rho = 1 + \delta \) (the density field), \( \Phi = \log \rho - \langle \log \rho \rangle \), and \( \sigma_\phi = \langle \Phi^2 \rangle \), the SLN3 model reads as

\[
p_3(\delta) d\delta = \left[ 1 + \frac{1}{3!} T_3 \sigma_\phi^3 H_3(\nu) + \frac{1}{4!} T_4 \sigma_\phi^4 H_4(\nu) \right] G(\nu) d\nu,
\]

where \( \nu \equiv \Phi/\sigma_\phi \), \( H_m(x) \) is an Hermite polynomial of degree \( m \), and \( G(x) \) is a Gaussian with zero mean and a variance of unity. The quantities \( T_3 \) and \( T_4 \) are the renormalized skewness and kurtosis of the field \( \Phi \), respectively (Colombi 1994):

\[
T_3 = \frac{\langle \Phi^3 \rangle}{\sigma_\phi^4}, \quad T_4 = \frac{\langle \Phi^4 \rangle - 3 \sigma_\phi^4}{\sigma_\phi^4}.
\]

| Cell Size R (h⁻¹ Mpc) | \( \langle N \rangle \) | \( \langle \log \rho \rangle \) | \( \sigma_\phi \) | \( T_3 \sigma_\phi^3 \) | \( T_4 \sigma_\phi^4 \) |
|-----------------------|-----------------|------------------|--------|-----------------|------------------|
| 2.21…………………. 64  | −1.157         | 1.185            | 1.110  | 1.289           |                   |
| 4.42………………… 512 | −0.798         | 1.108            | 0.757  | 0.781           |                   |
| 8.83………………… 4096 | −0.463         | 0.906            | 0.397  | 0.300           |                   |
| 17.66……………… 32768 | −0.201         | 0.633            | −0.008 | −0.080          |                   |
The convolution of $p_3(\delta)$ with the Poisson kernel, $\tilde{P}_N$, corresponds to a model of observed $P_N$ that depends on four parameters: $(\log \rho)_C, \sigma_8, T_3$, and $T_4$. These parameters can be fitted to the observed distribution, which in turn yields the best approximation of the real PDF by $p_3(\delta)$. The following Poisson likelihood function was used to fit the parameters:

$$L = \prod_N \frac{(MP_N)^{MP_N} e^{-MP_N}}{(MP_N)!},$$

where $M$ is the total number of cells used for the CIC measurements. Minimization of $-\ln L$ with respect to $(\log \rho)_C, \sigma_8, T_3$, and $T_4$ corresponds to a third-order SLN fit, which is realized with the Powell method (Press et al. 1992).

Note that we have experimented with other cost functions for fitting the parameters, including the conventional minimal $\chi^2$ and a likelihood function of Kim & Strauss (1998) similar to the above but with $MP_N$ replaced by $M\tilde{P}_N$. Numerical experience suggests that the above cost function, which is close to $\chi^2$ for large $MP_N$, gives the best bias reconstruction among the variations tested.

3. APPLICATION TO N-BODY SIMULATIONS

3.1. Simulations

The bias reconstruction method described above with both inversion methods at its core was extensively tested on a suite of dark matter mock galaxy catalogs extracted from $N$-body simulations by the GIF project of the Virgo Consortium (Kauffman et al. 1999). We used the $z = 0$ output for a $\Lambda$CDM universe with $\Omega_m = 0.3, \Omega_\Lambda = 0.7$, shape parameter $\Gamma = 0.21$, $\sigma_8 = 0.90$, $h = 0.7$, force soft length 20 $h^{-1}$kpc, and a simulation box $L = 141.3 h^{-1}$ Mpc. The simulations have $256^3$ particles of mass $1.4 \times 10^{10} M_\odot h^{-1}$. We used two mock galaxy catalogs, the GIF galaxy catalog and the GIF galaxy catalog for the Sloan Digital Sky Survey (SDSS). These
artificial galaxy catalogs have been produced using semi-analytical galaxy formation models (Kauffmann et al. 1999; Somerville et al. 2000; Benson et al. 2000) and have been studied extensively to explore biasing as functions of luminosity, scale, and redshift by Somerville et al. (2001).

Because of the computational difficulty of the RL inversion for the large dark matter simulations, we used the SLN3 model fit exclusively for the inversion of the their PDFs. We found that, above scales of $2.21 h^{-1}\text{Mpc}$, SLN3 provides an excellent model for the dark matter distribution, while on smaller scales, our model has increasing difficulty in fitting the tail of the distribution (see Fig. 3). On scales of $2.21 h^{-1}\text{Mpc}$, the measured CPDF develops a very long tail even in log space, which is not well fitted by the polynomial correction of the SLN3 beyond $N > 10^3$. However, since $\langle N \rangle = 64$ and we aim to use our method up to $\delta \leq 5$, we conclude that SLN3 will be a good approximation for our purposes. (See Table 1 for the parameters of SLN3 for the dark matter sample.)

### 3.2. A Null Test of Bias Reconstruction

In order to test the reliability of the bias extraction based on the SLN3 model fit, we selected several subsamples from the dark matter simulation by randomly sampling them at 10%, 1%, and 0.1%. The null test consists of recovering $b = 1$ from these catalogs; our method passed with flying colors.

#### TABLE 2

| Cell Size $R$ $(h^{-1}\text{Mpc})$ | $\langle N \rangle$ | $\langle \log \rho \rangle$ | $\sigma_\theta$ | $T_0 \sigma_\theta$ | $T_2 \sigma_\theta^2$ |
|---------------------------------|-------------------|-----------------|---------------|-----------------|-----------------|
| 2.21.......................... | 0.059             | −1.230          | 1.183         | 1.022           | 4.409           |
| 4.42.......................... | 0.471             | −1.318          | 1.591         | 0.752           | −1.749          |
| 8.83.......................... | 3.771             | −0.769          | 1.389         | −0.881          | 0.902           |
| 17.66......................... | 30.166            | −0.250          | 0.798         | −0.869          | 2.042           |
As a typical example, fits to the measured CPDFs at a 1% dilution level are presented in Figure 4. The fits are an excellent approximation to the measurements. The recovered \( C(\delta)/C_{14} \) is then used to reconstruct the bias of our diluted subsamples with respect to the full sample. Except for \( R = 2.21 h^{-1} \) Mpc, which has a ~10% peak difference, Figure 5 shows that the recovered bias is at most ~5% off from \( b = 1 \) for \( \delta < 5 \). We have performed the same study with several realizations, all giving similar results. The recovered bias at \( R = 2.21, 4.42, 8.83, \) and \( 17.66 h^{-1} \) Mpc are \( b = 0.96 \pm 0.12, 1.02 \pm 0.02, 1.00 \pm 0.03, \) and \( 1.01 \pm 0.02 \), respectively.

In addition, we have applied the Richardson-Lucy method to the 1% diluted subsamples. For \( \delta < 4 \), the inverted PDFs are identical to the PDFs from SLN3 model fits, except for the scale of \( 2.21 h^{-1} \) Mpc. Apparent differences occur at large \( \delta \). When the SLN3 model PDF is used as a reference for the full dark matter sample, the bias parameters recovered by Richardson-Lucy inversion are 1.16, 1.04, 1.03, and 0.94 on scales of \( R = 2.21, 4.42, 8.83, \) and \( 17.66 h^{-1} \) Mpc, respectively, for \( \delta < 5 \). From Figure 6, it is clear that, on large scales, both methods are in good agreement. On smaller scales, the deviation is due to the dominance of discreteness effects.

3.3. Bias from the GIF Mock Galaxy Catalog

In this section we subject our method to extensive testing. We aim to recover the bias function between a GIF mock galaxy catalog and the underlying dark matter catalog. Since we have both catalogs (obviously not the case with real data), we can test our reconstruction against a density-density scatter plot based on a cell-by-cell comparison of the two catalogs. Note that such direct comparison contains significant Poisson scatter, which appears to be an excellent approximation to the stochasticity of the bias in these simulations. The root mean square (rms) deviations of the reconstructed bias curves from the direct measurements are 0.35, 0.064, 0.033, and 0.059 for \( R = 2.21, 4.42, 8.83, \) and \( 17.66 h^{-1} \) Mpc (SLN3 fits) and 0.067, 0.021, and 0.021 for \( R = 4.42, 8.83, \) and \( 17.66 h^{-1} \) Mpc (RL fits), respectively.

Fig. 9.—Bias function \( \delta_g = f(\delta_m) \) for the mock galaxy catalog. Solid lines are from the SLN3 model galaxy PDFs; dashed lines are based on the Richardson-Lucy method (RL). The reference mass PDFs were obtained from an SLN3 fit. Circles are the \( \langle \delta_g | \delta_m \rangle \) measured directly from samples; error bars show their 1 \( \sigma \) scatter. The full width of \( \delta_m \) bins of the scatter plot is \( \Delta \delta_m = 0.22 \). The shaded area represents a simple Poisson scatter, which appears to be an excellent approximation to the stochasticity of the bias in these simulations. The root mean square (rms) deviations of the reconstructed bias curves from the direct measurements are 0.35, 0.064, 0.033, and 0.059 for \( R = 2.21, 4.42, 8.83, \) and \( 17.66 h^{-1} \) Mpc (SLN3 fits) and 0.067, 0.021, and 0.021 for \( R = 4.42, 8.83, \) and \( 17.66 h^{-1} \) Mpc (RL fits), respectively.

No. 1, 2004

RECOVERING THE NONLINEAR BIAS FUNCTION

31
seen in § 2, SLN3 is a good approximation for the dark matter distribution. For the galaxy distribution, SLN3 fits the tail of the distribution less accurately, even on large scales. As long as the Poisson kernel is not too broad, we can still use the SLN3 fit as an estimation of the PDF for small $\delta$. We can estimate the limit of the applicability of the SLN3 estimation as $\delta_{\text{max}} \leq N_{\text{max}} / \langle N \rangle - 1$, with $N_{\text{max}}$ being the $N$ at which our fit breaks off from the measured tail of the CPDF. (See Table 2 for parameters of SLN3 for the mock galaxy catalog.)

Cumulative PDFs based on the SLN3 model fit of the mock catalog and dark matter are shown in Figure 8. The bias function directly follows from equation (1). We plot $\delta$ as a function of $\delta_m$ in Figure 9. This is contrasted with the galaxy density—mass density scatter plot. Note that error bars represent the scatter due to “stochastic bias.” The shaded area on the plot represents simple considerations for the stochasticity of the bias based on the assumption that all scatter is due to Poisson noise. It is an excellent approximation to the measured scatter (the error bars), which appears to show that the scatter is indeed dominated by Poisson noise. We have checked that the measured error bars are only weakly dependent on the bin width $\Delta \delta$, represented by horizontal error bars: doubling it produced hardly noticeable effects. We have repeated the calculations using the Richardson-Lucy method for the galaxies only while still using an SLN3 model for the dark matter. The bias function is displayed in Figure 9 by dashed lines, except for the smallest scale, where $\langle N \rangle < 0.1$, which we established as the limit of applicability of this method.

3.4. Bias from the GIF Mock Galaxy Catalog for SDSS

This mock catalog contains more galaxies than the previous GIF mock galaxy catalog so as to match the number density of galaxies in the SDSS. The fits with SLN3 to its CPDFs are shown in Figure 10. Fitted parameters of the model are listed in Table 3. Bias functions are displayed in Figure 11.

3.5. Redshift Space

All the tests we have done so far in real space have been repeated in redshift space. The plausible assumption about the application of our method in redshift space is that, in absence of significant velocity bias, the redshift-space CPDFs will be modified similarly in the dark matter and galaxy catalogs. As long as this is correct, the bias function can be obtained by direct application of the method in redshift space. However, even if this assumption is only approximate, the dark matter distribution in redshift space is still recovered, which means that the effects of galaxy formation are decoupled from the evolution of dark matter; this is the principal goal of bias recovery. The tests in this section show that redshift distortions have a small effect on our procedure; thus our recipe can be safely applied to redshift surveys without significant corrections. It has to be borne in mind, however, that this statement is somewhat model dependent; we assume that the simulations and the semianalytic models used to create mock catalogs are close enough to reality that this statement will hold for real data.

The CPDFs and their SLN3 fits for the dark matter sample, the GIF mock galaxy catalog, and the mock galaxy catalog for SDSS are shown in Figures 12, 13, and 14, respectively. The recovered bias functions of the GIF mock galaxies are displayed in Figure 15, while those of the SDSS mock galaxies are in Figure 16.

The recovery of bias in redshift space is just as successful as in real space. We have also compared the recovered redshift space bias function to that of real space. Results based on RL inversion only are shown in Figures 17 (GIF galaxy mock catalog) and 18 (mock galaxy catalog for SDSS). The difference between the two curves is generally small; in fact for $\delta_m \leq 2$, the effect of redshift distortions on the bias function is negligible. The apparent difference between the GIF curves on the scale of $4.42\ h^{-1}\ Mpc$ is mostly because the RL inversion is pushed to its limits at $\langle N \rangle = 0.47$.

4. CONCLUSION AND DISCUSSION

We have presented a new method to extract the bias function from galaxy catalogs. Since our method uses direct comparison statistics extracted from simulations and data, it is applicable to a large range of scales. In particular, its domain of validity includes the nonlinear regime, which has proved to be impenetrable to other previous methods. This is all the

| Cell Size $R$ ($h^{-1}\ Mpc$) | $\langle N \rangle$ | $\langle \log \rho \rangle$ | $\sigma_\Phi$ | $T \sigma_\Phi$ | $T \sigma_\Phi^2$ |
|-------------------------------|-------------------|-----------------|-------------|--------------|----------------|
| 2.21                         | 0.7198            | -2.446          | 2.167       | 0.181        | -0.300         |
| 4.42                         | 5.758             | -1.631          | 1.971       | -0.387       | 0.557          |
| 8.83                         | 46.064            | -0.796          | 1.347       | -0.338       | 0.968          |
| 17.66                        | 368.514           | -0.318          | 0.824       | -0.231       | 0.280          |
Fig. 11.—Same as Fig. 9, but for the SDSS mock galaxy catalog. The rms deviations for increasing scales are 0.13, 0.1, 0.022, and 0.045 (SLN3), and 0.11, 0.019, 0.024, and 0.05 (RL), respectively.

Fig. 12.—CPDFs and their SLN3 model fits for dark matter in redshift space. Solid lines correspond to the measurements, while dashed lines show SLN3 model fits. For these curves from left to right, cell sizes at which CPDFs are measured are $R = 2.21$, 4.42, 8.83, and 17.66 $h^{-1}\text{Mpc}$.

Fig. 13.—Same as Fig. 12, but for the GIF mock galaxy catalog in redshift space.
more important, since the most reliable data are available on nonlinear scales. In addition, most available data on the largest scales still have significant nonlinear “contamination,” to which our method is completely insensitive. In addition to expanding the range of applicability, our method has an accuracy that rivals all other methods, because it is uses the full counts-in-cells distribution, which is sensitive to very high order statistics.

The new technique is based on comparing the cumulative probability distribution functions in simulations and data. To turn this idea into a robust and reliable method, a difficult technical challenge had to be met: the reconstruction of the continuous probability distribution of density fluctuations from counts-in-cells measurements. This is a delicate and potentially unstable inversion, for which we have proposed two solutions. One is a model-independent inversion using a Richardson-Lucy iteration, while the other is a model-dependent fit based on the skewed lognormal approximation (SLN3). The former method is useful down to scales where \( N > 0.5 \) and for relatively smaller number of particles because of computational constraints. The SLN3 fitting is useful possibly to even slightly smaller scales, and it is
Fig. 16.—Same as Fig. 9, but for the GIF mock galaxy catalog for the SDSS in redshift space. The rms deviations for increasing scales are 0.022, 0.22, 0.024, and 0.053 (SLN3) and 0.16, 0.038, 0.012, and 0.017 (RL).

Fig. 17.—Comparison of recovered bias functions of the GIF galaxy catalog in real space (solid lines) and in redshift space (dashed lines).

Fig. 18.—Same as Fig. 17, but for the mock galaxy catalog for the SDSS.
feasible for large simulations of arbitrarily high particle number. Both methods fail on very small scales, \( \langle N \rangle \leq 0.1 \), where Poisson noise dominates and the \( \delta \) resolution is very sparse. Therefore, the former lends itself naturally to fitting the CPDF in galaxy catalogs, while the latter is useful for large simulations. The range of reconstruction of the bias function is \(-1 \leq \delta \leq 5\), and typically we could recover the bias from fairly realistic simulations at the 5\% level. This suggests that our application of our method to contemporary catalogs, such as the SDSS and the Two Degree Field (2dF), will constrain bias at an accuracy close to the absolute limit determined by systematic errors.

Note that the above considerations are valid, even when the SLN3 fits have slight discrepancies at the low \( N \) tail, which is in our target range of \(-1 \leq \delta \leq 5\). The low \( N \) discrepancies exists only for large scales with a correspondingly large mean. In these cases, the failure interval is limited to a very small region, \(-1 \leq \delta \leq \delta_{\text{min}} = N_{\text{min}}/\langle N \rangle - 1\), where \( N_{\text{min}} \) is the lowest value for which the fit is good. For example, the SDSS mock catalog for 17.66 \( h^{-1} \) Mpc has \( N_{\text{min}} \sim 20 \) and \( \langle N \rangle = 368.5\); thus the failure range is \(-1 \leq \delta \leq -0.95\), i.e., negligible.

Most of our efforts have been centered on reconstructing the bias function that itself represents the mean bias. We have found, however, that part of the scatter, if not most, is due to Poisson scatter. Our reconstruction method corrects for this source of error to the fullest possible extent. In addition, we have found that a simple shot-noise model gives an excellent approximation to the stochastic component of the bias.

The expected number of galaxies in a region with galaxy density fluctuation \( \delta \) is simply \( N_g = \langle N_g \rangle (1 + \delta_g) \). If we assume a Poisson variance around this value (and neglect discreteness in the dark matter catalog, which is a good approximation), we have

\[
\frac{\Delta N_g}{N_g} = \frac{\Delta \delta_g}{1 + \delta_g} = \frac{1}{\sqrt{\langle N_g \rangle (1 + \delta_g)}}.
\]

It follows that \( \Delta \delta_g = \left[ \frac{1}{\langle N_g \rangle (1 + \delta_g)} \right]^{1/2} \). This simple formula is shown in Figures 9, 11, 15, and 16 as a shaded area and provides an approximation to the error bars typically at the 10\% level, with the largest deviation being 50\%, mainly at large \( \delta \). It appears that Poisson scatter provides the dominant fraction of the stochasticity of the bias. Note that there are signs of sub-Poisson scatter in some of the Figures. The above is hardly more than a toy model, and its degree of success is remarkable.

A convenient parametrization of the (mean) bias function relies on a Taylor series expansion,

\[
\delta_g = f(\delta_m) = \sum_{k=0}^{\infty} \frac{b_k}{k!} \delta_m^k.
\]

We adopted this form to fit our results empirically from the GIF and GIF-SDSS mock catalogs for \( \delta_m \in [-1, 5]\). We have found that an expansion up to \( k = 2 \) is always sufficient. The results, based on RL inversion, are shown in Tables 4 and 5. It also quantifies the difference between redshift and real space. For instance, \( b = b_1 \) typically within a few percent for the two cases. For analysis of real data, if \( \sigma_b \) in the simulation does not match the true underlying dark matter distribution, \( b_1 \) would shift accordingly, but the other parameters describing the nonlinearity of bias would remain reasonably robust (E. Branchini 2003, private communication).

A subtlety with the above formula is worth emphasizing again: in this paper, we have related the smoothed galaxy density field to the smoothed dark matter density field. Other methods might use “unsmoothed” fields, which really means that smoothing is done on much smaller scales than those scales considered in the measurement. The meaning of a truly unsmoothed galaxy density field, let alone its Taylor series expansion in terms of a truly unsmoothed dark matter density field, is somewhat dubious. As is well known, bias and smoothing do not commute, which means that our bias coefficients might have slightly different meanings than the quantities noted with the same letters in other works.

No statistical method would be complete without a way of placing error bars on the estimates derived from the method. Since our method is fully nonlinear and relies on direct

| \( R \) (\( h^{-1} \) Mpc) | \( h_0 \) | \( h_1 \) | \( h_2/2! \) | \( h_0 \) | \( h_1 \) | \( h_2/2! \) |
|-----------------|--------|--------|--------|--------|--------|--------|
| 4.42........... | 0.0085 | 1.085  | 1.085  | 0.0085 | 1.085  | 1.085  |
| 8.83........... | 1.016  | 1.18   | 1.085  | 0.10   | 1.17   | 0.066  |
| 17.66...........| 0.022  | 1.00   | 0.062  | 0.022  | 1.00   | 0.047  |

**TABLE 5**

**BIAS PARAMETERS FOR THE MOCK SDSS GALAXY CATALOG**

| \( R \) (\( h^{-1} \) Mpc) | \( h_0 \) | \( h_1 \) | \( h_2/2! \) | \( h_0 \) | \( h_1 \) | \( h_2/2! \) |
|-----------------|--------|--------|--------|--------|--------|--------|
| 2.21........... | 0.155  | -0.052 | 0.155  | -0.11  | 1.57   | -0.037 |
| 4.42........... | 0.071  | 1.12   | 0.01   | -0.084 | 1.10   | 0.032  |
| 8.83........... | 0.062  | 1.20   | 0.032  | -0.038 | 1.19   | 0.013  |
| 17.66...........| 0.015  | 1.20   | 0.059  | 0.005  | 1.19   | 0.014  |
comparison of galaxy catalogs with simulations, the only robust way of producing error bars is a Monte Carlo estimate. The procedure is obvious: one has to repeat all the calculations using a suite of simulations representing the data. In fact, we demonstrated this in § 3.2, when we measured bias in several realizations of the randomly diluted samples. If the dark matter simulation is large and dense enough, most of the variance will come from data. While for most of our present investigation, only one realization was at our disposal, the CPU budget would allow the analysis of a large number of simulations. It took about 1 hr to measure CIC in dark matter and galaxy catalogs, about 30 minutes for one SLN3 fit, and up to a few hours for RL inversion (a typical value for our calculation of the mock SDSS). We have used a 2.4 GHz dual Xeon workstation with 2 GBytes of memory.

In this work, we exclusively used the Poisson model to relate the discrete galaxy distribution to the underlying continuous field. This approximation might break down, especially on very small scales, where halo models are known to predict sub-Poisson scatter (Casas-Miranda et al. 2002; Berlind & Weinberg 2002). Such theories, once firmly established, can be naturally incorporated into our formalism by simply generalizing the kernel. Another application, for which the kernel will need modification, is recovery of the bias from two-dimensional angular catalogs. In that case, the corresponding kernel would depend on the selection function, but the core idea of the method would still work. Note that our method is well suited for determination of relative bias between different populations of objects without the use of any simulation. These generalizations, as well as applications to real data, are in preparation.

This work was supported by NASA through grants AISR NAG5-11996, ATP NAG5-12101, and by NSF through grants AST 02-06243 and ITR 1120201-128440. The simulations in this paper were carried out by the Virgo Supercomputing Consortium using computers based at Computing Centre of the Max-Planck Society in Garching and at the Edinburgh Parallel Computing Centre. The data are publicly available at http://www.mpa-garching.mpg.de/Virgo.

REFERENCES

Bardeen, J., Bond, J. R., Kaiser, N., & Szalay, A. 1986, ApJ, 304, 15
Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M., & Lacey, C. G. 2000, MNRAS, 311, 793
Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587
Binney, J., & Merrifield, M. 1998, Galactic Astronomy (Princeton: Princeton Univ. Press)
Branchini, E., Zehavi, I., Plionis, M., & Dekel, A. 2000, MNRAS, 313, 491
Casas-Miranda, R., Mo, H. J., Sheth, R. K., & Boerner, G. 2002, MNRAS, 333, 730
Colombi, S. 1994, ApJ, 435, 536
Dekel, A., & Lahav, O. 1999, ApJ, 520, 24
Efstathiou, G., et al. 2002, MNRAS, 330, L29
Feldman, H. A., Frieman, J. A., Fry, J. N., & Scoccimarro, R. 2001, Phys. Rev. Lett., 86, 1434
Fry, J. N. 1994, Phys. Rev. Lett., 73, 215
Fry, J. N., & Gaztañaga, E. 1993, ApJ, 413, 447
Gaztañaga, E., & Frieman, J. A. 1994, ApJ, 437, L13
Kaiser, N. 1984, ApJ, 284, L9
Kaufmann, G., Colberg, J. M., Diaferio, A., & White, S. D. M. 1999, MNRAS, 303, 188
Kim, R. S. J., & Strauss, M. A. 1998, ApJ, 493, 39
Lahav, O., & Saslaw, W. 1992, ApJ, 396, 430
Lahav, O., et al. 2002, MNRAS, 333, 961
Matsubara, T. 1995, ApJS, 101, 1
———, 1999, ApJ, 525, 543
Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. H. 1992, Numerical Recipes: The Art of Scientific Computing (2nd ed.; Cambridge: Cambridge Univ. Press)
Sigad, Y., Branchini, E., & Dekel, A. 2000, ApJ, 540, 62
Somerville, R. S., Lemson, G., Kolatt, T., & Dekel, A. 2000, MNRAS, 316, 479
Somerville, R. S., et al. 2001, MNRAS, 320, 289
Szapudi, I. 1995, Ph.D. thesis, Johns Hopkins Univ.
———, 1998a, MNRAS, 300, L35
———, 1998b, ApJ, 497, 16
Szapudi, I., & Kaiser, N. 2003, ApJ, 583, L1
Ueda, H., & Yokoyama, J. 1996, MNRAS, 280, 754
Verde, L., et al. 2002, MNRAS, 335, 432
———, 2003, ApJS, 148, 195
Zaroubi, S., Branchini, E., Hoffman, Y., & da Costa, L. N. 2002, MNRAS, 336, 1234