

Algorithms for Higher Order Spatial Statistics

István Szapudi

Institute for Astronomy
University of Hawaii

CIDU/AISRP Oct 14-16,2009

Outline

- 1 Introduction
- 2 Three-point Algorithm

Random Fields

Definition

A random field is a spatial field with an associated probability measure: $\mathcal{P}(A)DA$.

- Random fields are abundant in Cosmology.
- The cosmic microwave background fluctuations constitute a random field on a sphere.
- Other examples: Dark Matter Distribution, Galaxy Distribution, etc.
- We measure particular realization of a random field (ergodicity helps but we cannot avoid “cosmic errors”)

Definitions

- The ensemble average $\langle A \rangle$ corresponds to a functional integral over the probability measure.
- Physical meaning: average over independent realizations.
- Ergodicity: (we hope) ensemble average can be replaced with spatial averaging.
- Symmetries: translation and rotation invariance

Joint Moments

$$F^{(N)}(x_1, \dots, x_N) = \langle T(x_1), \dots, T(x_N) \rangle$$

Connected Moments

These are the most frequently used spatial statistics

- Typically we use fluctuation fields $\delta = T/\langle T \rangle - 1$

Connected moments are defined recursively

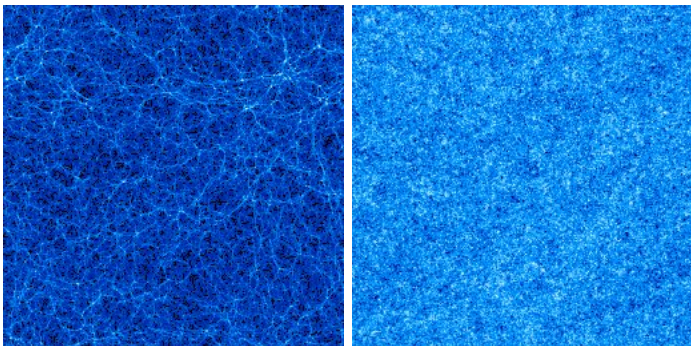
$$\langle \delta_1, \dots, \delta_N \rangle_c = \langle \delta_1, \dots, \delta_N \rangle - \sum_P \langle \delta_1 \dots \delta_i \rangle_c \dots \langle \delta_j \dots \delta_k \rangle_c \dots$$

- With these the N -point correlation functions are

$$\xi^{(N)}(1, \dots, N) = \langle \delta_1, \dots, \delta_N \rangle_c$$

Gaussian vs. Non-Gaussian distributions

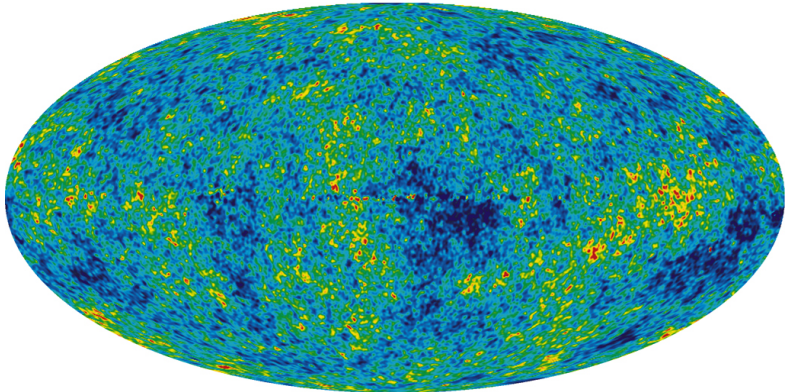
These two have the same two-point correlation function or $P(k)$



- These have the same two-point correlation function!

Cosmic Microwave Background

Gaussian to a good approximation



- But: inflationary predictions, f_{NL}

A Popular Parametrization

Phenomenology motivated by inflationary models

- $\Phi(k)$ is the primordial curvature perturbation

Perturbative Model

$$\Phi(x) = \Phi_L(x) + f_{NL}(\Phi(x)_L^2 - \langle \Phi_L^2(x) \rangle)$$

- The leading order effect will be bispectrum, (three-point correlations)

Bispectrum (Komatsu and Spergel 2001)

$$b_{l_1 l_2 l_3} = 2 \int r^2 dr (b_{l_1}^L(r) b_{l_2}^L(r) b_{l_3}^{NL}(r) + \text{perm.})$$

- where b^L and $b^{NL} = \frac{2}{\pi} \int k^2 dk f(k) g_{\mathcal{T}l}(k) j_l(k)$ with $f = P_\Phi(k)$ or f_{NL} respectively.
- Prediction through modification of CAMB (or CMBFAST).

Basic Objects

These are N -point correlation functions.

Special Cases

Two-point functions	$\langle \delta_1 \delta_2 \rangle$
Three-point functions	$\langle \delta_1 \delta_2 \delta_3 \rangle$
Cumulants	$\langle \delta_R^N \rangle_c = \mathbf{S}_N \langle \delta_R^2 \rangle^{N-1}$
Cumulant Correlators	$\langle \delta_1^N \delta_2^M \rangle_c$
Conditional Cumulants	$\langle \delta(0) \delta_R^N \rangle_c$

- In the above δ_R stands for the fluctuation field smoothed on scale R (different R 's could be used for each δ 's).
- Host of alternative statistics exist: e.g. Minkowski functions, void probability, minimal spanning trees, phase correlations, etc.

Complexities

Combinatorial explosion of terms

- N -point quantities have a large configuration space: measurement, visualization, and interpretation become complex.
- CPU intensive measurement: M^N scaling for N -point statistics of M objects.
- Theoretical estimation
- Estimating reliable covariance matrices

Algorithmic Scaling and Moore's Law

- Computational resources grow exponentially
- (Astronomical) data acquisition driven by the same technology
- Data grow with the same exponent

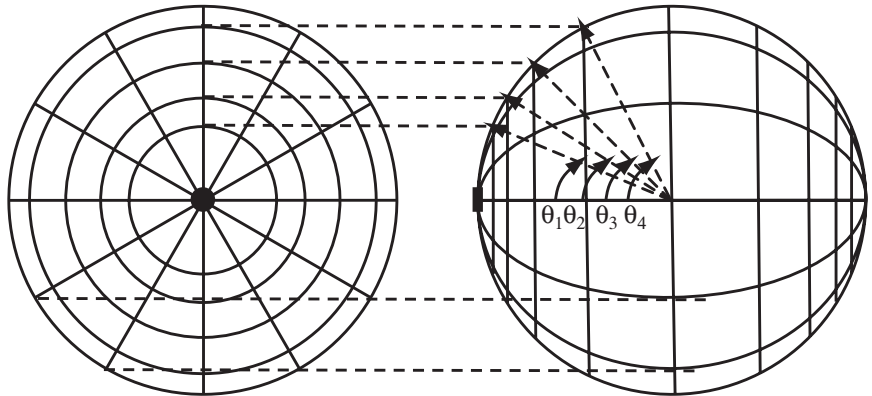
Corrolary

Any algorithm with a scaling worse then linear will become impossible soon

- Symmetries, hierarchical structures (kd-trees), MC, computational geometry, approximate methods

Example: Algorithm for 3pt

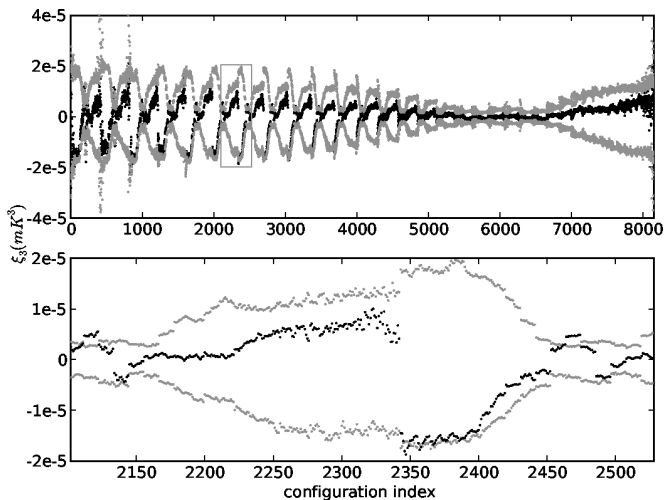
Other algorithms use symmetries



Algorithm for 3pt Cont'd

- Naively N^3 calculations to find all triplets in the map: overwhelming (millions of CPU years for WMAP)
- Regrid CMB sky around each point according to the resolution
- Use hierarchical algorithm for regridding: $N \log N$
- Correlate rings using FFT's (total speed: 2 minutes/cross-corr)
- The final scaling depends on resolution
$$N(\log N + N_\theta N_\alpha \log N_\alpha + N_\alpha N_\theta (N_\theta + 1)/4)/2$$
- With another cos transform one and a double Hankel transform one can get the bispectrum
- In WMAP-I: 168 possible cross correlations, about 1.6 million bins altogether.
- How to interpret such massive measurements?

3pt in WMAP

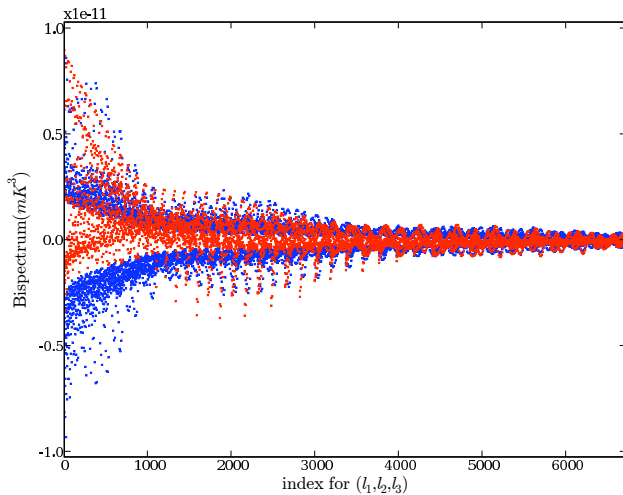


Recipe for Measuring the Bispectrum

With Edge Correction

- Measure the **full** three-point correlation function
- All configurations included with a fast algorithm (scaling is $N \log N$ times a factor depending resolution)
- Calculate Legendre moments of the three-point functions
- Double-Hankel (Bessel) transform each moment with J_l (or j_l for 3D)
- Optionally, recover the full bispectrum from its moments at a given resolution

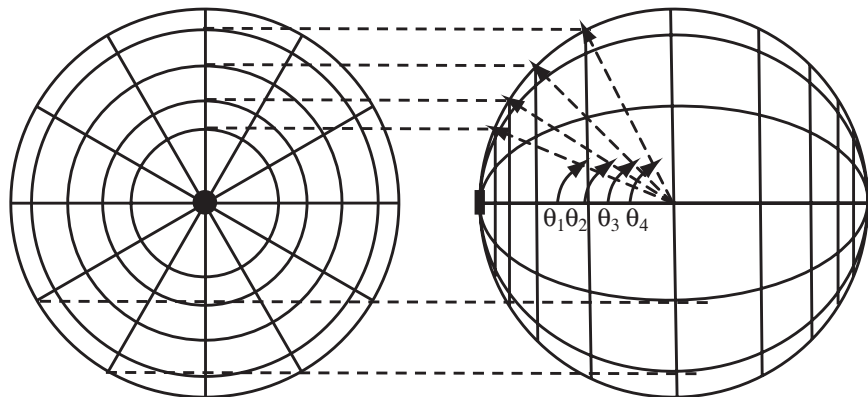
Full WMAP-bispectrum



Recent Challenges

- Processors becoming multicore (CPU and GPU)
- To take advantage of Moore's law: parallelization
- Disk sizes growing exponentially, but not the IO speed
- Data size can become so large that reading might dominate processing
- Not enough to just consider scaling

Alternative view of the algorithm: lossy compression



Compression

- Compression can increase processing speed simply by the need of reading less data
- The full compressed data set can be sent to all nodes
- This enables parallelization in multicore or MapReduce framework
- Specific (possibly lossy) can improve many algorithms

Compression Related Ideas

- Trading CPU for IO: Which astro/CMB algorithms can benefit from compression in a brute force way (e.g. image processing in a multicore framework)
- Which algorithms can benefit from **lossy pre-compression** phase for speed and parallization (e.g. three-point)
- What lossy compression algorithms can be found that are linear and by construction preserve certain quantities (e.g. flux),
- or in general contain the scientifically interesting information (e.g. storing all orbits in an N-body simulation).
- What information can be thrown away from data without loosing scientifically important information?
- What compression formats enable certain operations without decompressing?

Summary

- Fast algorithm for calculating 3pt functions with $N \log N$ scaling instead of N^3
- Approximate algorithm with a specific lossy compression phase
- Scaling with resolution and not with data elements
- Compression in the algorithm enables multicore or MapReduce style parallelization
- Planck Collaboration interest
- Trading CPU for IO