FAST EDGE-CORRECTED MEASUREMENT OF THE TWO-POINT CORRELATION FUNCTION AND THE POWER SPECTRUM

ISTVÁN SZAPUDI,¹ JUN PAN,¹,² SIMON PRUNET,³ AND TAMÁS BUDÁVÁRI⁴

Received 2005 May 16; accepted 2005 August 8; published 2005 August 31

ABSTRACT

We present a pair of related techniques to measure the two-point correlation function and the power spectrum with edge correction in any number of spatial dimensions. The underlying algorithm achieves unprecedented speed by using fast Fourier transforms for calculating a heuristically weighted, edge-corrected estimator for the two-point function. With its \( N \log N \) scaling, it will be able to keep up with the expected exponential increase in astronomical data. Such a speedup is especially important for massive Monte Carlo studies with large data sets. In addition, we present a new method to estimate the power spectrum by means of numerical integration of the Hankel transform of the measured two-point correlation function. Stability and accuracy are ensured by a novel numerical technique based on Gauss-Bessel quadrature and double exponential transformation. The resulting edge-corrected estimator for the power spectrum reduces the smearing effect of the survey window function. The increased resolution will help to better constrain the shape of the power spectrum, such as features arising from baryonic oscillations.

Subject headings: cosmology: theory — large-scale structure of universe — methods: statistical

Online material: color figures

1. INTRODUCTION

The two-point correlation function and its Fourier transform, the power spectrum, are the most measured statistics in studies of large-scale structure. In theory, these two statistics are simply related, yet they are perceived as complementary in practice. The correlation function can be measured with the nearly optimal edge-corrected estimator of Landy & Szalay (1993) or Szapudi & Szalay (1998, 2000) (cf. eq. [1]); for an alternative, see Hamilton (1993). On the contrary, power spectrum estimators (e.g., Feldman et al. 1994) do not provide for edge correction: the estimates are convolved with the survey window function, which in turn translates into a loss of resolution compared with the two-point function.

The power spectrum expresses naturally the translational invariance of space, and it can be calculated using fast Fourier transforms (FFTs) with \( N \log N \) scaling, where \( N \) is the number of data elements (pixels) in the survey. Correlation function estimators are based on the intuitive notion of pair counting: their naïve realization requires \( O(N^2) \) operations. This, although not yet prohibitive, is increasingly becoming a burden with the exponential growth of astronomical data sets. Pair counting can be sped up, if one is interested in measuring the correlation function on small scales with reasonably large bins. In that case, the double-tree algorithm by Moore et al. (2001) scales approximately as \( O(N \log N) \). However, if all scales are considered with high resolution, the tree-based algorithm slows down, matching the \( O(N^2) \) scaling of naive counting (since it has to split all nodes of the tree).

The above principal differences account for the prevailing wisdom that the two-point correlation function and the power spectrum are complementary. Correlation functions are used on smaller scales, where naïve estimation is fast and high resolution is desired; power spectra are estimated mainly on large scales, taking advantage of the fast estimation but without correction for the smearing effects of the survey window.

The simple relation between the two statistics, and the entirely technical nature of the complementarity, motivates us to ask the following questions: Can we take advantage of the FFT algorithms and calculate the correlation function with a fast \( O(N \log N) \) algorithm even on large scales? Can we exploit the edge-corrected estimator for the two-point correlation function and estimate high-resolution (edge-corrected) power spectra? We will show that with careful analysis of the discrete Fourier transform, one can indeed design a fast algorithm that produces results that exactly match that of a direct calculation of the two-point function on a grid. Moreover, a recent advance in numerical mathematics allows us to devise a novel method to obtain power spectra from a measured two-point function in a robust and accurate way. Thus, even in practice, the two-point correlation functions and power spectra are equivalent. One can use whichever is more convenient to achieve a given scientific purpose. The new techniques are similar to recent developments in cosmic microwave background (CMB) research, where edge-corrected power spectra (e.g., Szapudi et al. 2001a; Hivon et al. 2002) are now preferred over uncorrected “pseudo–power spectra.”

In § 2, we describe a novel technique to measure edge-corrected correlation functions using an FFT-based algorithm, and a numerical method to obtain edge-corrected power spectra. We present numerical results in § 3 and summarize and discuss them in § 4.

2. DESCRIPTION OF THE TECHNIQUE

The power spectrum is the Fourier transform of the two-point correlation function (by the Wiener-Khinchin theorem). This basic idea was exploited in Szapudi et al. (2001a) to devise a fast method for estimating the angular power spectrum \( C_l \)'s with fast harmonic transforms. A crucial intermediate step in that method was to estimate correlation functions (Szapudi et
al. 2001b). Although the window function can be deconvolved in harmonic space (Hivon et al. 2002), the real-space analog corresponds to the edge correction put forward by Szapudi & Szalay (1998, 2000). The matrix representing the window function is inverted in pixel space, where it is diagonal. Next we present analogous methods for flat (Euclidean) space with arbitrary dimensions. In turn, we introduce two novel techniques, useful individually or as a pair: (1) fast estimation of the two-point correlation function by means of FFT and (2) estimation of the power spectrum by means of a numerical Hankel transform of the measured two-point function.

2.1. Two-Point Correlation Function

The minimum variance estimator of Szapudi & Szalay (1998) is most natural when defined on a grid. If \( \delta \) is the density field on grid point \( i \), we can define the estimator as

\[
\hat{\xi}_D = \sum_{i,j} f_{ij} \delta_i \delta_j / \sum_{i,j} f_{ij},
\]

where \( f_{ij} \) is the pair weight and we assumed that \( \langle \delta \rangle = 0 \). Naive estimation with general pair weights is \( O(N^2) \), where \( N \) is the number of pixels. As we will see next, fast estimation is possible in the special case in which the pair weight \( f_{ij} = f_{ij} \delta_{i-j,2} \), that is, the weight is multiplicative and depends only on a shift \( \Delta \). Here \( \delta_{i-j,2} \) is the Kronecker delta, taking values of 1 when \( k = 0 \) and 0 otherwise. The simplest example is flat weighting, \( f_{ij} = 1 \). The normalization in an arbitrary complex geometry is obtained in practice by calculating the raw (unnormalized) correlation function of the geometry, putting 1’s into grid points fully inside the mask and 0’s everywhere else. Although our notation suggests one dimension, higher dimensions \( D \) are included by replacing indices, say, \( i, i_1 \), and with tuples \( (i_1, \ldots, i_D) \).

Computationally, one needs to calculate pair summations of the form \( \Sigma a_i a_{i+1} \). These are first reformulated to make use of fast Fourier transforms. If \( P(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1} \) is a polynomial and \( e^{\epsilon x} \) are unit roots, the coefficients of a discrete Fourier transform of the series \( a_i \) can be defined as

\[
a_{\kappa} = P(e^{\epsilon}) = a_0 + a_1 \epsilon + \ldots + a_{n-1} \epsilon^{(n-1)\kappa}.
\]

Direct calculation confirms that \( \Sigma a_i b_{i+1} \) can be calculated by Fourier transforming the series \( a_i \) and \( b_i \), calculating a discrete, anisotropic “pseudo–power spectrum” \( \hat{P}(k) = \hat{a}_i \hat{b}_{i+1}^* \), and finally inverse Fourier transforming back. The algorithm can be trivially extended for multiplicative weights by simply pre-multiplying each pixel by its weight.

According to the foregoing, an FFT-based algorithm for the two-point function has the following steps: (1) sum the objects in a sufficiently fine grid, storing the value \( N_i \), the number of objects (e.g., galaxies) at (vector) grid point \( k \) (this step is omitted if the density field is given a priori, such as in Euclidean approximation of CMB maps), (2) calculate fluctuations of the field as \( \delta = (N - \langle N \rangle)/\langle N \rangle \) for each grid point, (3) zero out all pixels that are masked out, (4) optionally weight each point with a minimum variance weight (e.g., \( J_i \) weighting, or Feldman et al. [1994]/Percival et al. [2004] weighting), (5) discrete Fourier transform with an FFT engine, (6) calculate \( \hat{a}_i \hat{b}_{i+1}^* \), and (7) Fourier transform back. To avoid aliasing effects, the survey needs to be put into a zero-padded box larger than the largest bin of the correlation function. The exact same procedure is followed for the mask/weight map. Finally, the raw correlation function is divided by the correlation function of the mask, according to equation (1). The result is the anisotropic correlation function \( \hat{\xi}(r) \), depending on vector shifts \( r \). The traditional correlation function is obtained as a sum in rings and spheres for \( D = 2 \) and \( D = 3 \), respectively.

2.2. Power Spectrum

The power spectrum is a Fourier transform of the two-point correlation function. With rotational invariance this becomes a Hankel transform in \( D = 2 \) or \( D = 3 \) dimensions.

\[
\xi(r) = \begin{cases} \int \frac{k \, dk}{2\pi} P(k) J_0(kr), & \text{if } D = 2, \\ \int \frac{k^2 \, dk}{2\pi^2} P(k) j_0(kr), & \text{if } D = 3, \end{cases}
\]

where \( J_0 \) and \( j_0 \) are the ordinary and spherical Bessel functions, respectively. We will use the self-inverse property of the Hankel transform to obtain an estimator of the power spectrum.

In practice, it is a delicate numerical procedure to calculate a Hankel transform numerically, which is the main reason that the above formula, while well known, has not been exploited in the past for estimating the power spectrum. We use the quadrature formula of Ogata (2005), which uses the roots of the Bessel functions as well as the double exponential transformation of Ooura & Mori (1999). These recipes are new developments in numerical mathematics, which we summarize next. To integrate over an arbitrary function \( f \) multiplied by a Bessel function, we use the formula

\[
\int_0^\infty f(x) J_0(rx) \, dx = \pi \sum_{k=1}^\infty w_k f(\pi\psi(hr_k)/h) J_1(\pi\psi(hr_k)/h) \psi(hr_k),
\]

where \( r_k \) are the roots of the Bessel function \( J_0 \), divided by \( \pi \), \( \psi(t) = t \tan(\frac{\pi}{2} \sin t) \) is the double exponential transformation, and \( h \) is the step of the integration (analogous to that of the trapezoidal rule). The weights are calculated as \( w_k = Y_1(\pi r_k)/|J_1(\pi r_k)| \), where the \( Y_1 \) are Bessel functions of the second kind. For inverting the two-point function, \( v = 0 \) and \( r = \frac{1}{2} \) Bessel functions should be used for two and three spatial dimensions according to equation (4).

3. RESULTS

The C implementation of the above algorithm, eSpICE (Euclidean version of SpICE, the spatially inhomogeneous correlation estimator of Szapudi et al. 2001a), uses the FFTW3 package (Frigo & Johnson 2005). The program was extensively tested and validated by comparison with naive pair-counting codes in two and three dimensions; Budavári et al. (2003) applies the algorithm to real data. Next we illustrate its speed and accuracy in three dimensions.

We performed measurements in \( \Lambda \)CDM simulations by the Virgo Supercomputing Consortium (Jenkins et al. 1998). We used the VLS (Very Large Simulation) with the following cosmological parameters: \( \Omega_m = 0.3, \Omega_{b} = 0.7, \Gamma = 0.21, h = 0.7 \), and \( \sigma_8 = 0.9 \). In order to estimate measurement errors, we divided the VLS into eight independent subsets; thus, each subset was a \( (239.5 \, h^{-1} \, \text{Mpc})^3 \) cube. In addition, a \( (119.75 \, h^{-1} \, \text{Mpc})^3 \) subcube was removed from each of the
We have measured the correlation function in logarithmic bins in the range 0.3–162 h⁻¹ Mpc. A measurement on a 768³ grid took about 15 minutes on one Opteron processor. We estimate that same measurement would have taken about 50 yr on the same computer with our naive counting code—a ≈2 million-fold speedup. We also performed additional measurements with the Moore et al. (2001) tree-based algorithm in logarithmic bins between 0.1 and 38 h⁻¹ Mpc. The required CPU time of this latter algorithm increases drastically on large scales; hence, we stopped at smaller scales. To be able to run the algorithm in approximately 1 day, we also needed to dilute the simulations to 2%. Even though the two runs were not identical, the two algorithms produced virtually identical results, as shown in Figure 1. In addition, we measured the correlation function without edge corrections, using the pseudo–power spectrum directly. The difference is quite significant, especially on larger scales.

Then we calculated the edge-corrected power spectrum. First, we performed a cubic spline interpolation on the measured correlation function at the roots of Bessel functions. Then we followed equation (4) for \( p \) = \( \frac{1}{3} \) to integrate the first case. We used \( h = 1/32 \) and truncated the sum at \( N = 200 \); doubling these parameters had no significant effect on our results. We have checked the accuracy or our integration routine with \( N = 10,000 \) point direct double exponential integration following Ooura & Mori (1999), and with Mathematica. This showed that this fast 200-point integration is better than 0.1% accurate.

We applied the inversion to the results both from eSpICE and from the tree-based algorithm (see Fig. 2). We used a conservative scale range of \( (k_{\text{min}}, k_{\text{max}}) = (2\pi/r_{\text{max}}, 1/r_{\text{min}}) \). In addition, we plotted the linear-theory and nonlinear Smith et al. (2003) power spectra. The agreement of our reconstruction with the phenomenology is excellent. For comparison, the pseudo–power spectrum (obtained from the correlation function without edge correction) is shown. It is slightly biased compared with the edge-corrected power spectrum, although the difference is less apparent than for the two-point correlation function, since the window function effect is a convolution (i.e., smoothing) of an already smooth power spectrum.

4. SUMMARY AND DISCUSSION

We have presented an FFT-based grid algorithm that is complementary to the tree-based pair counting of Moore et al. (2001). While the latter slows down dramatically on large scales, no corresponding effect constrains our technique. On the other hand, the grid used in our method has a smoothing effect at a few grid spacings, which restricts its applicability on the smallest scales. There is ample overlap between the two algorithms where they perform similarly. Our algorithm yields a substantial (order of a million) speedup compared with traditional methods. Given the exponential growth of astronomical data in accordance with Moore’s law, speed becomes increasingly significant: future data sets will not be analyzable with naive techniques. In addition, our high-speed technique opens new possibilities for exploring systematic and statistical effects through massive Monte Carlo simulations. In the age of “high-precision cosmology,” control of errors to the utmost degree is as important as the measurement itself. As an intermediate step, our algorithm produces a novel anisotropic correlation function of vector shifts \( \xi(r) \). These were used in Budavári et al. (2003) to filter out systematic effects from the drift-scanning operations of the Sloan Digital Sky Survey camera, and it has potential use in constraining systematics in other surveys as well. In three dimensions, the same technique can be used for estimation of the redshift-space correlation function in the distant-observer approximation, for narrow, deep redshift surveys.

We also presented a new method to turn a measured correlation function into an edge-corrected power spectrum. It exploits numerical Fourier-Hankel transforms to obtain a robust and accurate estimate of the power spectrum from the two-point function. Although we did not correct for grid effects in the inverted power spectrum, it should scale as \( W(k)^2 = \text{sinc}^2(kR) \) in \( D \) dimensions. The resulting technique can be used to process the two-point function measurements obtained with...
our algorithm, or from any other method. In particular, from large simulations it is difficult to obtain the power spectrum for large \( k \) (or small scales), because of memory requirements. Previously, Jenkins et al. (1998) used a folding technique to measure \( P(k) \) for wavenumbers larger than the Nyquist wavenumber of the largest grid fitting in the memory of the computer: such a folding procedure erases a large number of modes. In contrast, our inversion of the correlation function includes all modes. Thus, coupled with the algorithm of Moore et al. (2001), it is able to estimate the power spectrum in high-resolution simulations with unprecedented accuracy, even for the highest wavenumbers. Ours is by no means the only way to obtain the power spectrum from two-point functions. Baugh & Efstathiou (1993; also Maddox et al. 1996) and Dodson & Gaztañaga (2000) have obtained three-dimensional power spectra from angular correlation functions with Lucy’s deconvolution and brute-force maximum likelihood, respectively. Their focus on deprojection necessitated the use of these nonlinear, iterative (and thus slow) methods. FFTLog, by Hamilton (2000), relates to our method as discrete Fourier transforms to numerical integrals of oscillatory integrands; this means our method is likely to be more robust.

Starting with linear bins (an option in eSpICE with no effect on speed), we would obtain \( P(k) \) in linear bins as well. The maximum \( r_{\text{max}} \) at which the correlation function is measured determines the resolution and the smallest \( k_{\text{max}} \) \( = \Delta k = 1/r_{\text{max}} \) for \( P(k) \); the resolution and the smallest \( r_{\text{max}} \) at which the correlation function is measured determine \( k_{\text{max}} \) \( = 1/r_{\text{max}} \). However, emulating typical measurements of the past, we used logarithmic bins. As a result, we obtained \( P(k) \) in approximately logarithmic bins in \( k \). To see this, one has to note that the particular resolution corresponding to a given \( k \) in the power spectrum determines a maximum \( r \) up to which the correlation function is measured (at least) at that resolution. This maximum translates into an effective resolution \( \Delta k \) at the given \( k \).

Studies of large-scale structure, up until now, have estimated “pseudo-power spectra,” that is, a power spectrum convolved with the survey geometry. The potential of detecting baryonic oscillations motivates the measurement of power spectra with the highest resolution possible. In particular, if one of the survey dimensions is much smaller than the other two, the effective convolution kernel will be much wider than desirable.

Above, we demonstrated that the power spectrum can be recovered from steps not any more complicated than the measurement of the corresponding pseudo-power spectrum. While the information content is not increased with the inversion (e.g., Efstathiou 2004), higher resolution and correction for the window function effect are immensely useful when comparing measurements from different surveys and different geometries, as has been common in CMB research (Szapudi et al. 2001a).

Redshift distortions approximately break the translational invariance underlying the Fourier transform. This affects our method as it does the Feldman et al. (1994) type direct methods. We conjecture that our method of inverting the correlation function can be generalized for redshift distortions by means of the formulae of Szapudi (2004b); this is left for future research.

Note that the present technique can be used to estimate \( C_\ell \)’s in the flat-sky approximation, using the asymptotic behavior of the Legendre polynomials, \( P(\cos \theta) \approx (2/(\pi \sin \theta))^{1/2} \times \cos [(l + 1) \theta - \pi/4] \), and the Bessel functions, \( J_\ell(z) \approx (2/(\pi z))^{1/2} \cos (z - \pi/4) \). The correlation function can be measured with eSpICE, and integration of equation (3) gives the \( C_\ell \)’s with \( k = 1 + 1/\ell \). Generalization of the proposed computational and inversion techniques to gravitational lensing follows directly from the generalization of SpICE to CMB polarization (e.g., Chon et al. 2004). Finally, generalization of equation (3) for a relation between the three-point function and the bispectrum (Szapudi 2004a), together with a fast algorithm to measure the three-point function, yields a new, edge-corrected method to measure the bispectrum. These generalizations will be presented elsewhere.

I. S. thanks Alex Szalay for stimulating discussions, John Peacock and Jacek Guzik for helpful suggestions, and Josh Hoblit and Gang Chen for computing help. I. S. and J. P. were supported by NASA through Applied Information Systems Research Program grant NAG5-11996 and Astrophysics Theory Program grant NAG5-12101, as well as by NSF grants AST 02-06243, AST 04-34413, and ITR 112021-128440. J. P. acknowledges support from PPARC grant PPA/G/S/2000/00057. The eSpICE code will be made public upon publication of this Letter.\(^5\)

\(^5\) See http://www.ifa.hawaii.edu/~szapudi/ for details.

REFERENCES

Baugh, C. M., & Efstathiou, G. 1993, MNRAS, 265, 145
Budavári, T., et al. 2003, ApJ, 595, 59
Chon, G., Challinor, A., Prunet, S., Hivon, E., & Szapudi, I. 2004, MNRAS, 350, 914
Dodson, S., & Gaztañaga, E. 2000, MNRAS, 312, 774
Efstathiou, G. 2004, MNRAS, 349, 603
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23
Frigo, M., & Johnson, S. G. 2005, Proc. IEEE, 93, 216
Hamilton, A. J. S. 1993, ApJ, 417, 19
———. 2000, MNRAS, 312, 257
Hivon, E., Górski, K. M., Netterfield, C. B., Crill, B. P., Prunet, S., & Hansen, F. 2002, ApJ, 567, 2
Jenkins, A., et al. 1998, ApJ, 499, 20
Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
Maddox, S. J., Efstathiou, G., & Sutherland, W. J. 1996, MNRAS, 283, 1227
Moore, A. W., et al. 2001, in Mining the Sky, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann (Heidelberg: Springer), 71
Ogata, H. 2005, Publ. Res. Inst. Math. Sci., in press
Oura, T., & Mori, M. 1999, J. Comput. Appl. Math., 112, 229
Percival, W. J., Verde, L., & Peacock, J. A. 2004, MNRAS, 347, 645
Smith, R. E., et al. 2003, MNRAS, 341, 1311
Szapudi, I. 2004a, ApJ, 605, L89
———. 2004b, ApJ, 614, 51
Szapudi, I., Prunet, S., & Colombi, S. 2001a, ApJ, 561, L11
Szapudi, I., Prunet, S., Pogosyan, D., Szalay, A. S., & Bond, J. R. 2001b, ApJ, 548, L115
Szapudi, I., & Szalay, A. S. 1998, ApJ, 494, L41
———. 2000, Ann. Inst. Stat. Univ. Paris, 44, 1