Algorithms for Higher Order Spatial Statistics

István Szapudi

Institute for Astronomy
University of Hawaii

Future of AstroComputing Conference, SDSC, Dec 16-17

Outline

Random Fields

Definition

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

- 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.
- Astronomers 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

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

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

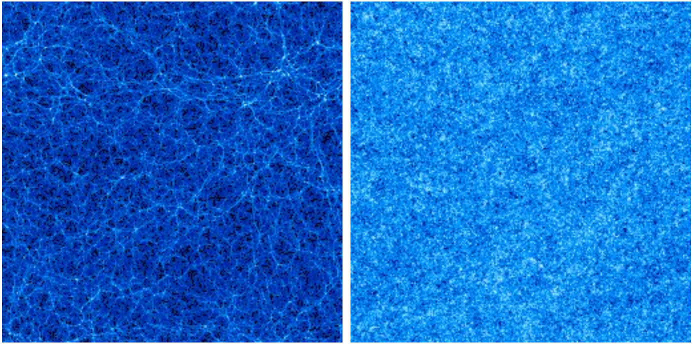
$$\langle \delta_1, \ldots, \delta_N \rangle_c = \langle \delta_1, \ldots, \delta_N \rangle - \sum_P \langle \delta_1 \ldots \ldots \delta_i \rangle_c \ldots \langle \delta_j \ldots \delta_k \rangle_c \ldots$$

- With these the N-point correlation functions are

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

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!

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	$\langle\delta_R^N\rangle_c = S_N\langle\delta_R^2\rangle^{N-1}$
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.
- e.g, already for CMB three-point, the total number of bins scales as $M^{3/2}$
- 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

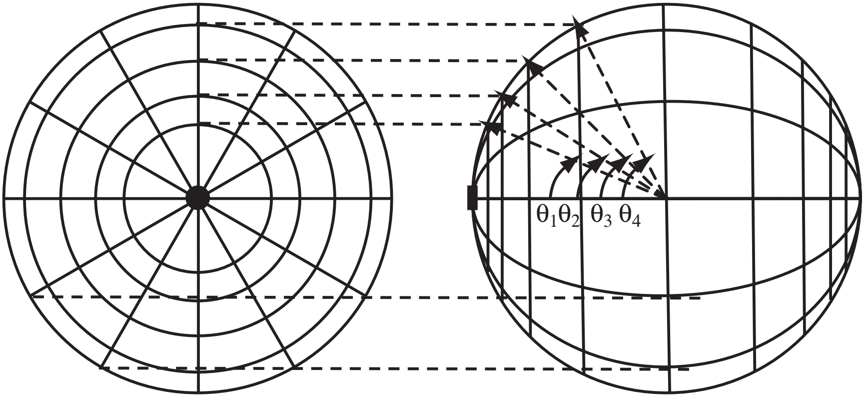
Corrolary

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

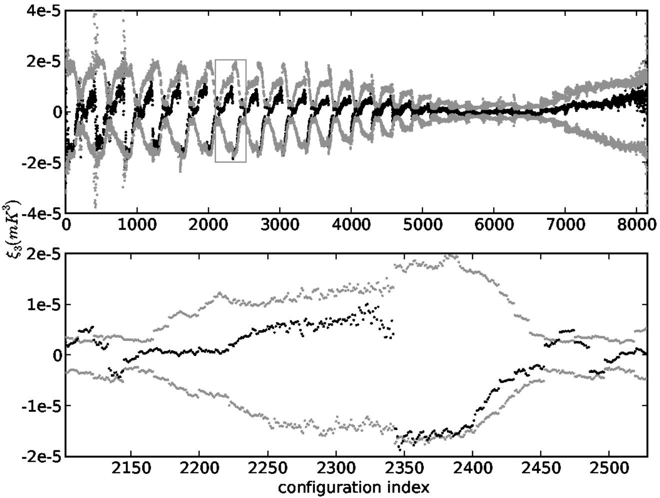
Other algorithms use symmetries



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?

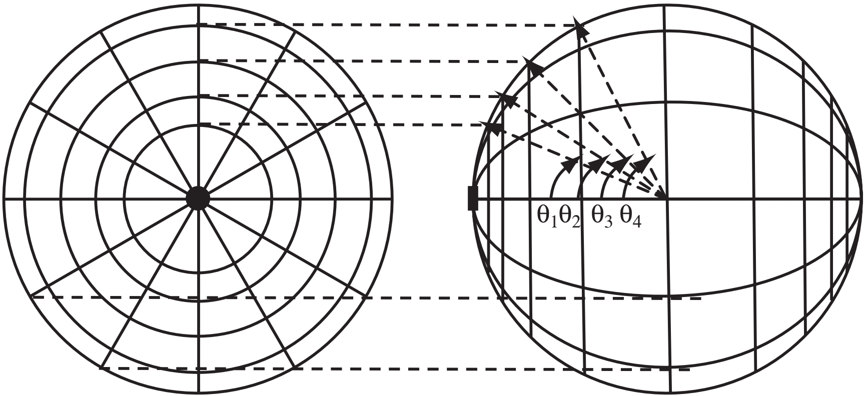
3pt in WMAP


4e-5
2e-5
0
-2e-5
-4e-5
0
1000
2000
3000
4000
5000
6000
7000
8000
$\xi_3(mK^3)$
2e-5
1e-5
0
-1e-5
-2e-5
2150
2200
2250
2300
2350
2400
2450
2500
configuration index

Recent Challanges

- 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

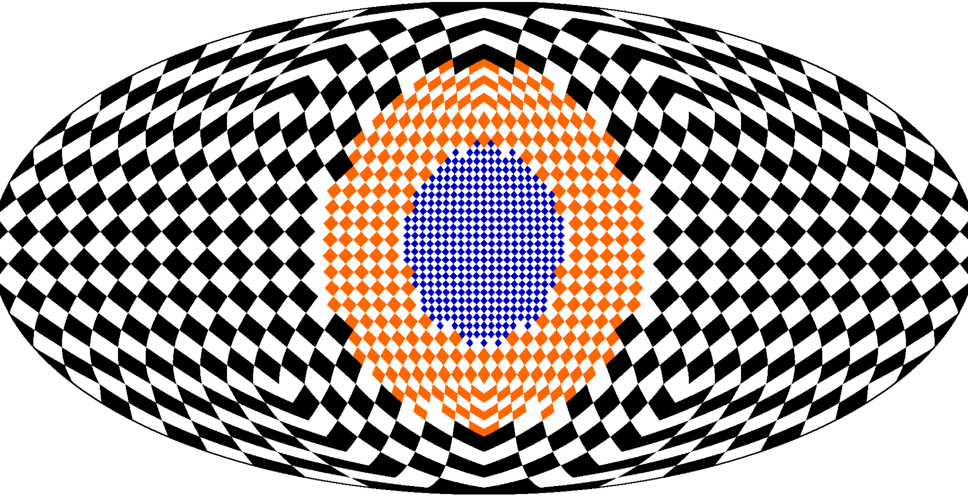
Alternative view of the algorithm: lossy compression



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
- For any algorithm specific (lossy compression) is needed

Another pixellization as lossy compression



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
- With a different compression we have done approximate likelihood analysis for CMB (Granett,PhD thesis)