PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Measuring large–scale structure from redshift surveys

Alexander S. Szalay

Phil. Trans. R. Soc. Lond. A 1999 **357**, 117-132
doi: 10.1098/rsta.1999.0318

THE ROYAL SOCIETY

Measuring large-scale structure from redshift surveys

By Alexander S. Szalay

Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218, USA

Observations of the large-scale distributions of galaxies in the universe indicate the presence of very large wall-like superclusters. In their distributions there is growing evidence that there is a characteristic scale in excess of 100 h^{-1} Mpc. This scale is rather interesting since it is too small to be well measured from fluctuations in the cosmic microwave background, but at the same time is large enough to be easily sampled in current redshift surveys. There are several physical processes in the universe at around recombination (equality of matter and radiation, the sound horizon, the recombination) that may have left an imprint on the fluctuation spectrum even in the galaxy distribution. We discuss how ongoing large-scale redshift surveys may be optimally analysed, via the Karhunen–Loève method, to provide high-precision information on the fluctuations on these interesting scales. These large surveys are the first where the dominant source of noise is systematic errors, requiring novel techniques of statistical analysis.

Keywords: redshift surveys; large-scale structure; cosmology; power spectrum

1. Introduction

The study of large-scale structure is one of the most dynamically evolving areas of astrophysics today. Cosmology and large-scale structure is growing into an accurate science and requires correspondingly more sophisticated methods of analysis. Twenty years ago the estimates of the fluctuation amplitude were about 10^{-3}, almost a factor of a hundred off of today's measurements. Ten years ago we could only hope for high-precision measurements of large-scale structure—there were less than 5000 redshifts measured—and only a handful of normal galaxies with $z > 1$ were known. Computer models of structure formation had just begun to consider non-power-law spectra based on physical models like hot/cold dark matter. As a consequence there was considerable freedom in adjusting parameters in the various galaxy-formation scenarios. In contrast, many of today's debates are about factors of two, and soon we will be arguing about 10% differences. The shape of the primordial fluctuation spectrum, first derived from philosophical arguments (Harrison 1970; Zel'dovich 1972) can now be quantified from detections of fluctuations in the cosmic microwave background (CMB) made by COBE (Smoot *et al.* 1992). The number of available redshifts is beyond 50 000, and soon we will have redshift surveys surpassing one million galaxies. N-body simulations are becoming more sophisticated, of higher resolution, and incorporating complex gas dynamics.

The unprecedented number of new observations currently under way gives us hope that over the next decade we will gain a clear understanding of the shape and evolution of the primordial fluctuation spectrum, understand from first principles how galaxies were formed, and make quantitative comparisons and tests to differentiate the various galaxy-formation scenarios. Up until now the CMB experiments have measured fluctuations mostly on scales above 300 Mpc, where the shape is expected to be primordial. Galaxy surveys were mostly sensitive to the regime below 100 Mpc, where strong clustering evolution does leave a significant imprint on the spectrum. *The least known, and at the same time the most interesting part* of the fluctuation spectrum is on scales between 100 and 300 Mpc, close to the horizon scale at equality and recombination, where we have the most reasons to believe that something may have left a detectable imprint, like the Doppler peaks in the CMB fluctuations. In the near future there is a good chance that this regime will be well studied by both CMB experiments and redshift surveys.

2. Quantifying large-scale structure

(*a*) *Key questions*

Structure in the universe evolves from the initially small primordial fluctuations. These fluctuations can arise during an inflationary expansion or come from topological defects later. They grow in amplitude, due to gravitational instability, and the shape of the fluctuation spectrum is altered by different physical processes. The nature of the dark matter, whether hot or cold, believed to dominate the mass density of the universe, determines the shape of the power spectrum on small (less than 100 Mpc) scales. On the other hand, the shape of the large-scale part of the fluctuations (greater than 300 Mpc) remains remarkably unchanged, because no scale in the evolutionary process becomes this large.

The COBE measurements constrain both the amplitude and the initial spectrum of the fluctuations in this regime, and demonstrate extremely good agreement ($n = 1.1 \pm 0.4$, Gorski *et al*. (1994)) with the Harrison–Zel'dovich predictions of $P(k) = k^n$, with $n = 1$. These fluctuations are due to differences in the gravitational potential at the surface of last scattering (Sachs & Wolfe 1967), reflecting the state of the universe at a redshift of *ca.* 1000. Galaxy surveys (at $z < 0.3$) are rapidly increasing in size, thus providing increasingly better measurements of the fluctuations on small scales (CfA slices (Geller & Huchra 1989); IRAS (Saunders *et al*. 1991); APM (Maddox *et al*. 1990); APM redshift surveys (Loveday *et al*. 1992); LasCampanas (Shectman *et al*. 1996)). One can use theoretical scenarios to evolve and extrapolate the large-scale CMB measurements into the structure of the local universe, but the two regimes do not yet overlap directly.

Currently, the most popular scenario is the cold dark matter dominated universe, where most of the mass is dark, interacting only via gravity, consisting of particles of such a large mass that their thermal motion is negligible. To match the observed clustering of galaxies without producing too large a velocity dispersion, the concept of 'biasing' has been invoked (Kaiser 1984; Bardeen *et al*. 1986): mass is converted into light only at the densest regions in the universe, creating a luminous component more clustered than the mass. This scenario, modulo a properly chosen initial normalization, has been remarkably successful over the past 12 years.

The COBE measurements create a conflict with the minimal biased CDM model: if a Harrison–Zel'dovich spectrum is assumed and the normalization is locked to COBE, then the biasing parameter must be unity to match the small-scale part of the fluctuation spectrum, leading to very large small-scale velocities. Several alternative models have been rapidly suggested. Gravity waves, which decay with time, may contribute to the largest scale modes observed by COBE and produce a 'tilt' of the spectrum (Davis *et al*. 1992). Alternative scenarios invoke a large cosmological constant (Kofman *et al*. 1993) or a Hubble constant as low as 30 km s^{-1} Mpc^{-1} (Bartlett *et al*. 1995). A mixture of cold and hot dark matter would also help, because the growth of fluctuations on small scales would be retarded due to the presence of a hot component with $\Omega_\nu \approx 0.2$ (Klypin *et al*. 1993).

(*b*) *Power on COBE scales*

What are the most important measurements we can make in order to differentiate between proposed models? Overlap between scales probed by CMB experiments and redshift surveys in the 'local' universe would place strong constraints on the power spectrum. While the CMB experiments measure the fluctuations at the surface of last scattering, redshift surveys measure the fluctuations today; thus their combination is sensitive to the total fluctuation growth since recombination. The power spectrum on scales of 200–500 Mpc from both redshift surveys and CMB would also tell us whether the gravity wave/tilted model is correct, measure the bias factor, and determine the shape of the spectrum on scales where most of today's models differ but which are too small for COBE and beyond the scale of current galaxy measurements. For the same reason, many CMB experiments are probing 1–2° scales, corresponding to a co-moving scale of about 120 Mpc.

In the next sections we outline how novel statistical techniques that we are currently developing will bring this goal within reach, using galaxy redshift surveys. The combination of the two types of measurements will reveal unprecedented details about the fluctuation spectrum. However, current power-spectrum estimation techniques are optimized to compensate for shot noise, while the dominant sources of errors at this point will no longer be statistical, but rather dominated by systematics. Our proposed power-spectrum analysis technique has been designed with this in mind.

(*c*) *Observing walls*

Several surveys have now found evidence for sharp wall-like structures in the universe. The existence of such features is by no means unexpected. Zel'dovich (1970) predicted that the generic features in a pressure-free gravitational collapse will be highly flattened 'pancakes'. Observational confirmation took a few years, Chincarini & Rood (1976), and Gregory & Thompson (1978) identified the excess of galaxies between Coma and A1367 with a supercluster, resembling a 'pancake'.

Kirshner *et al*. (1983) identified the first big 'void' in the galaxy distribution. A major breakthrough in our understanding of large-scale structure came from the CfA 'slice' by deLapparent *et al*. (1986), 6° wide in declination but over 100° in right ascension. At a radial distance of 70 h^{-1} Mpc a distinct pattern appears: a 'great wall' containing hundreds of galaxies, connecting several of the known Abell clusters. Its transverse spatial extent exceeds 100 by 50 h^{-1} Mpc. The general trend has been summarized by Geller & Huchra (1989): *'all surveys have detected structures as large as they could . . . .'*

If the universe were full of 'great walls', i.e. if they are typical of the very-large-scale structure, one can get an estimate of what a 'fair sample' would consist of

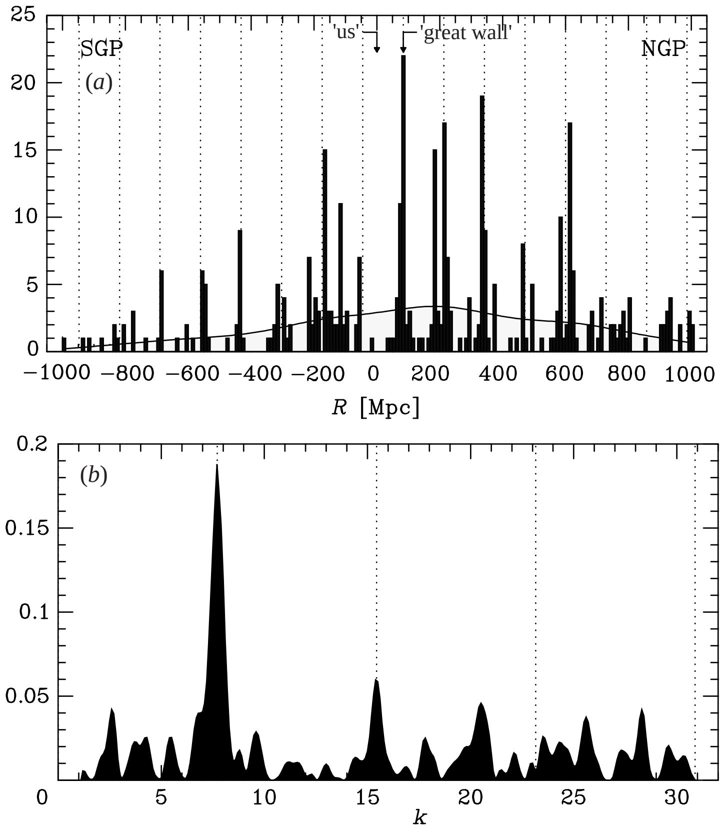

Figure 1. The redshift distribution of galaxies in the BEKS survey, comprised of two narrow pencil beams towards the galactic poles: (a) the histogram of all galaxies; (b) the one-dimensional power spectrum ($P(k) = |\delta(k)|^2$, $r = 1000/k$ Mpc. The big spike corresponds to the co-moving scale of 128 h^{-1} Mpc. Wavenumbers are in units of $k = 1000/\lambda\, h^{-1}$ Mpc^{-1}.

from the surface density of galaxies. If we assume that the fraction of all bright galaxies, *ca.* $f = 0.5$, is on these surfaces, with the surface density of galaxies as $\mu = 0.4h^2$ Mpc^{-2}, we can estimate the characteristic 'cell' size by requiring that the corresponding 'local' volume density of bright galaxies, $n = 0.02$ galaxies h^3 Mpc^{-3}, be approximately reproduced. Assuming spherical bubbles, and counting only half of the surface area, since the walls separate two volumes, the typical size of the voids is $\lambda = 2R = 3\mu/nf = 120\, h^{-1}$ Mpc. This gives us some idea of what cell sizes one can expect in a universe dominated by 'great walls', derived solely from the observations.

Broadhurst *et al.* (1990, hereafter BEKS) published results from a redshift survey in two opposite pencil beams. The angular diameter of the survey is $30'$ and the depth is about 0.5 in redshift, both at the North and South Galactic Poles. The combined surveys have a joint length in excess of $2000\, h^{-1}$ Mpc, considerably deeper than any other previous survey. To compensate for the small physical size of the survey at low redshifts, data from two bright surveys in almost the same directions were used, resulting in a combined selection well approximated by a cylinder of constant co-moving radius.

The northern pencil beam is in the CfA slice, and one can find the 'great wall' without much difficulty. Surprisingly, however, at very large radial distances one still cannot see a homogeneous distribution, rather most galaxies are in a few large 'spikes' along the line of sight, separated typically by more than 100 h^{-1} Mpc. The simplest explanation was that further 'walls' were found, meaning that the 'great wall' is by no means unique, and that these structures contain a large percentage (*ca.* 50%) of the galaxies.

(*d*) *Observing bumps*

Even more surprising was the fact that in the one-dimensional Fourier transform of the redshift distribution, a highly significant peak was found at the wavelength of 128 h^{-1} Mpc, with a probability of $P < 3 \times 10^{-4}$. This observation prompted many debates, and even more exotic theories. The main question was, of course, whether the peak in the Fourier spectrum is just a random accident, or does this scale arise as a result of a physical process? Extending the BEKS survey to nine pencil beams, randomly distributed over a $6^\circ \times 6^\circ$ region at both galactic poles, it was shown that the cross-correlation signal stays strong up to about 60 h^{-1} Mpc transverse separation (Broadhurst *et al.* 1995), indicating that the redshift spikes are indeed 'great wall'-like structures, also that the power spectrum peak was not due to a random alignment of small groups.

Several years later bigger redshift surveys became available. The Las Campanas survey (Shectman *et al.* 1996), consisting of six slices of 450 h^{-1} Mpc depth, found evidence for statistically significant excess power on 100 h^{-1} Mpc scales (Landy *et al.* 1996). A similar slice near the South Galactic Pole, the ESP project (Vettolani *et al.* 1997), confirms the BEKS spikes in the overlap region. A recent survey of the great attractor/Shapley concentration shows structure on 100 h^{-1} Mpc scales (Proust 1999). Deeper surveys on the Keck telescope (Cohen *et al.* 1996), and the CFRS (Lilly *et al.* 1995), found evidence for the existence of sharp walls at $z = 1$. In the distribution of clusters, Tully *et al.* (1992), Guzzo *et al.* (1992) and recently Einasto *et al.* (1997) have found a signature of a sharp spectral feature beyond 100 Mpc. Excess power on greater than 100 h^{-1} Mpc scales is present at even higher redshifts in QSO absorption systems (Quashnock *et al.* 1996) and in galaxies (Steidel *et al.* 1999). These high-redshift observations are extremely important—if the bumps appear on the same co-moving scale at much earlier Hubble times, then there is a built-in feature in the power spectrum! In any case, it is obvious now that there is excess power just beyond 100 Mpc, but it is not clear what the precise shape of this feature in the power spectrum is.

3. Measuring structure beyond 100 Mpc

(*a*) *Power spectrum*

To estimate the power spectrum from a galaxy redshift survey, we must take into account the sampling density (determined by the magnitude limit) and geometry of the survey (determined by the angular coverage and depth). The sampling process, and the fact that only integer numbers of galaxies can be counted, introduces shot noise into the power spectrum (the noise per mode is constant and thus easily subtracted, but contributes to the uncertainty). The observed power spectrum is a convolution of the true power with the Fourier transform of the spa-

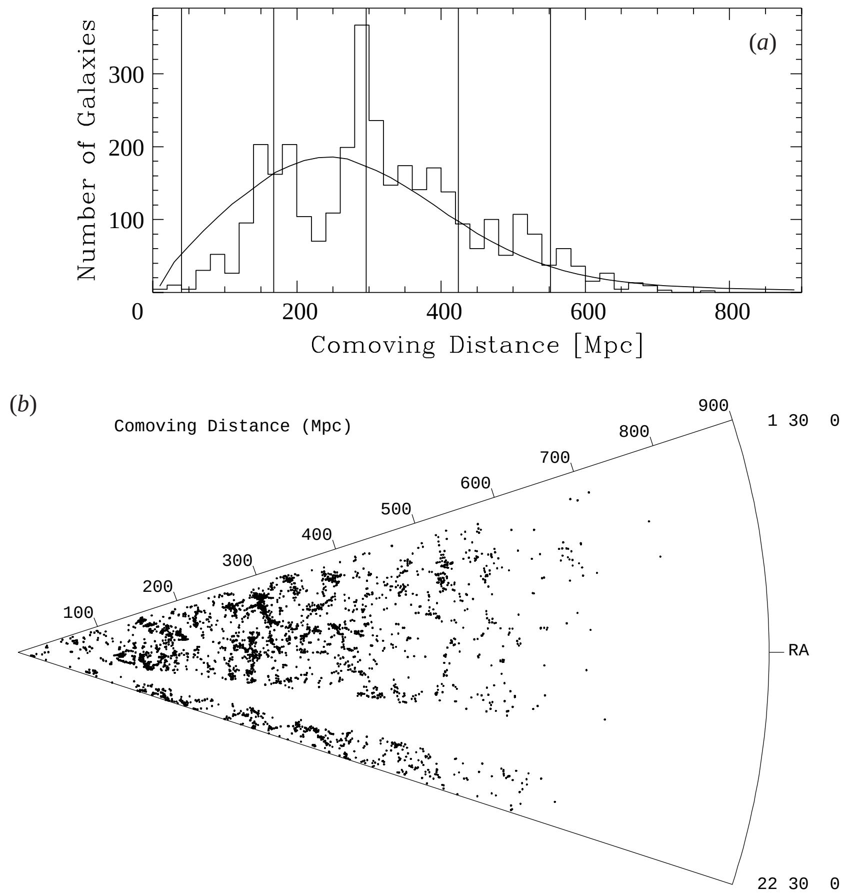

Figure 2. The redshift distribution of galaxies in the ESP survey, consisting of a slice at $\delta = -30°$. (a) The histogram of all galaxies. The vertical lines indicate the location of the BEKS spikes. (b) A wedge diagram of the actual redshifts.

tial window function of the survey ($W(\boldsymbol{x}) = 1$ inside the survey and 0 outside), $P_{\mathrm{obs}}(\boldsymbol{k}) = \int P_{\mathrm{true}}(\boldsymbol{k}')|W(\boldsymbol{k}-\boldsymbol{k}')|^2\,\mathrm{d}^3k'$. One can attempt to deconvolve the true power spectrum or compare it with convolved theoretical spectra, but in either case the survey geometry limits both the resolution and the largest wavelength for which an accurate measurement can be obtained.

The standard methods for power spectrum estimation (see, for example, Park *et al*. 1994; Feldman *et al*. 1994; Fisher *et al*. 1993) work reasonably well for data in a large contiguous three-dimensional volume with homogeneous sampling of the galaxy distribution. The weighting scheme is optimized for shot-noise-dominated errors. Using these techniques, nearby wide-angle redshift surveys (CfA, SSRS, IRAS 1.2, QDOT) yield strong constraints on the power spectrum on scales up to 100 h^{-1} Mpc. Because the uncertainty in the power spectrum depends on the number of independent cells of a given wavelength that we sample, constraints on larger scales require deeper surveys. Due to the difficulty of obtaining redshifts for fainter galaxies and limited

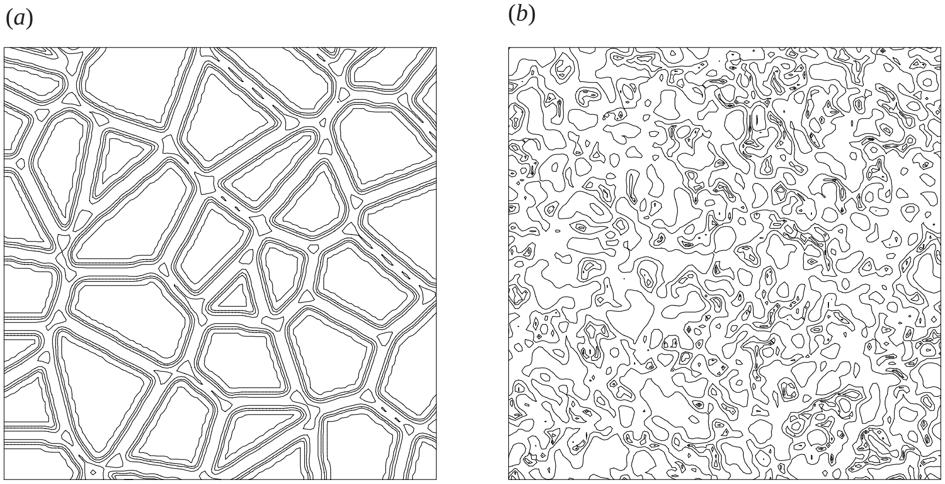

Figure 3. Two simple realizations of a two-dimensional universe with identical second-order statistical properties: (*a*) a two-dimensional Voronoi foam generated by the median surfaces between Poisson 'seeds' at a mean separation of 100 Mpc. In this simple toy model, galaxies reside only on the walls of the smoothed foam, so the walls have a finite thickness. The structure has a well-defined second-order statistic, but also has well-correlated phases. This picture has been Fourier transformed, all the phases randomized, then transformed back again. The result is shown in (*b*) with the same second-order properties, but with a Gaussian distribution. It is easy to see that placing well-sampled pencil beams across both surveys will easily distinguish between the two, whereas a sparse sample drawn from the two realizations cannot differentiate.

telescope time, deep redshift surveys typically have complex geometry, e.g. deep pencil beams or slices.

However, the standard methods are not efficient when applied to data in oddly shaped and/or disjoint volumes, or when the sampling density of galaxies varies greatly over these regions. Systematic effects, like extinction or calibration zero points can substantially contribute to the errors. Convolution of the true power with the complex window function causes power in different modes to be highly coupled. In other words, plane waves do not form an optimal eigenbasis for expansion of the galaxy density field sampled by redshift surveys. We desire methods for power-spectrum estimation that optimally weight the data in each region of the survey, taking into account our prior knowledge of the nature of the noise and clustering in the galaxy distribution. A detailed comparison of all available power-spectrum estimation methods is given in Tegmark *et al.* (1998). There are several contaminating effects, like aliasing from non-spherical survey geometries, extinction, redshift distortions, nonlinear fluctuation growth, etc. In the section below, we outline how to create an analysis tool that goes considerably beyond the present state of the art, and can take all these effects into consideration.

(*b*) *Walls and sampling effects*

If fluctuations in the universe are strictly Gaussian, their full statistical description is contained in the two-point correlation function or in its Fourier transform, the power spectrum. The phases of the individual Fourier components are random for such a process, and all high-order measures of clustering vanish. Averaging over an infinite number of finite-size realizations, the correct power spectrum is recovered. On the other hand, if there is a network of sharp 'walls' present, they are manifested as a set of sharp 'spikes' in Fourier space. These sharp spikes will vary from realization to realization, and in an ensemble average they will converge to the underlying power spectrum. Even though both scenarios converge to the true power spectrum in the infinite limit, it is much harder to tell from a small number of observations whether the detected sharp Fourier spikes are a genuine part of $P(k)$ or if they are due to the nonlinearities of the walls. The sampling rate will also dramatically affect how well sharp peaks can be measured. This is why well-sampled pencil beams may yield seemingly quite different results for the statistics of power-spectrum amplitudes than wide-angle sparsely sampled surveys. If there are even weak nonlinearities present in the density, their effect on the power spectrum is quite surprising (Amendola 1994): rare 'hotspots', high and narrow peaks, will emerge in every realization of the power spectrum. It is possible that the current physical scale of the k-space bumps does not exactly coincide with the broader peak of the ensemble-averaged power spectrum.

4. The Karhunen–Loève transform

One can find an optimal set of spatial filters to probe the density fluctuations. Rather than directly compute the Fourier transform of the distribution of objects, we expand the observed density field in the natural orthonormal basis determined for each survey from our prior knowledge of the survey geometry, selection function and clustering of galaxies, and find the most likely power-spectrum model in a Bayesian fashion (Vogeley & Szalay 1996). Expansion of the observed density field in this basis is known as the Karhunen–Loève (KL) transform (see, for example, Therrien 1992). Dividing the survey volume into cells V_i, we compute the correlation matrix of expected counts as

$$C_{ij} = \langle N_i N_j \rangle = \langle N_i \rangle \langle N_j \rangle (1 + \langle \xi_{ij} \rangle) + \delta_{ij} \langle N_i \rangle + \eta_{ij},$$

where $\delta_{ij} = 0$ for $i \neq j$, N_i is the galaxy count in the ith cell, η_{ij} is additional noise arising from systematic effects and

$$\langle \xi_{ij} \rangle = \frac{1}{V_i V_j} \int \xi(\boldsymbol{x}_i - \boldsymbol{x}_j)\, \mathrm{d}V_i\, \mathrm{d}V_j .$$

We compute ξ from a model, which is our null hypothesis. The eigenvectors $\boldsymbol{\Psi}_j$ that diagonalize the correlation matrix are the signal-to-noise eigenfunctions of the density field of the survey (solving the equation $\boldsymbol{C} \cdot \boldsymbol{\Psi}_j = \lambda_j \boldsymbol{\Psi}_j$). The eigenvectors have a very simple physical meaning: they contain the optimal weight of a given cell associated with each mode. This weight—via the matrix diagonalization—automatically considers all the different sources of errors, incorporated in the shot-noise term, and η, and the asymmetric geometry of the survey, then computes the optimal weight for each cell and each mode. The eigenvalues represent the statistical information content of the given mode. One can also see that, by ranking the modes by decreasing eigenvalues, the list begins with the modes containing large-scale power. The eigenvectors of larger rank mostly describe shot noise.

We expand the observed counts in this orthonormal basis $N_i = B^j \Psi_{ij}$ (Einstein summation convention), which defines the transform $B^j = \Psi^{ij} N_i$. Sorting these functions by decreasing eigenvalue λ yields the set of eigenfunctions in order of

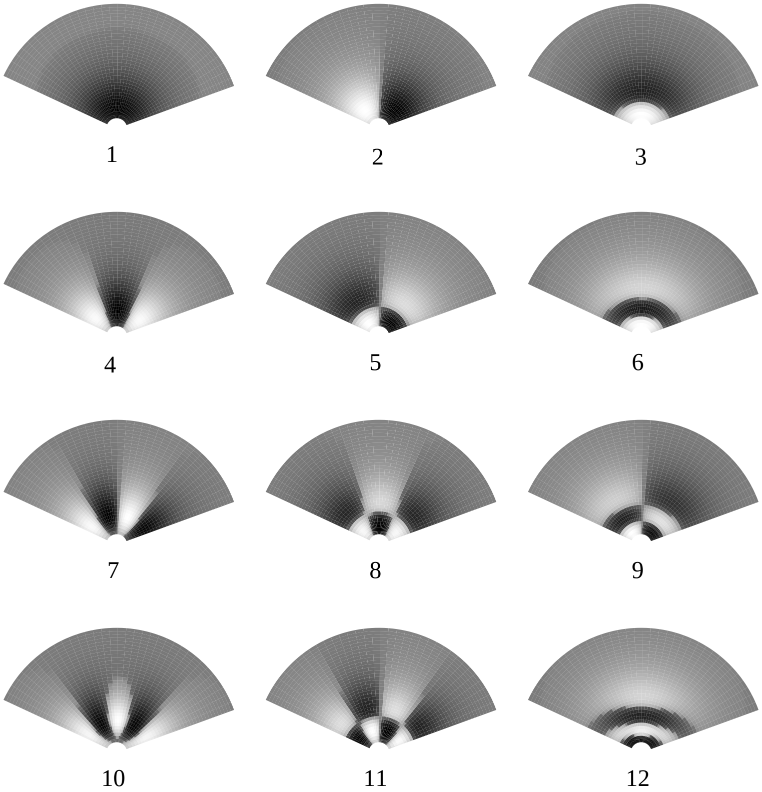

Figure 4. The figure shows the first 12 density eigenmodes for the geometry, selection function and correlation function of the first CfA slice (deLapparent *et al*. 1986). Each of these eigenmodes samples a narrow range of Fourier wavemodes.

decreasing signal to noise. Because the B^j are statistically orthogonal and because we can easily compute the expectation value and variance of the power per eigenmode for any power-spectrum model

$$\langle B_j^2\rangle = \boldsymbol{\Psi}_j^{-1}\cdot\boldsymbol{C}^{\mathrm{model}}\cdot\boldsymbol{\Psi}_j,$$

hypothesis testing is a straightforward process. Note that this method requires an initial guess at the power spectrum, but the form of the eigenfunctions does not depend sensitively on this assumption, and we can easily iterate the process. Summarizing the main features, the KL transform automatically determines the 'correlation eigenmodes' of a complex survey geometry, *each optimally weighted* to measure power on a certain scale. The expansion of the density field in terms of these modes still contains phase information, and the modes are orthogonal and independent, and thus statistical hypothesis testing is quite easy. *What are the problems where further improvements are necessary?* These will be discussed in the next sections, and we will outline the way these aspects can be improved.

(*a*) *Adaptive pixelization*

The method in its present form requires a pixelization, which was assumed to be given. By changing to smaller cells, the resolution is increasing, but at the cost of

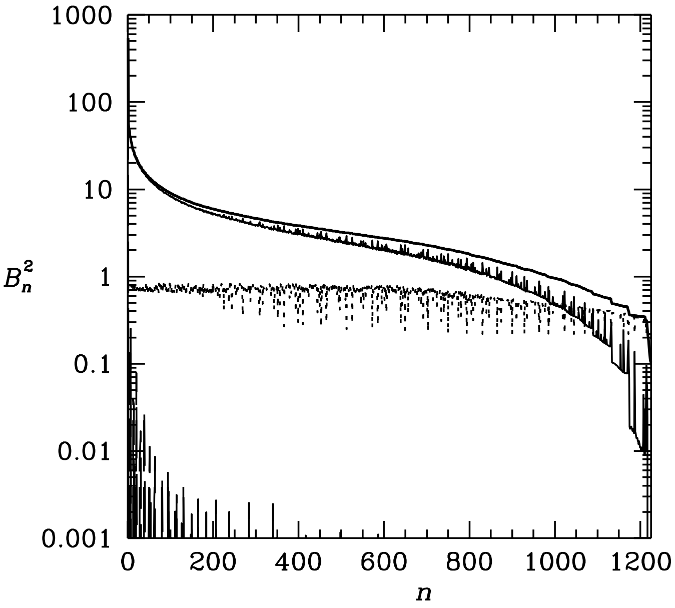

Figure 5. The figure shows the expected power per mode, $\langle B_n^2 \rangle$, of the KL transform analogous to the power spectrum of the Fourier expansion. The highest curve is the total power per mode, the sum of the true clustering power, shot noise, and the mean density. The component at the low wavenumbers is highly suppressed and is the contribution from the mean density, entering through aliasing via the side-lobes of the respective modes. The relatively flat component is the shot-noise contribution; it has a slight variation, since the cell sizes are not quite identical. The remaining component is the cosmological fluctuation spectrum, dominating the signal out to quite large wavenumbers.

a bigger matrix. At the same time figure 5 shows that the largest eigenvalues carry most of the clustering signal; most of the higher-ranked eigenvectors deal with the representation of the shot noise. Let us start out with a coarse grid, and compute the eigenvalues and eigenvectors. Next, we subdivide each cell into two halves. This can be considered as a perturbation on the eigensystem, just like the level splitting of the H atom in an external magnetic field. The resulting eigenvalues and eigenvectors can be computed from perturbation theory. If we are only interested in the first few thousand eigenvectors, we can set an accuracy threshold, beyond which we do not subdivide the cells any further. This threshold should be set to the sum of the first M eigenvalues. This technique will automatically guarantee the coarsest pixelization that is still within our required error bounds. The sensitivity of the eigensystem with respect to splitting individual cells can also be computed this way.

(*b*) *Diagonalizing large matrices*

Since the standard techniques of matrix diagonalization (SVD, Jacobi, Gauss–Seidel) are typically proportional to N^3, where the matrix is $N \times N$, computation times start to become prohibitive beyond matrix sizes of $N > 8000$. On the other hand, there is an algorithm, developed by C. Lanczos in the 1920s, that is widely used in various areas of computational physics, like nuclear physics and QCD to diagonalize matrices several million in size. The technique can compute an approximation to the first M eigenvalues and eigenvectors of large matrices, with a saving of over a factor of a hundred in CPU time. A variant of this technique can also be used at the likelihood computation stage, where we calculate directly the (approximate) scalar product of the data vectors with a hypothesis matrix.

(*c*) *Redshift space distortions*

The pixelization will occur in redshift space, thus the computation of the correlation matrix has to be done in redshift space. On large scales the effects from the thermal motion of galaxies ('fingers of God') are negligible, but the linear infall effects can be considerable. At the same time, the distribution of angles between lines of sight depends on the survey geometry; redshift distortions are greatest in pencil beams, smaller in slices, and even smaller, but non-negligible, in wide-angle surveys. Most of the simple results for redshift distortions have been computed using the plane-parallel approximation (Kaiser 1986; Hamilton 1992), where the lines of sight to the two galaxies are close to each other. In a general wide-angle survey this is not the case, and explicit expressions are needed for the redshift space correlations. A numerical computation has been done by Zaroubi & Hoffman (1996) and recently a simple analytic expression has been obtained by Szalay *et al.* (1999) for cells with an arbitrary angle between the lines of sight, just what is needed here. We will use this expression to compute the KL correlation matrix. Our prior model will incorporate the usual parameter $\beta = \Omega^{0.6}/b$, connecting density perturbations to peculiar velocities. This also means that in our parameter estimation not only the shape of the real-space power spectrum, but also β are simultaneously recovered.

(*d*) *Applying inverse nonlinear corrections*

The power spectrum is affected by the growth of fluctuations due to gravity. For fully linear growth, the shape of the spectrum remains unchanged; only its amplitude varies with time. On the other hand, the amplitude of fluctuations today is such that the scale of nonlinearity $k_{\text{nl}} = 0.125\,h\,\text{Mpc}^{-1}$, only a factor of two away from $k = 2\pi/100\,\text{Mpc} = 0.063\,h\,\text{Mpc}^{-1}$. This means that mildly nonlinear effects will modify the shape of the spectrum even in this regime, given the high accuracy we seek with our method. For a long time only N-body simulations provided a solution. Recently, Jain *et al.* (1995) and Peacock & Dodds (1996) provided a simple semi-analytic expression, using the effective spectral index n_{eff}, to compute the nonlinear shape of the power spectrum. It is straightforward to compute the inverse expressions for the power spectrum, which will enable us to go from the mildly nonlinear estimated spectrum to the linear power spectrum, to be compared with the CMB measurements.

(*e*) *Including systematic effects*

There are various systematic effects, which can easily become the dominant source of error for the next-generation surveys. These include zero-point errors in the photometric calibrations, which are typically done over fields several degrees in size, all the way to every fourth of the 6° plates in the APM survey. A zero-point error causes a correlated shift in every magnitude; thus in certain areas of the sky the survey goes deeper. A similar large angular scale error can be caused by the galactic extinction, which is quite clumpy and can be several tenths of a magnitude. As the selection function starts to fall steeply, these effects modulate the outer edges of the survey, resulting in mock large-scale features in the power spectrum. On scales beyond 200 h^{-1} Mpc this is a huge effect (Vogeley & Connolly 1999). One can compensate by correcting with an extinction map, but still the errors in the galaxy counts in cells at the same part of the sky will be correlated. This can be taken into account by an additional variance to the affected cells in the KL correlation matrix, which will effectively down-weight these cells.

Similarly, constraints from a fixed number of fibres in a given patch of the sky (like in the LCRS) can be considered by using an increased multivariate variance instead of Poisson, for the cells along the same line of sight, since the sum of the galaxies must add up to the total number of fibres! Data from different surveys with differing systematics can also be easily combined into a common analysis. These and many other effects can be taken into account in a simple fashion in the KL framework and are impossible to incorporate into any other algorithm currently available.

(*f*) *Using gappy data*

As redshift surveys are under way, the sky coverage is incomplete, containing gaps, only N' cells out of a total N, even though in the end there will be a contiguous area. Since we are interested in large-scale structure, we can (and will) use a truncated KL basis for our analysis, consisting of a certain number $M < N' < N$ of eigenvectors. Let us assume that we built our KL basis for the whole survey, including also the cells where redshift have not yet been measured (or maybe never will be). In order to determine the expansion coefficients on the truncated basis we need to estimate M numbers, based on N' measurements. This is trivial, and as a result we obtain smooth information about galaxy counts in all N cells, even where no data were available. This represents an optimal extrapolation over gappy areas and also enables us to perform analysis of incomplete redshift surveys while they are in progress.

(*g*) *Parameter estimation: hypothesis testing*

The KL transform represents a useful mapping of the data into a set of orthogonal expansion coefficients, which make hypothesis testing quite convenient. On the other hand, when we try to estimate certain parameters, not all modes contain the same information. For example, in the estimation of β, the radial KL modes will be quite sensitive to the value of β, unlike the transverse modes. How can one create the optimal combination of modes for a particular set of parameters? The Fisher information matrix, the Hessian of the log-likelihood with respect to the parameters of interest, tells us the quantitative relevance of the various modes (Vogeley & Szalay 1996). The practical ways of combining the various KL modes are described in detail in Tegmark *et al.* (1998).

5. Are primordial sound waves the source of the bumps?

Here we would like to discuss how such 100 h^{-1} Mpc bumps can arise in the power spectrum. It has been understood for a long time (Peebles 1968; Sunyaev & Zel'dovich 1970) that around recombination due to the high pressure in the photon–baryon plasma, fluctuations oscillate like sound waves. On smaller wavelengths, these oscillations damp, but on larger scales, near the horizon scale at recombinations, they may survive longer. The motion of baryons due to these sound waves gives rise to the Doppler peaks in the CMB fluctuations. At the same time it was understood early on that after recombination, as the sound speed approaches zero, different sound waves transform into a different mix of growing and decaying modes, depending on the actual phases of the waves. An 'interference pattern' may emerge, the so-called Sakharov oscillations (Sakharov 1966). Sound waves that go entirely into growing mode are amplified, others with opposite phases will cancel (Hu & Sugiyama 1996). Since the horizon scale at recombination is very close to the regime of interest (between 100–200 h^{-1} Mpc, depending on $\Omega_0 h$), it is worth considering what it takes for these sound waves to have an appreciable effect on not only the CMB but on the galaxy distribution!

Since galaxy clustering is only affected by gravity, the fluctuations in the baryons due to the sound waves need to leave an imprint in the gravitational potential. This requires as high a baryon fraction as possible. How high can this number be? From observations of the primordial deuterium (Tytler 1997), the 1σ limit, $\Omega_{\mathrm{B}}\, h^2 \leqslant 0.025$, can be combined with reasonably low estimates of the Hubble constant, and values of $\Omega_{\mathrm{B}} \approx 0.1$ are not unimaginable. At the same time, in order for a large imprint, Ω_{t} has to be low, in the range of $\Omega_{\mathrm{t}} \approx 0.3$. Given the faint number counts of galaxies, this is no longer an outrageous idea.

(*a*) *Linear theory*

Recent linear calculations of Eisenstein & Hu (1998) provide analytic approximations for the shape of the transfer function in the parameter range, when Ω_{B} is a substantial fraction of $\Omega_{\circ}$. These transfer functions have been applied by Eisenstein *et al*. (1998) in an attempt to explain the existing bumps in the power spectrum based upon a fully linear calculation. The results are intriguing: they indicate that for a reasonable choice of the parameters it is difficult to associate the first acoustic peak directly with the observed bump in the power spectrum. There is still some freedom in adjusting the precise value of the Hubble constant h, or giving a small blue tilt to the spectrum, although the available range in these parameters is becoming smaller every year. There are several caveats though, mostly related to how well the current redshift surveys sample k-space and what effect nonlinearities can have on the bumps.

(*b*) *Nonlinear enhancements*

There are also several other amplification mechanisms at work. The surveys measure the distribution of galaxies, while the above calculations refer to the linear fluctuations in all the mass. First of all, the formation of the walls is a highly nonlinear process, which will amplify fluctuations if there is a distinct scale associated with them. Second, the galaxy surveys are analysed in redshift space; thus infall on to the walls will enhance these structures and will result in a further amplification. This effect will depend on the survey geometry: it is very important for pencil beams, less so for slices and spherical volumes. Even in the Las Campanas survey one can notice that some of the walls curve to stay perpendicular to the line of sight—a consequence of redshift space enhancements.

The most important nonlinear phenomenon is the Zel'dovich displacement, due to the coherent infall onto the 'walls'. The expectation value of the infall velocity is the ensemble average $\langle \rho v_r \rangle$, the density–velocity correlation function. The peak of this integral is phase shifted from the peak of the correlation function. Expressing this another way: characteristic scales in the Zel'dovich approximation tend to appear not at the maximum of the power, but where the spectrum is steepest. We are currently undertaking detailed analysis of these effects, including large-scale numerical simulations using the Zel'dovich approximation with a high baryon transfer function and large survey volumes. This should provide a final answer to whether this explanation of the bumps is feasible. All other alternatives (isocurvature fluctuations, tilts) are much less attractive, since they would require new physics on these large scales.

6. Conclusions

Over the next few years several new large-scale surveys will start producing data, like the Sloan Digital Sky Survey and the 2dF (Colless, this issue). The analysis techniques outlined here, based on the KL transform, combined with the new dataset, could result in major new developments in understanding the nature of the fluctuations on scales over 100 Mpc. They can measure the shape of the fluctuation spectrum in an overlap region with COBE. The method is capable of including systematic effects, redshift distortions and incompletenesses in the data, representing considerable improvements over current state of the art techniques.

Several observations are pointing to excess power on 100–130 h^{-1} Mpc scales, which manifests itself in a small number of sharp spikes in Fourier space. These reflect the presence of walls and voids on similar scales. The emergence of this comoving scale at high redshift implies that this is imprinted on the fluctuations. Such a scale occurs naturally at recombination. The Sakharov oscillations, remnants of the sound waves at that epoch, may provide an intriguing explanation. In such scenarios the baryon content of the universe must be high, the Hubble constant low, and the universe open. This family of models deserves further investigation, and it is just barely possible that the 100 Mpc bumps may be the first preview of the elusive Doppler peaks—a fascinating preview of further connections between the galaxy distribution and the CMB.

References

Amendola, L. 1994 *Astrophys. J. Lett.* **430**, 9.

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

Bartlett, J. G., Blanchard, A., Silk, J. & Turner, M. S. 1995 *Science* **267**, 980.

Broadhurst, T. J., Ellis, R. S., Koo, D. C. & Szalay, A. S. 1990 *Nature* **343**, 726.

Broadhurst, T. J., Ellis, R. S., Ellman, N. E., Koo, D. C. & Szalay, A. S. 1995 *Proc. of Wide Field Spectroscopy in the Distant Universe* (ed. S. Maddox & A. Aragon Salamanca), p. 178. Singapore: World Scientific.

Chincarini, G. & Rood, H. J. 1976 *Astrophys. J.* **206**, 30.

Cohen, J. G., Cowie, L. L., Hogg D. W., Songalia, A., Blandford, R., Hu, E. M. & Shopbell, P. 1996 *Astrophys. J.* **471**, 5.

Davis, R. L., Hodges, H. M., Smoot, G. F., Steinhardt, P. J. & Turner, M. S. 1992 *Phys. Rev. Lett.* **69**, 1856.

deLapparent, V., Geller, M. J. & Huchra, J. P. 1986 *Astrophys. J. Lett.* **301**, 1.

Einasto, J., Gottloeber, S., Mueller, V., Saar, V., Starobinsky, A. A., Tago, E., Tucker, D., Andernach, H. & Frisch, P. 1997 *Nature* **385**, 139.

Eisenstein, D. & Hu, W. 1998 *Astrophys. J.* **496**, 60.

Eisenstein, D., Hu, W., Silk, J. & Szalay, A. S. 1998 *Astrophys. J. Lett.* **494**, 1.

Feldman, H. A., Kaiser, N. & Peacock, J. A. 1994 *Astrophys. J.* **426**, 23.

Fisher, K. B., Davis, M., Strauss, M. A., Yahil, A. & Huchra, J. P. 1993 *Astrophys. J.* **402**, 42.

Geller, M. J. & Huchra, J. P. 1989 *Science* **246**, 897.

Gorski, K. M., Hinshaw, G., Banday, A. J., Bennett, C. L., Wright, E. L., Kogut, A., Smooth, G. F. & Lubin, P. 1994 *Astrophys. J. Lett.* **430**, 89.

Gregory, S. A. & Thompson, L. A. 1978 *Astrophys. J.* **222**, 784.

Guzzo, L., Collins, C. A., Nichol, R. C. & Lumsden, S. L. 1992 *Astrophys. J. Lett.* **393**, 5.

Hamilton, A. J. S. 1992 *Astrophys. J. Lett.* **385**, 5.

Harrison, E. R. 1970 *Phys. Rev.* D **1**, 2726.

Hu, W. & Sugiyama, N. 1996 *Astrophys. J.* **471**, 542.

Jain, B., Mo, H. J. & White, S. D. M. 1995 *Mon. Not. R. Astr. Soc.* **276**, L25.

Kaiser, N. 1984 *Astrophys. J. Lett.* **284**, 9.

Kaiser, N. 1986 *Mon. Not. R. Astr. Soc.* **219**, 785.

Kirshner, R. P., Oemler, A. J., Schechter, P. L. & Shectman, S. A. 1983 *Astron. J.* **88**, 1285.

Klypin, A., Holtzman, J., Primack, J. & Regös, E. 1993 *Astrophys. J.* **416**, 1.

Kofman, L., Gnedin, N. & Bahcall, N. A. 1993 *Astrophys. J.* **413**, 1.

Landy, S. D., Shectman, S. A., Lin H., Kirshner, R. P., Oemler, A. A. & Tucker, D. 1996 *Astrophys. J. Lett.* **456**, 1.

Lilly, S. J., Tresse, L., Hammer, F., Crampton, D. & Le Fevre, O. 1995 *Astrophys. J.* **455**, 108.

Loveday, J., Efstathiou, G., Peterson B. A. & Maddox, S. J. 1992 *Astrophys. J.* **400**, L43.

Maddox, S. J., Efstathiou, G., Sutherland, W. J. & Loveday, J. 1990 *Mon. Not. R. Astr. Soc.* **242**, 43P.

Park, C., Vogeley, M. S., Geller, M. J. & Huchra, J. P. 1994 *Astrophys. J.* **431**, 569.

Peacock, J. A. & Dodds, S. J. 1996 *Mon. Not. R. Astr. Soc.* **280**, 19.

Peebles, P. J. E. 1968 *Astrophys. J.* **153**, 1.

Proust, D. 1999 *Proc. Postdam Cosmology Meeting* (ed. S. Gottloeber). (In the press.)

Quashnock, J. M., Van Den Berk, D. E. & York, D. G. 1996 *Astrophys. J.* **472**, 69.

Sachs, R. K. & Wolfe, A. M. 1967 *Astrophys. J.* **147**, 73.

Sakharov, A. 1966 *Soviet Phys. JETP* **22**, 241.

Saunders, W., Frenk, C., Rowan-Robinson, M., Lawrence, A. & Efstathiou, G. 1991 *Nature* **349**, 32.

Shectman, S. A., Landy, S. A., Oemler, A., Tucker, D., Lin, H., Kirschner, R. L. & Schechter, P. L. 1996 *Astrophys. J. Suppl.* **470**, 172.

Smoot, G. F. (and 27 others) 1992 *Astrophys. J. Lett.* **396**, 1.

Sunyaev, R. A. & Zel'dovich, Ya. B. 1970 *Astrophys. Space Sci.* **7**, 3.

Steidel, C., Adelberger, K., Dickinson, M., Giavalisco, M., Pettini, M. & Kellogg, M. 1999 *Astrophys. J.* (In the press.)

Szalay, A. S., Matsubara, T. & Landy, S. A. 1999 *Astrophys. J.* (In the press.)

Tegmark, M., Hamilton, A. J. S., Strauss, M., Vogeley, M. S. & Szalay, A. S. 1998 *Astrophys. J.* **499**, 555.

Therrien, C. W. 1992 *Discrete random signals and statistical signal processing.* Englewood Cliffs, NJ: Prentice-Hall.

Tully, R. B., Scaramella, R., Vettolani, G. & Zamorani, G. 1992 *Astrophys. J.* **388**, 9.

Tytler, D. 1997 *Proc. 18th Texas Symp. on Rel. Astrophys* (ed. J. Frieman & A. Olinto).

Vettolani, G. (and 18 others) 1997 *Astron. Astrophys.* **325**, 954.

Vogeley, M. S. & Connolly, A. J. 1999 *Astrophys. J.* (Submitted.)
Vogeley, M. S. & Szalay, A. S. 1996 *Astrophys. J.* **465**, 34.
Zaroubi, S. & Hoffmann, Y. 1996 *Astrophys. J.* **462**, 25.
Zel'dovich, Ya. B. 1970 *Astron. Astrophys.* **5**, 84.
Zel'dovich, Ya. B. 1972 *Mon. Not. R. Astr. Soc.* **160**, 1P.