Mon. Not. R. astr. Soc. (1991) **248,** 1–13

A lognormal model for the cosmological mass distribution

Peter Coles[1]★ and Bernard Jones[2]†

[1]*Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH*
[2]*Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen 0, Denmark*

Accepted 1990 June 18. Received 1990 June 18; in original form 1990 May 14

SUMMARY

We discuss the use of a lognormal (LN) random field as a model for the distribution of matter in the Universe. We find a number of reasons why this should be a plausible approximation to the distribution of density irregularities obtained by evolving from Gaussian initial conditions. Unlike straightforward linear theory, the model always has $\rho > 0$ but is arbitrarily close to the Gaussian at early times. It has the added advantage that, like the Gaussian model, all its statistical properties can be formulated in terms of one covariance function.

A number of interesting and important difficulties with the statistical treatment of density perturbations are revealed by an analysis of this model. In particular, the LN model is not completely specified by its moments. We explain why this could be true for the actual matter field. We also show that the usual method of representing the three- and four-point correlation functions of galaxies, in terms of the parameters Q and R, is not useful for discriminating between Gaussian and non-Gaussian fluctuations, and propose better parameterizations in terms of the skewness and kurtosis of the three- and four-point distributions, respectively.

Other characteristics of the model, such as topology (genus curves, etc.), multifractal behaviour, void probabilities and biasing (behaviour of 'peaks' relative to background fluctuations) are also discussed. The model also provides a way of checking the consistency of treatments of large-scale streaming motions in the Universe by allowing us to determine the scale at which linear theory cannot be accurate for both the matter and velocity fields.

We discuss a possible model for the number-count distribution of galaxies, based on the LN distribution but allowing for discreteness effects which can make the distribution of log n appear non-Gaussian, and show how to construct Monte-Carlo simulations of point patterns (in one-, two, or three-dimensions) which contain correlations of all orders.

1 INTRODUCTION

Ever since the pioneering studies of Neyman & Scott (1952) it has been understood that our knowledge of the distribution of galaxies is statistical and, therefore, incomplete in the sense that we will never be able to predict the specific locations of galaxies around us. It is also true that our knowledge is incomplete even in a statistical sense: the distribution of galaxies in space is only completely specified by the hierarchy‡ of n-dimensional joint probability density functions, $f_n(\rho)$, connecting the density at different spatial positions and our knowledge of these is restricted to low-order moments and related functions. Just as the analysis of galaxy catalogues gives us only partial information about the f_n, so is it true that physical models do not allow the f_n to be expressed in any analytic form. The late stages of galaxy formation are usually modelled by discrete numerical simulations (Efstathiou *et al.* 1985, and references therein) and one cannot extract any more information about the f_n from these than one could from a galaxy catalogue. Analytic approximations for the growth of non-linear structure such as the Zel'dovich approximation (Zel'dovich 1970; Bond & Couchman 1988; Shandarin & Zel'dovich 1989), second-order perturbation theory (Juszkiewicz, Sonoda & Barrow 1984; Coles 1990) or those techniques based upon the

★Present address: Astronomy Unit, Queen Mary and Westfield College, Mile End Road, London E1 4NS.
†Present address: Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH.
‡Doob (1953) first demonstrated this fact. Doob defines two stochastic processes to be *stochastically equivalent* only if all finite dimensional joint distributions are identically equal for the two distributions.

Burgers equation (Kofman & Shandarin 1988; Gurbatov, Saichev & Shandarin 1989; Shandarin & Zel'dovich 1989; Kofman, Pogosyan & Shandarin 1990) allow some progress but still do not allow us to specify the f_n completely. Both theory and observations are therefore statistically underspecified.

The only fully specified model for the density field of widespread use in cosmology is the Gaussian random field. The finite-dimensional joint distributions of such a field are all multivariate Gaussian pdf's and the model is therefore statistically complete. This complete analytical specification allows many complex properties of the local geometry of such fluctuations to be calculated analytically which has made them a favourite amongst applied mathematicians (Rice 1945; reprinted in Wax 1954; Cartwright & Longuet-Higgins 1956; Longuet-Higgins 1957; Adler 1981; Vanmarcke 1983). The numerous physical motivations for assuming Gaussian statistics for the smaller linear density perturbations that are commonly supposed to have been present at recombination, particularly from inflationary models (Barrow & Coles 1990, and references therein), have also led to a vast output of cosmological literature on the statistical properties of such fluctuation fields (Kaiser 1984a; Politzer & Wise 1984; Peacock & Heavens 1985; Bardeen *et al.* 1986, hereafter BBKS; Jensen & Szalay 1986; Coles 1986; Couchman 1987a,b; Lumsden, Heavens & Peacock 1989; Coles 1989).

A Gaussian random field, however, can only be a model for the linear density field, i.e. in the limit of zero fluctuation amplitude. As soon as we start to deal with finite rms fluctuations, σ, the Gaussian model assigns a non-zero probability to regions of negative density (e.g., Fry 1986). Although this probability might be acceptably small when σ is tiny, it must grow with time as gravitational instability takes over, so that the non-linear density field at late times cannot be described consistently by a Gaussian random field. As we discussed above, we have no theoretical means for specifying the f_n in this regime.

This difficulty highlights the importance that should be attached to the construction of self-consistent *stochastic* models for the density perturbation field. That is, models which are not necessarily physically motivated but which are completely specified statistically and which do not violate common-sense conditions such as $\rho > 0$. Such models allow us to see what it is that various clustering statistics are (or are not) measuring and suggest better ways of discriminating between theory and observations.

The usefulness of this approach was realized by Neyman & Scott (1952) who constructed models for the discrete galaxy distribution which are based upon the Poisson cluster formula (see also Peebles 1980). Further discrete models are the Thermodynamic model (Saslaw & Hamilton 1984), the Bose–Einstein or negative binomial model (Carruthers & Shih 1983) and the scaling ansatz of Schaeffer (1985). Not all of these are, in fact, completely specified in the sense given above but they are useful in illustrating some properties of statistical tools commonly used in the analysis of galaxy catalogues (e.g., Fry 1985). Unfortunately, few consistent stochastic models for the continuous density fluctuation field exist and those that do are difficult to handle analytically (Coles & Barrow 1987; Coles 1988; Coles 1989). Furthermore, with the currently fashionable theoretical emphasis upon biased galaxy formation, the relationship between the matter distribution and the galaxy distribution has become obscure. A consistent model for the density field should be the starting point for studies of biasing.

In this paper, we shall study one particular stochastic model for the cosmological density field, which we call a lognormal (LN) random field. The model is fully specified statistically; its intimate relationship to a Gaussian random field allows many properties of the model to be calculated in the same way as for a Gaussian field (Coles & Barrow 1987; Coles 1988; Lucchin & Matarrese 1988; Coles 1989). The model always has $\rho > 0$ but becomes arbitrarily close to Gaussian statistics at early times. After giving some technical background in Section 2 and illustrating some of the motivations for thinking that the LN model might be a useful one (Section 3), we study this field in some depth to see what we can learn about the analysis of real density fields. In Section 4 we look at properties of the moments of the one-point distribution. In Section 5 we calculate correlation functions of all orders. In Section 6 we show how one can connect our model for the density field to a Gaussian model for the peculiar velocity field. In Section 7 we show how to construct a model for the discrete number-count distribution from our definition of the continuous density field. Some miscellaneous properties (maxima, topology, simulations and multifractal behaviour) are discussed in Section 8. Finally, in Section 9, we summarize the main conclusions and outline areas for further work in the light of our analysis.

2 DEFINITIONS AND TECHNICAL BACKGROUND

This section contains some background material required for understanding the technical details of the model we are discussing. We shall take as our starting point the definition of a Gaussian random field (Adler 1981; Vanmarcke 1983; Peacock & Heavens 1985; BBKS). Such a field, $X(r)$ will have a one-point probability density function (pdf) $f_1(x)$ given by a normal distribution with some mean and variance; that is $X \sim N(\mu, \sigma^2)$:

$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2]. \tag{1}$$

Furthermore, all the higher order n-point pdf's, $f_n(\boldsymbol{x})$, of field values at different positions $\boldsymbol{r}_i$ are multivariate normal:

$$f_n(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{M}|^{-1/2} \exp\left(-\frac{1}{2}\sum_{i,j} \boldsymbol{M}^{-1}_{ij} x_i x_j\right), \tag{2}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i = x(\boldsymbol{r}_i)$ and $\boldsymbol{M}$ is the covariance matrix

$$\boldsymbol{M}_{ij} = \langle\langle (X_i - \mu)(X_j - \mu)\rangle\rangle.$$

Note that $M_{ii} = \sigma^2$ and we have assumed that each of the n variates has the same mean $\mu_i = \mu$, i.e. that the field is stationary or statistically homogeneous. For a Gaussian random field the covariances $\boldsymbol{M}_{ij}$ are determined completely by the covariance function, $\Xi(r)$, which depends only on $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ if the field is statistically isotropic. Thus

$$\Xi(r_{ij}) = \{[X(\boldsymbol{r}_i) - \mu][X(\boldsymbol{r}_j) - \mu]\} = \boldsymbol{M}_{ij}. \tag{3}$$

The covariance function completely specifies all the finite dimensional pdf's for a Gaussian random field. The relatively simple form of the f_n for this random field allows one to obtain exact solutions for many of the properties of such fields (Rice 1945, reprinted in Wax 1954; Adler 1981; Vanmarcke 1983; Peacock & Heavens 1985; BBKS). Arbitrary non-Gaussian fields offer no such possibilities. However, in Coles & Barrow (1987) we showed how to construct a wide variety of model non-Gaussian fields by using various non-linear transformations of Gaussian fields. One particular field we looked at was the lognormal random field which we obtained by transforming a Gaussian field via

$$Y(\boldsymbol{r}) = \exp[X(\boldsymbol{r})]. \tag{4}$$

The transformation (4) results in the following one-point distribution of Y:

$$f_1(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right]\frac{dy}{y}, \tag{5}$$

where μ and σ^2 are the mean and variance of the underlying Gaussian field X. A variate drawn from such a distribution is usually denoted $Y \sim \Lambda(\mu, \sigma^2)$ by the statisticians. The resulting mean and variance of Y will be discussed later in Section 4. Note that equation (5) is not enough to specify all the statistical properties of the lognormal random field. We need also to extend the definitions of the multivariate normal distribution to

$$f_n(y_1, \ldots, y_n) = (2\pi)^{-n/2}|\boldsymbol{M}|^{-1/2} \times \exp\left[-\frac{1}{2}\sum_{i,j} \boldsymbol{M}^{-1}_{ij} \log(y_i)\log(y_j)\right] \prod_{i=1}^{n} \frac{1}{y_i} \tag{6}$$

where $\boldsymbol{M}$ is the covariance matrix of the X-values (3). The covariance function for y, which will be denoted $\xi(r)$, is discussed later in Section 5.

3 MOTIVATIONS FOR THE LOGNORMAL MODEL

Now that we have explained how the lognormal random field is constructed, we shall review the motivations for thinking that it might provide a good model for the distribution of density fluctuations in the Universe. It must be said at the outset that some of these suggest only that $f_1(\rho)$ might be described by a lognormal distribution $\Lambda(\mu, \sigma^2)$ (5) and say nothing about the n-point distributions in general. However, (6) gives the most straightforward extension of (5) and this simple model still retains many interesting features.

3.1 Observations

Hubble (1934) first noticed that the distribution of galaxy counts in two-dimensional cells on the sky could be well approximated by (5). This possibility has also been discussed by Peebles (1980). Recent more extensive galaxy redshift surveys by the Durham group (Hale-Sutton *et al.* 1989; Shanks *et al.* 1989) do, in fact, reveal a distribution of log N that looks rather skewed. We are proposing the lognormal as a model for the continuous density field, however, and it may be that number counts are a biased tracer of the underlying matter (see below) or that the discreteness of the distribution causes some spurious skewness (see Section 6) or, of course, both. Although the evidence is therefore circumstantial, it certainly does not contradict the model.

3.2 The central limit theorem

One of the primary motivations for supposing that primordial density fluctuations might be Gaussian is the Central Limit Theorem. This states that, if

$$Y = \sum_{i=1}^{n} X_i,$$

where the X_i are independent variates with finite variance, then $Y \to N(\mu, \sigma^2)$ as $n \to \infty$, independent of the distributions of the underlying X. In inflationary models, each Fourier mode of the perturbation field comes inside the horizon with a random phase and the resulting superposition inside the horizon leads to a Gaussian random field (Barrow & Coles 1990, and references therein).

We can build a non-linear analogue of this theorem by assuming that, instead of a superposition, we have a product of independent influences:

$$Y = \prod_{i=1}^{n} X_i. \tag{7}$$

If this is the case then log Y conforms to the requirements of the central limit theorem so that $Y \to \Lambda(\mu, \sigma^2)$. It may be seen then that the lognormal is a paradigm for non-linear noise just as the Gaussian is for linear noise. In an astrophysical context, Zinnecker (1984) showed how hierarchical fragmentation or coagulation models lead to lognormal mass functions; his model is just a special case of the above argument. In the context of the explosion models of large-scaled structure, Ostriker (1986) showed how the distribution of masses of structures formed by fragmentation of gas in colliding shock fronts should be roughly lognormal. For an extensive discussion of the genesis of lognormal and related models in this way see Aitchison & Brown (1969) and Crow & Shimizu (1988).

3.3 Kinematics

Let us consider the growth of inhomogeneities in an expanding universe (Peebles 1980; Shandarin & Zel'dovich 1989). We shall work with coordinates comoving with the Hubble expansion $\boldsymbol{x} = \boldsymbol{r}/a(t)$ where $a(t)$ is the cosmic scale factor and t is proper time. We split the total velocity field $\boldsymbol{u}$ into Hubble expansion and peculiar velocity components:

$$\boldsymbol{u} = \dot{a}(t)\boldsymbol{x} + \boldsymbol{v}(\boldsymbol{x}, t). \tag{8}$$

The equation of continuity for the matter flow is

$$\frac{\partial \rho}{\partial t} + 3\left(\frac{\dot{a}}{a}\right)\rho + \frac{1}{a}\nabla\cdot(\rho\boldsymbol{v}) = 0. \tag{9}$$

We shall now switch time coordinate to conformal time: $a d\tau = dt$ and work with a different density variable: $\varrho = \rho a^3$. The equation (9) then becomes

$$\frac{\partial \varrho}{\partial \tau} + (\boldsymbol{v}\cdot\nabla)\varrho + \varrho(\nabla\cdot\boldsymbol{v}) = 0. \tag{10}$$

Writing the total time derivative as $d/d\tau$, this becomes

$$\frac{1}{\varrho}\frac{d\varrho}{d\tau} = -(\nabla\cdot\boldsymbol{v}). \tag{11}$$

If the initial peculiar velocity field is Gaussian then so is $(\nabla\cdot\boldsymbol{v})$ (Rice 1945, in Wax 1954). If it continues to grow linearly then the peculiar velocity divergence stays Gaussian and scales with τ: $(\nabla\cdot\boldsymbol{v})_\tau = (\nabla\cdot\boldsymbol{v})_{\tau_0}(\tau/\tau_0)$. We therefore find that

$$\varrho(\boldsymbol{x}) = \varrho_0 \exp[\varepsilon(\boldsymbol{x})\tau^2], \tag{12}$$

where ε is a Gaussian random field $= -\frac{1}{2}(\nabla\cdot\boldsymbol{v})_{\tau=\tau_0}\tau_0^{-1}$. The expression (12) is of a similar form to equation (4). We can therefore regard a lognormal distribution as being a kinematical model for the density distribution, obtained by extrapolating the equation of continuity into the non-linear regime by assuming linear velocity fluctuations. Fry (1986) argued that $\boldsymbol{v}$ should be more robust to higher order corrections than ϱ; there is certainly a less pressing need for $\boldsymbol{v}$ to depart from Gaussian behaviour since $\boldsymbol{v}$ is not constrained to be >0. The model therefore seems rather natural. It is, however, a very crude model: no account is taken of the geometry of the local velocity field so, unlike the Zel'dovich approximation (Zel'dovich 1970), it cannot reproduce pancaking in three dimensions. It should, however, give a good description of the general tendency of the matter to clump in the weakly non-linear regime. It is interesting to compare this model with a one-dimensional version of the Zel'dovich approximation (Shandarin & Zel'dovich 1989). In this approximation, the local matter density grows as

$$\varrho = \varrho_0(1-\varepsilon\tau^2)^{-1}. \tag{13}$$

The full (3D) version of the Zel'dovich approximation has a rather more complex dependency on the geometry of the local ε-field but behaves in qualitatively the same way (Shandarin & Zel'dovich 1989). Expanding each of these in powers of τ and noting that $\tau \propto t^{1/3}$ when the Universe is matter-dominated, we find that both approximations reproduce linear theory:

$$\delta = (\varrho-\varrho_0)/\varrho_0 \sim t^{2/3} \tag{14}$$

although the higher order contributions are different. Note that the Zel'dovich approximation predicts the formation of caustics ($\rho\to\infty$) which the lognormal model does not. We shall return to this problem in Section 4.

Note also that the treatment above is entirely Lagrangian – the coordinates we are using will fall into developing non-linear overdensities – so cannot be accurate to treat structures in Eulerian space except possibly in the weakly non-linear regime.

3.4 Biased galaxy formation

Various aspects of lognormal distributions have been discussed in the literature in connection with biased galaxy formation. In the simplest biased models, galaxy or cluster formation occurs only where the local mass density exceeds some threshold level ν. This is known as 'sharp' or 'step-function' thresholding. Various authors have attempted to refine this approach by arguing that, in practice, there is more likely to be a non-linear relation between the mass-density and the resulting number of galaxies. Kaiser & Davis (1985) proposed an extension of Kaiser's (1984a,b) biased model of cluster formation that would work for massive superclusters of galaxies. Their model relates the number of galaxies per unit mass inside and outside the supercluster via

$$\left(\frac{N_g}{M}\right)_{sys} = \exp(\kappa_g\Delta_l)\left(\frac{N_g}{M}\right)_{global}, \tag{15}$$

where Δ_l is the linear theory density contrast and κ_g depends on the model of galaxy formation: in simple models $\kappa_g \simeq \nu^2(1+z_f)^{-1}$ where z_f is the redshift of galaxy formation. Szalay (1988) has discussed this in terms of using an exponential threshold function rather than a step-function (see also Borgani & Bonometto 1990). Note that equation (15) leads to a lognormal distribution of N_g/M if the linear density fluctuations are Gaussian. This is not quite the same as saying that the density fluctuations are lognormal, which is what our model involves, but the similarities are sufficient to provide further reasons for such an in-depth study.

3.5 Simplicity

All the above discussion provides only partial motivation for our model. The lognormal is one of the simplest ways of defining a fully self-consistent random field which always has $\rho>0$ and, most importantly, is one of the few non-Gaussian random fields for which interesting properties are calculable analytically. (For an interesting application in a different cosmological context, see Barrow & Morgan 1983.) It has no more free parameters than the Gaussian from which it is derived and in many cases analytic results are obtained more easily for this model than the Gaussian (Vanmarcke 1983; Coles & Barrow 1987; Coles 1988; Szalay 1988).

In subsequent sections we shall see that even such a simple model exhibits some strange behaviour and teaches us much about the limitations of statistical treatments of non-linear density fluctuations.

4 PROPERTIES OF THE ONE-POINT LN DISTRIBUTION

Here we shall discuss some of the properties of $\Lambda(\mu,\sigma^2)$ relevant to large-scale structure. The most obvious starting point is to consider the moments of the distribution about the origin, usually denoted μ_n':

$$\langle Y^n\rangle = \mu_n' = \int_0^\infty y^n\Lambda(\mu,\sigma^2)\,dy. \tag{16}$$

These are easily evaluated explicitly in terms of μ and σ^2,

$$\mu_n' = \exp(n\mu + n^2\sigma^2/2). \tag{17}$$

The moments about the mean, which we shall denote μ_n, are obtained in the usual way. We shall only be interested in μ_2 and μ_3. The variance is $\Sigma^2 = \mu_2' - (\mu_1')^2$ so that

$$\mu_2 = \exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]. \tag{18}$$

The third moment about the mean is

$$\begin{aligned}\mu_3 &= \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 \\ &= \exp(3\mu+3\sigma^2/2)[\exp(\sigma^2)-1]^2[\exp(\sigma^2)+2].\end{aligned} \tag{19}$$

It is interesting also to compute the skewness of this distribution (Kendall & Stuart 1988):

$$\Gamma = \left[\frac{(\mu_3)^2}{(\mu_2)^3}\right]^{1/2}$$

$$= [\exp(\sigma^2) - 1]^{1/2}[\exp(\sigma^2) + 2]. \quad (20)$$

Note that $\Gamma \to 0$ as $\sigma \to 0$. These results obtain for a general lognormal field. When we discussed kinematics in Section 3, we suggested that the best model for ϱ was given by equation (12). For simplicity we shall work with a new variable $\chi = \varrho/\varrho_0$. According to our model we find that $\chi \sim \Lambda(0, \sigma_\varepsilon^2\tau^4)$, where σ_ε^2 is the variance of the ε-field introduced in Section 2. This, however, leads to an immediate problem when we calculate the mean. Substituting these parameters in (17) we get

$$\langle\chi\rangle = \exp(\sigma_\varepsilon^2\tau^4/2). \quad (21)$$

In other words the mean density of the Universe increases with time. The reason for this problem is that our extrapolation of the continuity equation into the non-linear regime incorporated no requirement that matter be conserved globally. We can circumvent the problem by 'renormalizing' χ so that it always has the same mean $\langle\chi\rangle = 1$: $\chi \mapsto \chi \exp(-\sigma_\varepsilon^2\tau^4/2)$. That is, $\chi \sim \Lambda(-\sigma_\varepsilon^2\tau^4/2, \sigma_\varepsilon^2\tau^4)$. The moments of χ about the origin become

$$\langle\chi^n\rangle = \exp[n(n-1)\sigma_\varepsilon^2\tau^4/2]. \quad (22)$$

The variance, Σ^2, is given by

$$\Sigma^2 = \exp(\sigma_\varepsilon^2\tau^4) - 1. \quad (23)$$

Recall that $\tau \propto t^{1/3}$ during matter domination. We can see that the rms fluctuation of the density $\Sigma \sim \tau^{2/3}$ to lowest order in t so that, again, we are reproducing linear theory. The skewness of χ is just

$$\Gamma = [\exp(\sigma_\varepsilon^2\tau^4) - 1]^{1/2}[\exp(\sigma_\varepsilon^2\tau^4) + 2] \quad (24)$$

so that the skewness vanishes at early times. We shall return to this in Section 5. Fig. 1 shows the behaviour of the distribution of χ with time: it looks Gaussian at early times and becomes progressively more and more skewed as time progresses.

Many properties of statistical fluctuations can be expressed succinctly in terms of the Moment Generating Function, $M(s)$, defined by

$$M(s) = \sum_0^\infty \frac{\mu_n' s^n}{n!} = \int d\chi f(\chi) e^{s\chi} = \langle e^{s\chi}\rangle, \quad (25)$$

where the μ_n' are defined by equation (17) (see Peebles 1980; Fry 1985). Inserting the form of the lognormal into this equation leads to a divergent integral for any real, positive (non-zero), s. [For negative s, the integral in (25) is finite, but the series expansion diverges. See the remarks below concerning the Void Probability Function.] This strange behaviour is related to an important property of the lognormal distribution: *it is not completely determined by its moments.* In other words, one could construct at least one different distribution with all moments identical to those of a given lognormal distribution. This problem is discussed extensively in the statistical literature (Aitchison & Brown 1969; Kendall & Stuart 1977; Crow & Shimizu 1988). The

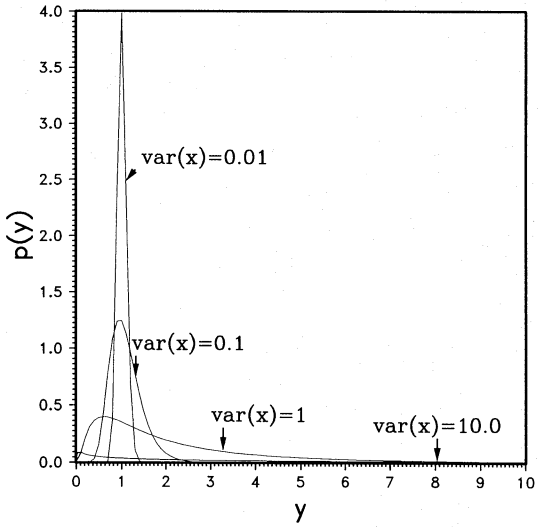

Figure 1. Lognormal distributions $P(y)$ generated from underlying Gaussian distributions with variance shown. The variable y is related to the Gaussian x via $y = \exp[x - \mathrm{var}(x)/2]$, as discussed in the text prior to equation (22), in order to conserve matter globally.

condition that any set of moments completely determine a probability distribution can be put in a number of ways. The simplest is the requirement that

$$\lim_{n\to\infty} \sup(\mu_n' s^n/n!)^{1/n} < 1 \quad (26)$$

or, using Stirling's approximation for $n!$

$$\lim_{n\to\infty} \sup(\mu_n')^{1/n}/n < k/s, \quad (27)$$

where k is a constant and s is any real non-zero s. These conditions are not satisfied for the lognormal since $(\mu_n')^{1/n} \sim \exp(\mu + n\sigma^2/2)$.

It seems therefore that our non-linear density distribution cannot be described completely, in a statistical sense, by a complete specification of all its moments. One might argue here that this does not have any implications for the description of the real distribution of matter in space because we just happen to have picked a pathological distribution in our choice of the exponential transformation of the initial conditions. Recall, however, how the Zel'dovich (1970) approximation boils down to a similar transformation that would have $\chi = (1 - \varepsilon\tau^2)^{-1}$ (13). If ε is Gaussian in this case, it is very straightforward to show that, far from being completely determined by its moments, none of the moments of the distribution of χ actually exist, even for arbitrarily small σ_ε. This is explained by noting that there will always be a finite number of caustics in a field evolved according to (13). These might be very improbable ($\varepsilon\tau^2 = 1$ for a caustic which is very unlikely if $\sigma_\varepsilon\tau^2 \ll 1$), but still cause a divergence of the integral when it is taken over an infinite space. The widespread use and success of the Zel'dovich approximation suggest that the problem we have discovered in connection with the lognormal is likely to be true of any non-linear density field with a long positive tail. Indeed, one way of stating the conditions (26, 27) is that the one-point pdf should not decay any slower than an exponential at large χ (Kendall & Stuart 1977, p. 423). Note, however, that it is the very high peaks which cause the divergence of the integral.

When such a high peak forms, the local density field will probably not be well-described by a lognormal and the Zel'dovich approximation certainly breaks down when a caustic forms. It may be that the strong clustering in such regions is better described, for example, by one of Fry's (1985) hierarchical distributions which are all uniquely determined by their moments. Whether the actual matter density field possesses this non-uniqueness property is therefore open to some doubt, but it is quite possible that it does.

In physical terms, the reason for non-uniqueness is that all the high-order moments are dominated by the very high-density regions of the density field. In a distribution where there are very high overdensities and no compensating underdensities (i.e. a skewed field bounded to the left at zero), the high-order moments tell us only about the densest clusters and practically nothing about the voids in between. We could therefore re-arrange the voids, keeping the high-order moments the same but obtaining a different overall distribution. This is the reason behind White's (1979) suggestion that void probabilities should provide a good discriminant between various types of model fields. One can actually see this problem explicitly in the lognormal model by looking at the void probability function which is simply written as

$$P_0(V) = M(-1)$$

after suitable scaling of the variable (White 1979; Fry 1985). We saw that for the LN distribution, the Moment Generating Function for negative values of s can only be defined in terms of an integral (which, unfortunately, cannot be done analytically) and the series expansion in terms of moments diverges. We can see therefore that a unique expression for the void probabilities exists, but it cannot be expressed in terms of all the moments of the distribution. This demonstrates that, in general, $P_0(V)$ contains *at least* information about all moments of the distribution and, in cases like the LN distribution, it in fact contains *more* information than is contained in all the moments.

As a final point before we go on to discuss correlation functions, it is worth stressing that, just as the complete set of moments does not completely characterize the one-point distribution, so the complete set of n-point covariance functions cannot fully specify the n-point distribution. (The moments basically determine the correlation functions at zero lag.) This places even more importance upon searches for alternative statistical measures of clustering than the traditional correlation function approach initiated by Peebles and his co-workers (Peebles 1980, and references therein).

5 CORRELATIONS IN THE LN MODEL

5.1 Two-point correlation functions

We use Peebles' (1980) method for covering the covariance functions of continuous random fields into the n-point correlations functions of a point data set. The two-point correlation function is calculated from the moments of the two-point pdf $f_2(\chi_1, \chi_2)$:

$$\xi(r) = \frac{\langle (\chi_1 - \langle\chi\rangle)(\chi_2 - \langle\chi\rangle)\rangle}{\langle\chi\rangle^2} \tag{28}$$

so that

$$1 + \xi(r) = \frac{\iint \chi_1 \chi_2 f_2(\chi_1, \chi_2)\, d\chi_1\, d\chi_2}{[\int \chi f_1(\chi)\, d\chi]^2}. \tag{29}$$

We shall work with the χ-field defined in Section 4 to avoid a proliferation of factors of ϱ_0. The χ_i correspond to different positions $\boldsymbol{r}_i$ and $r = |\boldsymbol{r}_1 - \boldsymbol{r}_2|$. The integral is easily evaluated by substituting the form (6) for the two-point distribution and then transforming back to a bivariate Gaussian. We find that

$$\xi(r) = \exp[\Xi(r)] - 1. \tag{30}$$

Devotees of biased clustering theory will recognize this as the Politzer & Wise (1984) approximation for the correlation function of high-level regions of a Gaussian random field; it is an exact expression for the (unbiased) correlation function of a lognormal field in terms of the covariance funcion of the underlying Gaussian field.

Our discussion of kinematics in Section 3.3 suggested that the generating Gaussian field should be related to the divergence of the peculiar velocity field $\varepsilon(\boldsymbol{r}) = (\boldsymbol{\nabla}\cdot\boldsymbol{v})$. If this field has covariance function $\Xi_0(r)$ at $\tau = \tau_0$ then the matter correlation function scales as

$$\xi(r, \tau) = \exp\left[\Xi_0(r)\left(\frac{\tau}{\tau_0}\right)^2\right] - 1. \tag{31}$$

Again, we find that linear theory is reproduced, with $\xi(r)$ scaling as $\sim t^{4/3}$. Furthermore, as we enter the non-linear regime the correlation function begins to steepen, as demonstrated in Fig. 2. This is a well-known feature of non-linear gravitational clustering, as evidenced by numerical simulations (see, Efstathiou *et al.* 1985). One difficulty with the model (31) is that it predicts that $\xi(r) = 0$ at the same r for all τ. Numerical experiments are generally too noisy to determine the zero-crossing with any real accuracy, but alternative

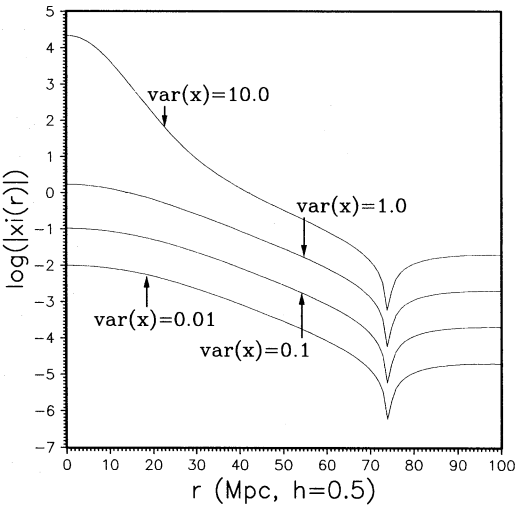

Figure 2. Evolution of the matter correlation function via the expression (31). The correlation functions are those generated by transformations of Gaussian processes with variances shown. Steepening of the correlation function is only perceptible for the highly non-linear model. The underlying correlation function is that of Cold Dark Matter correlation function smoothed on a scale of 10 Mpc ($h = 0.5$).

theoretical techniques do suggest that the zero-crossing should move in the non-linear regime (Bond & Couchman 1988; Coles 1990). However, we know already that this model is only going to be reliable in the weakly non-linear regime anyway and these theoretical studies show that the zero-crossing is robust to small-scale non-linear effects when the large-scale clustering is still weak.

The observational data suggest that the galaxy–galaxy two-point correlation function is well-fitted by the form

$$\xi_{gg}(r) \simeq \left(\frac{r}{r_0}\right)^{-1.8} \tag{32}$$

with $r_0 \simeq 5\ h^{-1}$ Mpc (see Shanks *et al.* 1989 for a discussion of the data). Our expression would suggest a break with any self-similar form for $\xi(r)$ as non-linear clustering develops and therefore seems to be at odds with the observations. In fact, the observations do suggest that there may be a break in the power law at $r \simeq 10\ h^{-1}$ Mpc (Shanks *et al.* 1989) and the power-law behaviour may well be generated by strong non-linear behaviour (Peebles 1980; see also the comments below about hierarchical higher order corelations) rather than evidence for any scale-invariance in the initial data, i.e. a power-law form for $\Xi_0(r)$. Our model therefore provides a simple not-too-unconvincing heuristic model for the growth of correlations into the non-linear regime.

5.2 Three-point correlation functions

The three-point correlation function $\zeta(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3) = \zeta_{123}$ is obtained by a generalization of (28):

$$\zeta_{123} = \frac{\langle(\chi_1 - \langle\chi\rangle)(\chi_2 - \langle\chi\rangle)(\chi_3 - \langle\chi\rangle)\rangle}{\langle\chi\rangle^3}. \tag{33}$$

Noting that $\langle\chi\rangle = 1$ for this field, we find that

$$\zeta_{123} = \langle\chi_1\chi_2\chi_3\rangle - \langle\chi_1\chi_2\rangle - \langle\chi_2\chi_3\rangle - \langle\chi_3\chi_1\rangle + 2. \tag{34}$$

If we denote $\xi_{ij} = \xi(r_{ij})$ and note that $1 + \xi_{ij} = \langle\chi_i\chi_j\rangle$, we find that

$$\langle\chi_1\chi_2\chi_3\rangle = \exp[\Xi_{12} + \Xi_{23} + \Xi_{31}] \tag{35}$$

and eliminating Ξ we find that

$$\zeta_{123} = \xi_{31}\xi_{12} + \xi_{12}\xi_{23} + \xi_{23}\xi_{31} + \xi_{12}\xi_{23}\xi_{31}, \tag{36}$$

so that the three-point correlations of the LN random field obey a well-known form known as the *Kirkwood scaling relation* (Kirkwood 1935; Peebles 1980; Matarrese, Lucchin & Bonometto 1986; Bonometto, Lucchin & Matarrese 1987; Szalay 1988).

Observational determinations of the three-point correlation function were carried out by Groth & Peebles (1977) who found that their results were consistent with a purely hierarchical form (Peebles & Growth 1975) for ζ_{123}:

$$\zeta_{123} \simeq Q(\xi_{31}\xi_{12} + \xi_{12}\xi_{23} + \xi_{23}\xi_{31}) \tag{37}$$

with $Q \simeq 1$ out to about $2\ h^{-1}$ Mpc. Importantly, these results are obtained only from two-dimensional data, which leads to some uncertainty in the interpretation (Bonometto *et al.* 1987). Whether the triple term absent from (36) is significant has been a question of considerable debate in recent times. Groth & Peebles (1977) argue strongly that this term is absent in the data; Shanks' (1979) analysis suggested that one cannot discriminate the two possiblities. Bonometto *et al.* (1987) found that their de-projection technique gave a best fit with the triple term present. More recently still, Tóth, Hollósi & Szalay (1989) have analysed the three-point correlations of rich clusters and find no evidence for the triple term, which seems bad news for biased theories of rich cluster formation which also predict the existence of this term (Kaiser 1984a; Politzer & Wise 1984; Jensen & Szalay 1986). In any case, the observations of galaxy triplet correlations are only reliable out to distances where the galaxy pair correlations are very strong. They do not strongly constrain the weakly non-linear behaviour we are interested in. The generation of hierarchical higher order correlations during the strongly non-linear regime is extremely plausible (Davis & Peebles 1977; Peebles 1980; Hamilton 1988). One also wonders what sense there is in quoting tight error bars on $Q(1 \pm 0.1$ is quoted by Tóth *et al.* 1989) when different determinations of $\xi(r)$ give a factor of ~ 10 variation at these distances (Shanks *et al.* 1989).

Numerical simulations (Fry 1985; Melott & Fry 1986) show that it is not difficult for any model to reproduce $Q \sim 1$. It may be that this is because of the generation of hierarchical clustering during the strongly non-linear phase, but we shall argue here that, if one wishes to discriminate between arbitrary models in the weakly non-linear regime, then Q is a bad statistic to choose.

For illustrative purposes we shall consider three-point correlations for point configurations which are equilateral triangles $\xi_{12} = \xi_{23} = \xi_{31} = \xi(r)$. The restricted three-point correlation function is then

$$\zeta^{eq} = \exp[3\xi(r)] - 3\exp[\xi(r)] + 2. \tag{38}$$

Putting $x = \exp[\xi(r)]$ we find that

$$\zeta^{eq} = (x-1)^2(x+2). \tag{39}$$

We thus obtain

$$Q^{eq} = \frac{\zeta^{eq}}{3\xi(r)^2} = (x+2)/3 \tag{40}$$

(remembering that $\langle\chi\rangle = 1$), so that Q should really be a function of distance; the uncertainties in the quoted observational results may be sufficient to conceal this behaviour in the data. This does not tend to zero at early times in our model, even though we know that the lognormal model looks arbitrarily close to a Gaussian when the generating variance is small. The statistic Q is often interpreted as a measure of the skewness of the distribution (e.g. Peebles 1980; Fry 1985). This analysis demonstrates that it is not a good measure of (multivariate) skewness because it does not go to zero even when the univariate distribution is unskewed (Section 4): $Q^{eq} \to 1$ as $\sigma_\varepsilon^2 \to 0$ and $\Gamma \to 0$. A more sensible way of discriminating between models would be to use an extension of the skewness we introduced in (20):

$$\Gamma^{eq}(r) = \left[\frac{(\zeta^{eq})^2}{\xi(r)^3}\right]^{1/2} = [(x-1)(x+2)]^{1/2} \tag{41}$$

which goes to zero as $r \to \infty$. The extension to the non-degenerate case is straightforward. We can actually calculate

Γ for the two-dimensional data of Groth & Peebles (1977), assuming that Q is actually constant as they claim. The result is that

$$\Gamma(\theta) \sim \theta^{(1-\gamma)/2}, \tag{42}$$

where $\gamma \sim 1.7$–1.8. Clearly the skewness is decreasing with angular separation so the distribution may be tending towards a multivariate Gaussian for larger separations.

We must stress here that we are not arguing that Q should not be used in galaxy clustering studies. If the correlations are strong then they might well be hierarchical in form and Q is then the best way to parameterize them. However, it would be misguided to attempt to continue its use into the region where the correlations are weak; it is an inappropriate statistical tool for measuring slight departures from Gaussian behaviour.

5.3 Higher order correlations

It is a straightforward matter to extend the calculations above to the general nth order correlation functions. We find that the Kirkwood scaling relation is generalized to

$$\xi_n(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n) = \prod_{i>j} [\xi(\boldsymbol{r}_i, \boldsymbol{r}_j) + 1] - 1. \tag{43}$$

Observational results have only been obtained with any confidence for the fourth-order term (Fry & Peebles 1978; Sharp, Bonometto & Lucchin 1984). The results are less certain but seem to be consistent with the hierarchical form

$$\eta_{1234} = R_a[\xi_{12}\xi_{23}\xi_{34} + \ldots(12 \text{ terms})] + R_b[\xi_{12}\xi_{13}\xi_{14} + \ldots(4 \text{ terms})], \tag{44}$$

where the first sum is over all 12 distinct 'snake' connections and the second is over the four distinct 'star' graphs. The values of $R_a \simeq 2.5$ and $R_b \simeq 4.3$ are uncertain by at least ~ 50 per cent. We can see an analogous problem to that which we found for Q by again specializing to the case where galaxies are at the corners of regular tetrahedra so that the ξ_{ij} are all equal. The model (44) would have $\eta^{\rm eq} = \text{constant}\ \xi(r)^3$, whereas our lognormal model has

$$\eta^{\rm eq} = (x-1)^3[x^3 + 3x^2 + 6x + 6] = \xi^3[\xi^3 + 6\xi^2 + 15\xi + 16] \tag{45}$$

giving us

$$R^{\rm eq} \propto \eta^{\rm eq}/\xi^3 \propto [\xi^3 + 6\xi^2 + 15\xi + 16] \tag{46}$$

so, like Q, the R would depend on distance in this model. More importantly, this parameter does not go to zero as $r \to \infty$ and $\xi \to 0$. Again the best discriminator would be obtained by defining a *kurtosis* coefficient:

$$\kappa^{\rm eq} = \eta^{\rm eq}/\xi^2 = \xi[\xi^3 + 6\xi^2 + 15\xi + 16] \tag{47}$$

which has the desired properties. Because of the possibility that 'snake' and 'star' configurations might cluster differently as in the model (44), it is more difficult to generalize this definintion to the case of non-degenerate separations. The simplest way is to define two coefficients κ_a and κ_b so that the denominator in (47) comprises a sum over snakes and stars, respectively as in (44).

6 CONSISTENT TREATMENT OF DENSITIES AND VELOCITIES

Considerable theoretical attention has been paid in recent times (Górski *et al.* 1989, and references therein) to the analysis of streaming motions revealed by redshift-independent estimators of galaxy distances (Collins, Joseph & Robertson 1986; Lynden-Bell *et al.* 1988; Górski *et al.* 1989, and references therein). These studies invariably involve the application of purely linear perturbation theory to the peculiar velocity field. As we discussed in Section 3.3, it seems more reasonable to invoke Gaussian statistics for the local velocities than for the matter density. Our ansatz for $\varrho(\boldsymbol{x})$ allows us to determine in a very straightforward way at what scale the observed (linear) velocity correlations imply a departure from linear theory for the matter correlations.

The scale at which non-linear effects become important is particularly relevant for studies of biasing (Coles 1986; BBKS; Coles 1989; Lumsden *et al.* 1989) which have demonstrated that the biased correlation function is sensitively dependent upon the shape of the underlying matter correlation function. Bond & Couchman (1988) and Coles (1990) have employed approximate analytical techniques (the Zel'dovich approximation and second-order perturbation theory respectively) to estimate the departures from linear theory and found them to be slight on scales $\sim 25\ h^{-1}$ Mpc for the popular CDM model. We shall see in this section how the model (12) provides a simple estimate of the scale at which departures from Gaussian statistics become important.

To show the consistency of our treatment of matter and velocity fields we shall use the notation of Górski (1988) (see also Górski *et al.* (1989). In Górski's notation, the covariance function of the total (3D) peculiar velocity field in linear theory is

$$\begin{aligned}\Psi(\boldsymbol{r}) &= \langle \boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x}+\boldsymbol{r}) \rangle \\ &= \Psi_\perp(r) + 2\Psi_\parallel(r) \\ &\simeq \frac{H_0^2 \Omega_0^{1.2}}{2\pi^2} \int_0^\infty P(k) \frac{\sin(kr)}{kr}\, dk,\end{aligned} \tag{48}$$

where $P(k)$ is just the power spectrum of the linear density fluctuations. Following the work of Rice (1945) (reprinted in Wax 1954; see also BBKS; Coles 1989) we find that

$$\begin{aligned}\Xi(r) &= \langle \varepsilon(\boldsymbol{x})\, \varepsilon(\boldsymbol{x}+\boldsymbol{r}) \rangle \\ &\sim \langle [\nabla \cdot \boldsymbol{v}(\boldsymbol{x})][\nabla \cdot \boldsymbol{v}(\boldsymbol{x}+\boldsymbol{r})] \rangle \\ &= -\, d^2\Psi(r)/dr^2\end{aligned} \tag{49}$$

so that

$$\Xi(r) \sim \int_0^\infty P(k)\, k^2 [\sin(kr)/kr]\, dk$$

which is just the linear theory matter correlation function, $\xi_{\rm l}(r)$. The usefulness of this connection is that we can estimate the non-linear contribution to $\xi(r)$ by comparing

$\xi_{\rm l}$ with $[\exp(\xi_{\rm l}) - 1]$.

This is a much simpler check on the validity of the linear approach than those by Bond & Couchman (1988) or Coles

(1990) and is still more rigorous than the usual assumption that one can use linear theory until $\xi_1 \simeq 1$. If we look at the normalization of the CDM model (with $h=0.5$ and $\Omega_0=1$, for example) we find that, in the $b=1$ case, the difference between linear and LN extrapolations is about 13 per cent at $r=10\ h^{-1}$ Mpc and falls to less than 1 per cent at $r=25\ h^{-1}$ Mpc, figures which are surprisingly close to those obtained by Bond & Couchman (1988) and Coles (1990). Our model seems to be a helpful consistency check on linear theory.

Note that we do not expect the actual shape of $\xi(r)$ to be given accurately at small separations by (31); what we do claim is that we can use (31) to estimate the order-of-magnitude departures from linear theory in the weakly non-linear regime.

7 A MODEL FOR THE GALAXY NUMBER-COUNT DISTRIBUTION

Our discussion of the lognormal model has so far been limited to the treatment of the fluctuations in total matter density with position in space. It may be that there is a very non-linear relationship between the local matter density and the local number-density of galaxies; this is the idea behind biased galaxy formation (see Section 8.3 and references therein). To avoid further complexities, however, we shall suggest a simple way of extending the (continuous) lognormal distribution to provide a description of the (discrete) number-count distribution which does not invoke biasing and which has, at least on average, a linear relationship between galaxy number density and mass density.

Let us suppose that ϱ is distributed according to $\varrho \sim \Lambda(\mu, \sigma^2)$ (see Section 2.1). An easy (but not unique) way of obtaining a discrete distribution of galaxies is to assume that the number of galaxies at a point is given by a Poisson process with some mean density λ but that this λ is given by the local mass density:

$$\lambda = \beta\rho, \tag{50}$$

where β is some constant, normalized to give the correct number density of the objects in question. Note that we actually assumed a relation of the form (50) in our treatment of the correlations in Section 5 (Peebles 1980). With these two assumptions the probability of finding r galaxies in a small volume δV, which is incorporated in β, at a point where the local mass density is ρ is just

$$\Pr(N=n) = \int_0^\infty p(\lambda(\ P(n|\lambda)\, d\lambda$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{n!} \int_0^\infty e^{-\lambda} \lambda^{n-1} \exp\left\{ -\frac{[\log(\lambda/\beta)-\mu]^2}{2\sigma^2} \right\} d\lambda. \tag{51}$$

There is no known simpler way of writing this distribution, which is an example of a unimodal lattice distribution (Johnson & Kotz 1970, p. 31; Crow & Shimizu 1988, pp. 195–210). Fry (1985) suggested a method for deriving such distributions from the moment generating function $M(s)$. Fry's prescription will not work for our model as the moment generating function does not exist for the lognormal distribution, as we found in Section 3.

From (51) it is straightforward to calculate the kth order factorial moments (Kendall & Stuart 1977, p. 65):

$$\mu_{[k]} = \alpha^k \omega^{k^2/2}, \tag{52}$$

where $\alpha = \beta e^{\mu}$ and $\omega = e^{\sigma^2}$. From these we quickly determine both the central and non-central moments. The results are

$$\begin{aligned}
\mu_1' &= \alpha\sqrt{\omega} \\
\mu_2' &= \alpha\sqrt{\omega}[\alpha\omega\sqrt{\omega}+1] \\
\mu_3' &= \alpha\sqrt{\omega}[\alpha^2\omega^4+3\alpha\omega\sqrt{\omega}+1] \\
\mu_4' &= \alpha\sqrt{\omega}[\alpha^3\omega^7\sqrt{\omega}+6\alpha^2\omega^4+7\alpha\omega\sqrt{\omega}+1]
\end{aligned} \tag{53}$$

and so on for the non-central moments (i.e. about the origin). For the first few central moments we find

$$\begin{aligned}
\mu_2 &= \alpha\sqrt{\omega}[\alpha(\omega-1)\sqrt{\omega}+1] \\
\mu_3 &= \alpha\sqrt{\omega}[\alpha^2\omega(\omega-1)^2(\omega+2)+3\alpha(\omega-1)\sqrt{\omega}+1] \\
\mu_4 &= \alpha\sqrt{\omega}[\alpha^3\omega\sqrt{\omega}(\omega^4+2\omega^3+3\omega^2-3) \\
&\quad +6\alpha^2\omega(\omega-1)(\omega^2+\omega-1)+\alpha\sqrt{\omega}(7\omega-4)+1].
\end{aligned} \tag{54}$$

Note that, when $\alpha\sqrt{\omega}$ is small, the last term dominates in each of these expressions; in particular, the non-central moments are equal and are therefore just those of a Poisson distribution. The distribution is then discreteness-dominated. When $\alpha\sqrt{\omega}$ is large the first term dominates and the non-central moments are just those of the lognormal distribution derived in Section 4. It can be seen, therefore, that the distribution of $\log n$ will look significantly non-Gaussian for some parameter choices. It may be that discreteness fluctuations of the sort we have introduced here could result in a skewed distribution of $\log n$. Furthermore, note that the procedure of 'binning' galaxies to produce a histogram of number-counts does not necessarily leave the distribution unchanged. If the bin size is greater than some scale length of order the coherence length of the fluctuations (Coles 1988), the binned distribution will tend to look like a Gaussian independently of the actual distribution. When looked at in terms of $\log n$, the distribution would therefore look skewed. The observations mentioned in Section 2.1 do not therefore rule out the use of a lognormal as a model for the underlying matter distribution; detailed modelling is needed to test this hypothesis.

8 FURTHER PROPERTIES OF LN DENSITY FLUCTUATIONS

8.1 Maxima

We discussed already (in Section 3) that the LN model can be regarded as a kind of biasing in itself. It is also possible to calculate the statistical properties of the *peaks* of LN fluctuations to the same extent as one can for Gaussian fluctuations. The details of the calculations are given in Coles (1989) and so will not be repeated here. The usefulness of such calculations is that one can examine how sensitively the clustering properties of maxima (or high-level regions) depend on the underlying statistics. The need for such an examination is clear because we know Gaussian statistics cannot be correct when $\delta \sim 1$ (Section 1; Fry 1986). In fact, the LN random field exhibits similar biased correlations to the Gaussian (Coles 1989) so this problem seems not to be

severe for the usual theoretical treatments of biased galaxy formation. Indeed, this behaviour seems to be part of a general tendency for the statistics of high peaks to be only rather weakly dependent on the exact form of the underlying statistics (Coles & Barrow 1987; Catelan, Lucchin & Matarrese 1988).

8.2 Topology

In Sections 4 and 5 we demonstrated the importance of finding alternative statistical characterizations of clustering patterns to the traditional approach based on moments and correlation functions. One promising avenue in recent years has been to look at the *topology* of the distribution obtained by smoothing the point distribution on some scale. Topological characteristics such as the genus, Euler–Poincaré characteristic, mean curvature etc. change when the field so obtained is 'sliced' at different threshold levels ν. The variation of topology with ν can be used to discriminate between different clustering models (Gott, Melott & Dickinson 1986; Gott, Weinberg & Melott 1987; Hamilton, Gott & Weinberg 1987; Weinberg, Gott & Melott 1987; Couchman 1987b; Coles 1988; Melott, Weinberg & Gott 1988). An explicit calculation of these topological characteristics for two-dimensional LN random field was carried out by Coles (1988). The extension to 3D is trivial when one notes that the unique mapping (4) of a Gaussian on to the LN field ensures that the topology of the LN field above the level x must be the same as that of the Gaussian field above the level $\log x$. We shall not give the detailed results here as they essentially involve a replacement of ν by $\log \nu$ in the genus curves of the above references.

What is worth pointing out, however, is that the technique used by the topologists cited above to calculate the slicing levels for comparing topologies is *not* the best way to utilize the topological behaviour for statistical discrimination between Gaussian and non-Gaussian fields. Gott *et al.* (1987) pick their level by choosing the same fraction of points to be above the level in each comparison case. For a LN field, this procedure would produce a genus curve *identical* to that of a Gaussian field (apart from a normalization factor): since there is a one-to-one mapping, (4), the top 1 per cent points of the X-field map to the top 1 per cent points of the Y-field and the mapping therefore preserves the topology above fixed fractile levels. One could therefore not discriminate between the Gaussian and a LN field by this procedure. Furthermore, any density field obtained by a unique mapping would behave in a similar way. A particular example is the Zel'dovich (1970) approximation: in the early stages of non-linear evolution the points contained in a volume element at Eulerian position x all arrive there from the same Lagrangian position $\boldsymbol{q}$ so there is a similar unique mapping. (The weak dependence of topology so defined in this regime can also be observed in numerical simulations; Melott, private communication.) Later on the matter flow becomes a multistream flow ('shell crossing' occurs: Shandarin & Zel'dovich 1989) and the uniqueness is broken. In general circumstances, however, the procedure outlined above does not make the most of the discrimatory power of topological characteristics. A better way of choosing the level ν would be to pick ν to be a fixed number of standard deviations above the mean. Although this makes the procedure into a parametric one, a substantial increase in the power of the discriminant would be obtained (see also Coles & Barrow 1987).

8.3 Simulating LN fluctuations

Although, as we have seen, many of the properties of LN fields are tractable analytically, there are circumstances where one might have to resort to Monte-Carlo simulations. Coles (1988, 1989) showed how to simulate continuous lognormal fields on a grid using FFT techniques. The extension to three dimensions is straightforward and will not be given here.

It is useful, however, to show how the technique can be extended to allow the simulation of clustered point patterns. Previously the techniques available for such simulations have been limited to variations on the Poisson cluster models pioneered by Neyman & Scott (1952) (see also Peebles 1980; Barrow & Bhavsar 1987, and references therein). A laborious trial-and-error computer model for galaxy clustering studies was presented by Soneira & Peebles (1978) which incorporated correlations up to fourth-order but no further. More recently, Messina *et al.* (1990) have performed N-body simulations for non-Gaussian initial perturbations. Their algorithm for generating the starting configuration is based on the Zel'dovich approximation (Section 3.3) and can only generate mildly non-Gaussian initial conditions. Furthermore, they are restricted to *white noise* fluctuations (i.e. possessing a correlation function which is identically zero everywhere) so that their method cannot reproduce initial fields possessing information about all the finite-dimensional distributions.

The LN field allows us to construct discrete simulations very easily. The 2D case is outlined here but the procedure can be generalized very straightforwardly. One simply picks the desired correlation function for the continuous LN density field, calculates the covariance of the Gaussian field required to generate it {i.e. $\Xi(r) \sim \log[1+\xi(r)]$}, performs a random-phase FFT to generate the ε-field on a grid and then maps this on to a χ-field, defined on a grid, via $\chi_{ij} = \exp(\varepsilon_{ij})$. If one desires the mean number of galaxies for each realization to be $\langle N_g \rangle$ then the probability of finding a galaxy at the position $\{i, j\}$ is just

$$p_{ij} = 1 - \exp(-\beta\chi_{ij}) \simeq \beta\chi_{ij} (\text{small } \beta), \tag{55}$$

where the choice of β determines $\langle N_g \rangle$:

$$\langle N_g \rangle = N_p - \sum_{i,j} \exp(-\beta\chi_{ij}), \tag{56}$$

where N_p is the number of grid positions ('pixels'). The actual number of galaxies will vary from realization to realization and care must be taken that the distribution is not discreteness-dominated (see Section 7). Examples of point patterns generated in this way are given in Fig. 3.

8.4 Multifractal description

Another alternative to traditional characterizations of galaxy clustering is the multifractal formalism (Jones *et al.* 1988, and references therein). We shall concentrate on its use for

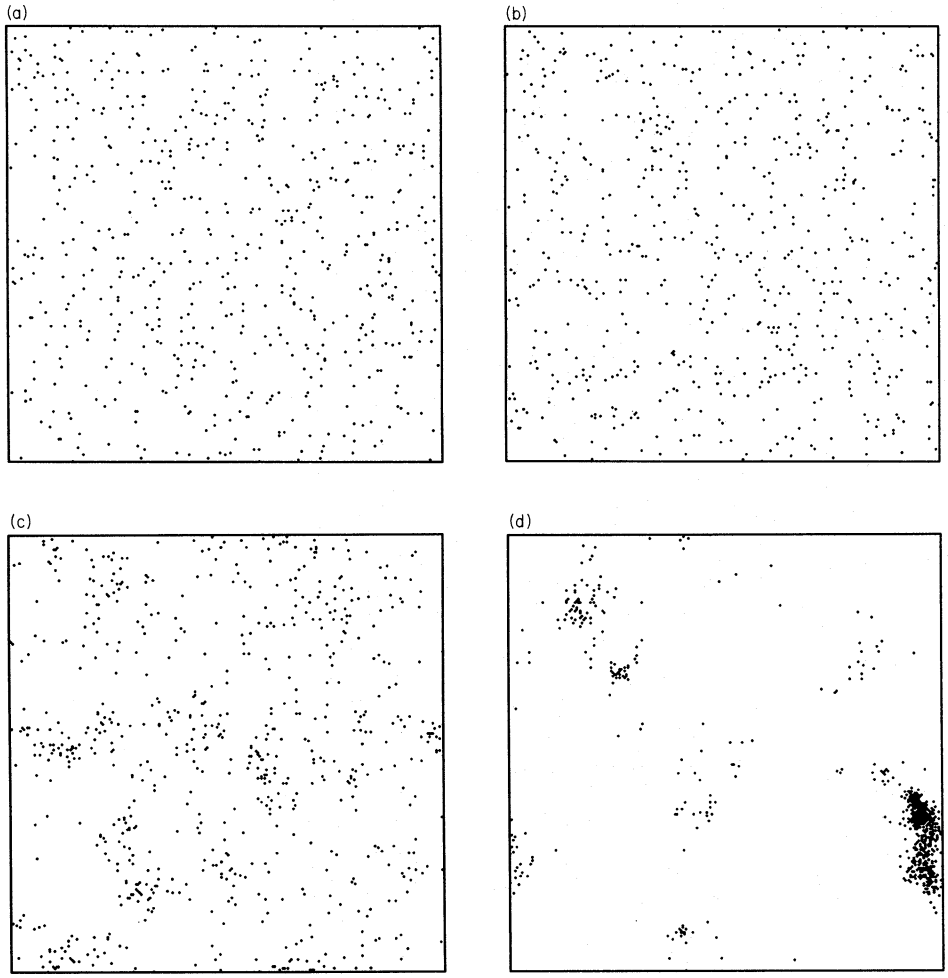

Figure 3. Simulations of clustered point patterns generated from lognormal random fields using the method described in Section 8.3. The underlying Gaussian fluctuations have variances (a) 0.01, (b) 0.1, (c) 1.0 and (d) 10.0 so these maps have one-point distributions like those shown in Fig. 1. They all possess the same shape correlation function but the amplitude is scaled according to the variance of the generated process. The correlation function is $\xi(r) \propto (1 + r^2/2r_c^2)^{-1}$, as used by Coles (1988), with r_c approximately 3 per cent of the map edge. Each is laid down on a 256×256 grid by an FFT technique as discussed in the text. The number of points varies slightly from simulation to simulation, but is about 650 in each of the cases considered. The development of clustering is clearly visible.

characterizing a point data set, such as that produced by the simulations of Section 8.3, although one could, in principle, apply it to smoothed density fields. There are many ways of defining the multifractal behaviour, all interrelated. We shall concentrate here on the function $\tau(q)$ which is defined in terms of the moments of the occupation probabilities of cells of vanishingly small size s:

$$\tau(q) = \lim_{s \to 0} \frac{\log \Sigma [p_i(s)]^q}{\log s}. \qquad (57)$$

When one is dealing with a point set, one cannot let ε tend to zero because the cell-counts then become discreteness-dominated. It is easy to see how this happens. The occupation probabilities will be given by an expression of the form (55). The value of β will just be proportional to the cell volume, s^3, and the ratio in (57) can be replaced by a derivative using L'Hopital's rule. As $s \to 0$, $p_i \sim s^3$. The result is a linear dependence of $\tau(q)$ on q characteristic of the scale-independent purely fractal behaviour exemplified by Poisson noise.

Fortunately, the behaviour of (57) at finite (but small) s seems also to be interesting (Jones *et al.* 1988). The multifractal behaviour can be calculated by noting that, not only does β scale with s, but so does the variance of the Gaussian field which generates the lognormal, previously denoted ε, so in general the pure fractal behaviour is broken. We find, for small s, that a term proportional to $q(q-1)h^2(t)d\sigma_\varepsilon^2/ds$ is added to the linear expression obtained for the Poisson-dominated case above [$h(t)$ is the time-scaling factor in (22)]. The homogeneous scaling behaviour is therefore broken and the departure is greater at late times and for large $|q|$. In other words, both the deep 'voids' (described by negative q) and the high 'peaks' (positive q) show the greatest departure. Note also that, when $q = 0$ or $q = 1$, $\tau(q)$ is unchanged from the Poisson scale-invariant case. This means that our exponential transformation leaves unchanged the Haussdorf dimension D_0 and the information dimension D_1 (using the

notation of Jones *et al.* 1988). We can estimate $d\sigma_\varepsilon^2/ds$ by modelling the 'binning' process as equivalent to smoothing the ε-field on a scale s. The power spectrum of ε [the Fourier transform of $\Xi(r)$] is thus, e.g. for Gaussian smoothing,

$$P_\varepsilon(k, s) = P_\varepsilon(k) \exp(-k^2 s^2)$$

so that

$$\frac{d\sigma_\varepsilon^2}{ds} \sim -2s \int_0^\infty P(k) k^4 \exp(-k^2 s^2)\, dk, \qquad (58)$$

and the manner of the multifractal scaling of the LN model depends on the power-spectrum of the underlying Gaussian field.

It is worth remarking here that one has to be careful in applying this sort of statistic in practice because, when one smooths the field on a scale of order the coherence length of the fluctuations, one changes the underlying statistics. If one smooths on a large scale the statistics of the smoothed field will approach the Gaussian form by virtue of the central limit theorem. One must tread a narrow line, therefore, between discreteness domination on the one hand and smoothing domination on the other. Further expertise in using multifractal techniques will be gained by using them on realizations of LN fields and we shall return to this in future work. In particular, such models can help us determine the best way of extracting the $\tau(q)$ from a data set. Bearing in mind the difficulties with the other statistics we have uncovered and the fact that so much useful information is related to the multifractal function $\tau(q)$ (Jones *et al.* 1988), we consider this to be an important avenue for exploration.

9 CONCLUSIONS

We have studied the properties of a particular non-Gaussian random field, the lognormal random field, and found several reasons (both statistical and kinematical) why it should be an interesting model to study with regard to the statistics of large-scale galaxy clustering.

The model's close relationship to a Gaussian random field (the general LN field includes the Gaussian case in the limit of zero rms fluctuation) allows to perform a complete statistical decomposition and calculate analytically how the usual statistical quantities used in analyses of galaxy catalogues would behave if the density fluctuations were as given by the model. The most important points of this analysis are the following.

(i) It is by no means obvious that the actual density distribution is such that it can be specified by all its moments. The LN distribution is in no way 'pathological' but does possess this disturbing property.

(ii) The 'Void Probability Function' cannot be expressed as a moment expansion in this model since the moment generating function does not exist. This function therefore contains more information than can be contained in the set of moments of the distribution. This is intimately related to (i).

(iii) A calculation of the higher order correlation functions for the model demonstrates that the usual parameterization of higher-order correlations in terms of Q and R, although possibly useful in the strongly non-linear regime, will not be useful for discriminating between Gaussian and non-Gaussian fluctuations in general. Alternative parameterizations are presented and discussed.

(iv) The model allows an heuristic treatment of the steepening of the two-point correlation function in the non-linear regime and the generation of higher-order correlations from initially Gaussian perturbations.

(v) A connection can be made between our model for the non-linear density field and the statistics of the linear velocity field. This allows us to determine the domain within which linear theory can simultaneously apply to both density and velocity.

(vi) The model reveals that current procedures for treating the 'topology' of the regions where the density exceeds a given threshold do not make the most of the information available from the topological measures used. We have proposed a new procedure.

(vii) We have outlined a new 'paradigm' for the simulation of clustered point data sets as an alternative to traditional methods based on Poisson clusters and given some examples of the results obtained.

(viii) The multifractal approach can provide important information about departures from purely fractal scaling behaviour. Simulations are needed to determine the computational procedures that best utilize the information contained in such measures.

It is possible to regard the LN model under different lights. To some it will be a plausible model for the large-scale distribution of matter that can be tested against observations. We have shown how such a test could be performed and will return to this question in later work when better clustering data are available. To others, more sceptical perhaps, the model is just a 'toy' model which has no real physical justification but which illustrates some of the likely pitfalls of currently popular statistical approaches when we try to extend them into the weak clustering regime, where there are currently very poor data. Toys such as this one (and the others mentioned in the introduction) should be taken seriously; like the best children's games, they are educational as well as entertaining.

ACKNOWLEDGMENTS

PC gratefully thanks NORDITA for their hospitality and generosity during the two visits during which this work was completed. We have benefited from discussions with John Barrow, Sabino Matarrese and Adrian Melott. PC is an SERC postdoctoral research fellow.

REFERENCES

Adler, R. J., 1981. *The Geometry of Random Fields,* John Wiley & Sons, New York.

Aitchison, J. & Brown, J. A. C., 1969. *The Lognormal Distribution with special reference to its uses in economics,* Cambridge University Press, Cambridge.

Bardeen, J. M., Bond, J. R., Kaiser, N. & Szalay, A. S., 1986. *Astrophys. J.,* **304,** 15 (BBKS).

Barrow, J. D. & Morgan, J., 1983. *Mon. Not. R. astr. Soc.,* **203,** 393.

Barrow, J. D. & Bhavsar, S. P., 1987. *Q. Jl R. astr. Soc.,* **28,** 109.

Barrow, J. D. & Coles, P., 1990. *Mon. Not. R. astr. Soc.*, **244,** 188.

Bond, J. R. & Couchman, H. M. P., 1988. In: *Proceedings of the Second Canadian Conference on General Relativity and Relativistic Astrophysics,* eds Coley, A. & Dyer, C., World Scientific, Singapore.

Bonometto, S. A., Lucchin, F. & Matarrese, S., 1987. *Astrophys. J.*, **323,** 19.

Borgani, S. & Bonometto, S. A., 1990. *Astrophys. J.*, **348,** 398.

Carruthers, P. & Shih, C. C., 1983. *Phys. Lett.*, **B127,** 242.

Cartwright, D. E. & Longuet-Higgins, M. S., 1956. *Proc. R. Soc. A.*, **237,** 212.

Catelan, P., Lucchin, F. & Matarrese, S., 1988. *Phys. Rev. Lett.*, **61,** 273.

Coles, P., 1986. *Mon. Not. R. astr. Soc.*, **222,** 9P.

Coles, P., 1988. *Mon. Not. R. astr. Soc.*, **234,** 509.

Coles, P., 1989. *Mon. Not. R. astr. Soc.*, **238,** 319.

Coles, P., 1990. *Mon. Not. R. astr. Soc.*, **243,** 171.

Coles, P. & Barrow, J. D., 1987. *Mon. Not. R. astr. Soc.*, **228,** 407.

Collins, C. A., Joseph, R. D. & Robertson, N. A., 1986. *Nature,* **320,** 506.

Couchman, H. M. P., 1987a. *Mon. Not. R. astr. Soc.*, **225,** 777.

Couchman, H. M. P., 1987b. *Mon. Not. R. astr. Soc.*, **225,** 795.

Crow, E. L. & Shimizu, K. (eds), 1988. *Lognormal Distributions: Theory and Applications,* Marcel Dekker Inc., New York.

Davis, M. & Peebles, P. J. E., 1977. *Astrophys. J. Suppl.*, **34,** 425.

Doob, J. L., 1953. *Stochastic Processes,* John Wiley & Sons, New York.

Efstathiou, G., Davis, M., Frenk, C. S. & White, S. D. M., 1985. *Astrophys. J. Suppl.*, **57,** 241.

Fry, J. N., 1985. *Astrophys. J.*, **289,** 10.

Fry, J. N., 1986. *Astrophys. J.*, **308,** L71.

Fry, J. N. & Peebles, P. J. E., 1978. *Astrophys. J.*, **221,** 19.

Górski, K. M., 1988. *Astrophys. J.*, **332,** L7.

Górski, K. M., Davis, M., Strauss, M. A., White, S. D. M. & Yahil, A., 1989. *Astrophys. J.*, **344,** 1.

Gott, J. R., Melott, A. L. & Dickinson, M., 1986. *Astrophys. J.*, **306,** 341.

Gott, J. R., Weinberg, D. H. & Melott, A. L., 1987. *Astrophys. J.*, **319,** 1.

Groth, E. J. & Peebles, P. J. E., 1977. *Astrophys. J.*, **217,** 385.

Gurbatov, S. N., Saichev, A. I. & Shandarin, S. F., 1989. *Mon. Not. R. astr. Soc.*, **236,** 385.

Hale-Sutton, D., Fong, R., Metcalfe, N. & Shanks, T., 1989. *Mon. Not. R. astr. Soc.*, **237,** 569.

Hamilton, A. J. S., 1988. *Astrophys. J.*, **332,** 67.

Hamilton, A. J. S., Gott, J. R. & Weinberg, D. H., 1987. *Astrophys. J.*, **309,** 1.

Hubble, E., 1934. *Astrophys. J.*, **79,** 8.

Jensen, L. G. & Szalazy, A. S., 1986. *Astrophys. J.*, **305,** L5.

Johnson, N. L. & Kotz, S., 1970. *Distributions in Statistics: Discrete Distributions,* John Wiley & Sons, New York.

Jones, B. J. T., Martinez, V. J., Saar, E. & Einasto, J., 1988. *Astrophys. J.*, **332,** L1.

Juszkiewicz, R., Sonoda, D. H. & Barrow, J. D., 1984. *Mon. Not. R. astr. Soc.*, **209,** 139.

Kaiser, N., 1984a. *Astrophys. J.*, **284,** L9.

Kaiser, N., 1984b. In: *Inner Space/Outer Space,* p. 258, eds Kolb, E. W., Turner, M. S., Olive, K., Seckel, D. & Lindley, D., University of Chicago Press, Chicago.

Kaiser, N. & Davis, M., 1985. *Astrophys. J.*, **297,** 365.

Kendall, M. & Stuart, A., 1977. *The Advanced Theory of Statistics,* Volume 1, 4th edn, Griffin & Co., London.

Kirkwood, J. C., 1935. *J. Chem. Phys.*, **3,** 300.

Kofman, L. A. & Shandarin, S. F., 1988. *Nature,* **334,** 129.

Kofman, L. A., Pogosyan, D. & Shandarin, S. P., 1990. *Mon. Not. R. astr. Soc.*, **242,** 200.

Longuet-Higgins, M. S., 1957. *Phil. Trans. R. Soc. Lond. A.*, **249,** 321.

Lucchin, F. & Matarrese, S., 1988. *Astrophys. J.*, **330,** 535.

Lumsden, S. L., Heavens, A. F. & Peacock, J. A., 1989. *Mon. Not. R. astr. Soc.*, **238,** 293.

Lynden-Bell, D., Faber, S. M., Burstein, D., Davies, R. L., Dressler, A., Terlevich, R. J. & Wegner, G., 1988. *Astrophys. J.*, **326,** 19.

Matarrese, S., Lucchin, F. & Bonometto, S. A., 1986. *Astrophys. J.*, **310,** L21.

Melott, A. L. & Fry, J. N., 1986. *Astrophys. J.*, **305,** 1.

Mellot, A. L., Weinberg, D. H. & Gott, J. R., 1988. *Astrophys. J.*, **328,** 50.

Messina, A., Moscardini, L., Lucchin, F. & Matarrese, S., 1990. *Mon. Not. R. astr. Soc.*, **245,** 244.

Neyman, J. & Scott, E. L., 1952. *Astrophys. J.*, **116,** 144.

Ostriker, J. P., 1986. In: *Galaxy Distances and Deviations from Universal Expansion,* eds Madore, B. F. & Tully, R. B., p. 273, Reidel, Dordrecht.

Peacock, J. A. & Heavens, A. F., 1985. *Mon. Not. R. astr. Soc.*, **217,** 805.

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

Peebles, P. J. E. & Groth, E. J., 1975. *Astrophys. J.*, **196,** 1.

Politzer, H. D. & Wise, M. B., 1984. *Astrophys. J.*, **310,** L21.

Rice, S. O., 1945. *Bell Systems Tech. J.*, **24,** 46.

Saslaw, W. C. & Hamilton, A. J. S., 1984. *Astrophys. J.*, **276,** 13.

Schaeffer, R., 1985. *Astr. Astrophys.*, **144,** L1.

Shandarin, S. F. & Zel'dovich, Ya. B., 1989. *Rev. Mod. Phys.*, **61,** 185.

Shanks, T., 1979. *Mon. Not. R. astr. Soc.*, **186,** 583.

Shanks, T., Hale-Sutton, D., Fong, R. & Metcalfe, N., 1989. *Mon. Not. R. astr. Soc.*, **237,** 589.

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

Soneira, R. M. & Peebles, P. J. E., 1978. *Astr. J.*, **83,** 845.

Szalay, A. S., 1988. *Astrophys. J.*, **333,** 21.

Tóth, G., Hollósi, J. & Szalay, A. S., 1989. *Astrophys. J.*, **344,** 75.

Vanmarcke, E. H., 1983. *Random Fields: Analysis and Synthesis,* MIT Press, Cambridge, Massachusetts.

Wax, N. (ed), 1954. *Selected Papers on Noise and Stochastic Processes,* Dover, New York.

Weinberg, D. H., Gott, J. R. & Melott, A. L., 1987. *Astrophys. J.*, **319,** 1.

White, S. D. M., 1979. *Mon. Not. R. astr. Soc.*, **186,** 145.

Zel'dovich, Ya. B., 1970. *Astr. Astrophys.*, **5,** 84.

Zinnecker, H., 1984. *Mon. Not. R. astr. Soc.*, **210,** 43.