Correlation Functions in Astrostatistics

Sten Delos

Department of Physics and Astronomy
UNC-Chapel Hill

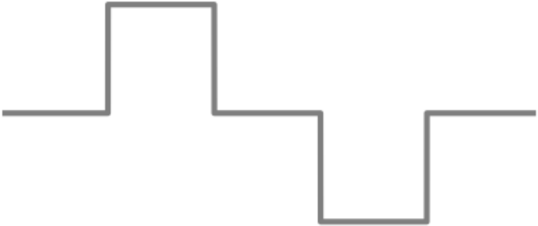
ASTR 703, Fall 2016

Outline

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad \text{("2-point autocorrelation")}$$

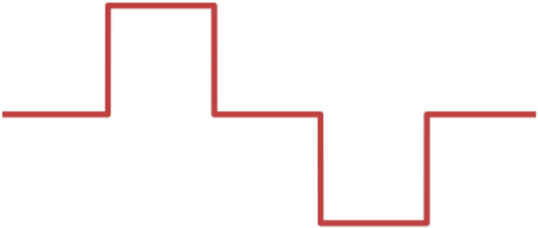
Average overlap between field and displaced field.



What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad \text{("2-point autocorrelation")}$$

Average overlap between field and displaced field.

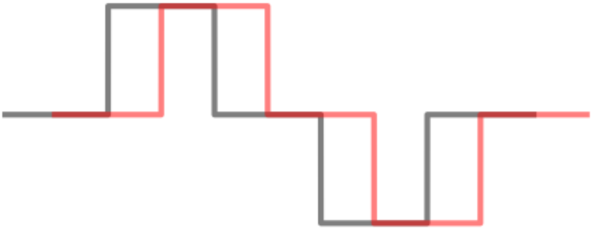


$$\xi(0) = \frac{1+1}{5} = 0.4$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad \text{("2-point autocorrelation")}$$

Average overlap between field and displaced field.

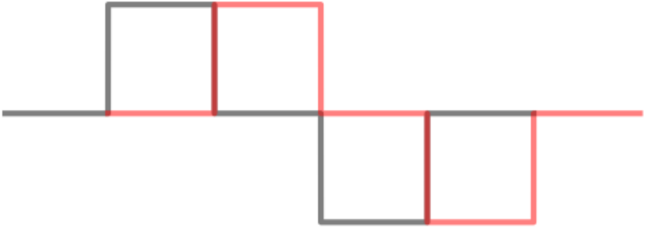


$$\xi(0.5) = \frac{0.5 + 0.5}{5} = 0.2$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle$$ ("2-point autocorrelation")

Average overlap between field and displaced field.

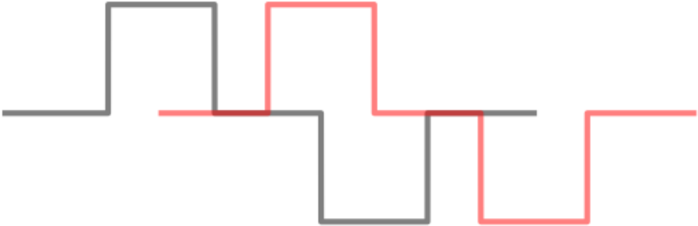


$$\xi(1) = 0$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad \text{("2-point autocorrelation")}$$

Average overlap between field and displaced field.

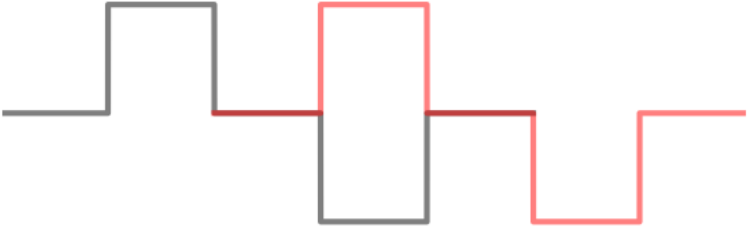


$$\xi(1.5) = \frac{-0.5}{5} = -0.1$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad (\text{"2-point autocorrelation"})$$

Average overlap between field and displaced field.

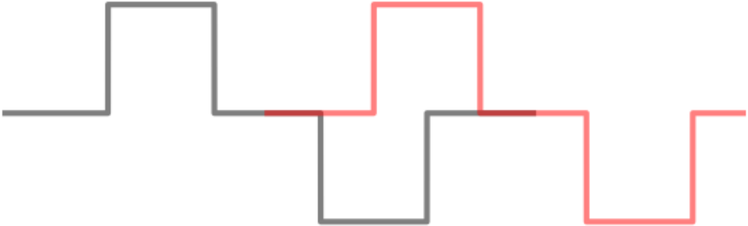


$$\xi(2) = \frac{-1}{5} = -0.2$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad \text{("2-point autocorrelation")}$$

Average overlap between field and displaced field.

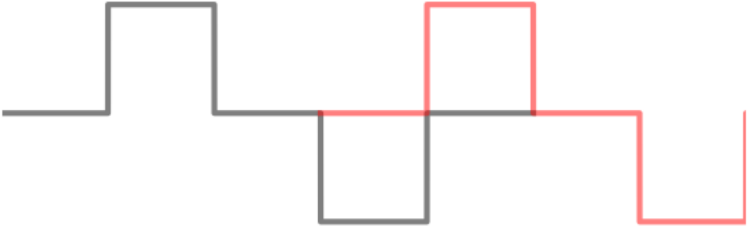


$$\xi(2.5) = \frac{-0.5}{5} = -0.1$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad \text{("2-point autocorrelation")}$$

Average overlap between field and displaced field.

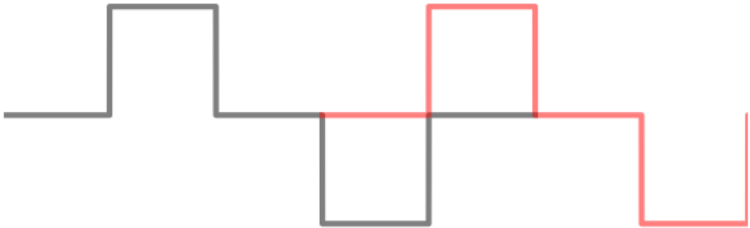


$$\xi(3) = 0$$

What is a correlation function?

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle \qquad (\text{"2-point autocorrelation"})$$

Average overlap between field and displaced field.



$$\xi(3) = 0$$

Often consider as function of $r \equiv |\vec{x}_1 - \vec{x}_2|$: tells you whether the field tends to have similar or dissimilar values over scale r.

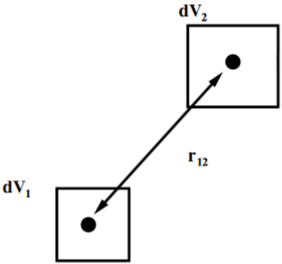
In astrostatistics

$\delta =$ fractional overdensity field:

$$\delta(\vec{x}) = \frac{n(\vec{x}) - \bar{n}}{\bar{n}}$$

Leads to probabilistic interpretation:

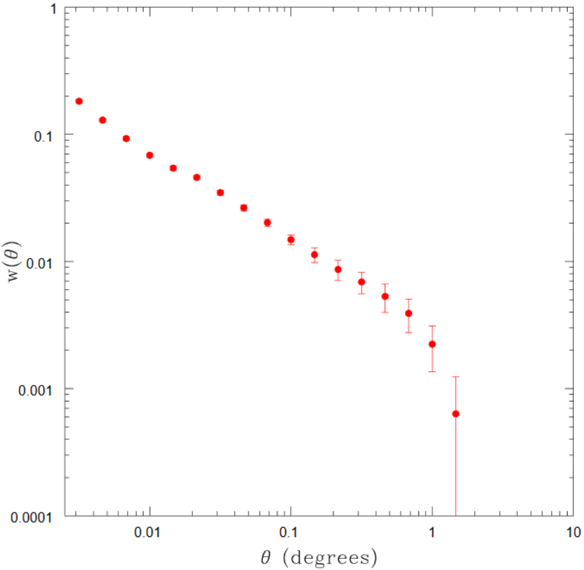
$$dP(\text{galaxy pair in volumes } dV_1 \text{ and } dV_2 \text{ separated by } r_{12})$$
$$= (1 + \xi(r_{12}))\bar{n}^2 dV_1 dV_2$$



Gives information about the scale of structure.

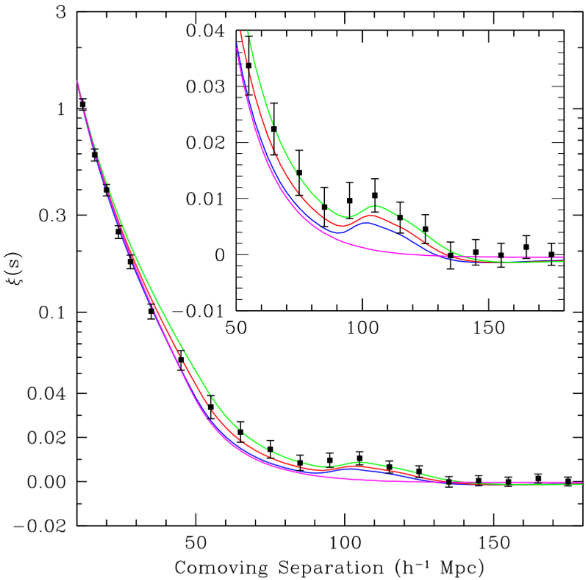
image source: http://dx.doi.org/10.1888/0333750888/2136

Example



Connolly et al 2002

Example



Eisenstein et al 2005

Outline

Pairwise computation

$$1+\xi(r_{12}) = \frac{dP(\text{galaxy pair in volumes } dV_1 \text{ and } dV_2 \text{ separated by } r_{12})}{\bar{n}^2 dV_1 dV_2}$$

Numerator:

- Bin distribution of distances of every data-data pair

Denominator:

- Build uniform random dataset in same volume
- Bin distribution of distances of every random-random pair

$$1 + \hat{\xi}(r) = \frac{\text{number of data-data pairs with binned distance r}}{\text{number of random-random pairs with binned distance r}}$$

Pairwise estimators

$$\hat{\xi}(r) = \frac{DD(r)}{RR(r)} - 1 \qquad \text{(Natural)}$$

$$\hat{\xi}(r) = \frac{DD(r)}{DR(r)} - 1 \qquad \text{(Davis \& Peebles)}$$

$$\hat{\xi}(r) = \frac{DD(r) - DR(r)}{RR(r)} \qquad \text{(Hewett)}$$

$$\hat{\xi}(r) = \frac{DD(r)RR(r)}{DR^2(r)} - 1 \qquad \text{(Hamilton)}$$

$$\hat{\xi}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \qquad \text{(Landy \& Szalay)}$$

Pairwise estimators

$$\hat{\xi}(r) = \frac{DD(r)}{RR(r)} - 1 \qquad \text{(Natural)}$$

$$\hat{\xi}(r) = \frac{DD(r)}{DR(r)} - 1 \qquad \text{(Davis \& Peebles)}$$

$$\hat{\xi}(r) = \frac{DD(r) - DR(r)}{RR(r)} \qquad \text{(Hewett)}$$

$$\hat{\xi}(r) = \frac{DD(r)RR(r)}{DR^2(r)} - 1 \qquad \text{(Hamilton)}$$

$$\hat{\xi}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \qquad \text{(Landy \& Szalay)}$$

To first order: Poisson variance, no bias, edge effects cancel
(Kerscher et al 2000)

Details of pairwise estimation

$$\hat{\xi}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)}$$

- Use much (~20x) larger random set than data set
- Scale DD, DR, RR to the total number of pairs

$$DD(r) = \frac{\text{number of data-data pairs with binned distance r}}{N_{data}(N_{data} - 1)/2}$$

$$RR(r) = \frac{\text{number of random-random pairs with binned distance r}}{N_{rand}(N_{rand} - 1)/2}$$

$$DR(r) = \frac{\text{number of data-random pairs with binned distance r}}{N_{data}N_{rand}}$$

Error bars

Pointwise bootstrap resample is not reliable

Options

- Bootstrap resample subvolumes
- Jackknife

Both exhibit "worrying failings" (Norberg et al 2009)

Optimal procedure: generate mock catalogues that look like the data

- Enormous challenge (N body simulations)

See Norberg et al for details

Optimization

Counting all pairs is expensive ($\sim N^2$) for large data sets

Typical optimization: consider only pairs within threshold radius

- Grids
- Binary space partitioning tree (e.g. k-d tree or ball tree)

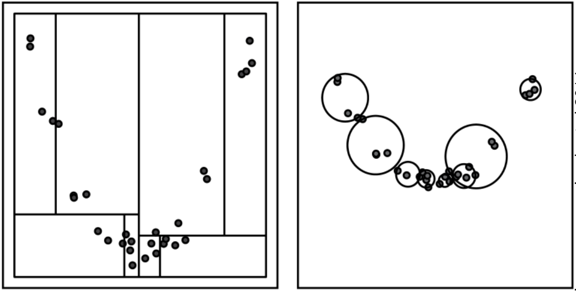


Image source: Ivezic et al 2014

Good for correlation function (small beyond ~30 Mpc)

Not good for power spectrum

Packages

astroML.correlation

- Pre-built pairwise estimator (Landy-Szalay, etc)
- Angular tools
- ~~bootstrap_two_point~~

sklearn.neighbors.KDTree (or BallTree)

- Method two_point_correlation(): counts pairs (write estimator yourself)
- Or query_radius to count yourself

And many more (Corrfunc?)

Convolution methods

$$\xi(\vec{x}_1 - \vec{x}_2) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2) \rangle$$

This is a convolution! Evaluate using FFT in $\mathcal{O}(N \log N)$ time:

$$\hat{\xi}(\vec{r}) = \frac{[(D-R) * (D-R)](\vec{r})}{[R * R](\vec{r})}$$

- Requires binning density field, potentially washing out small scale correlations
- R can be much larger at minimal cost
- Need to reverse one field (convolution vs correlation)

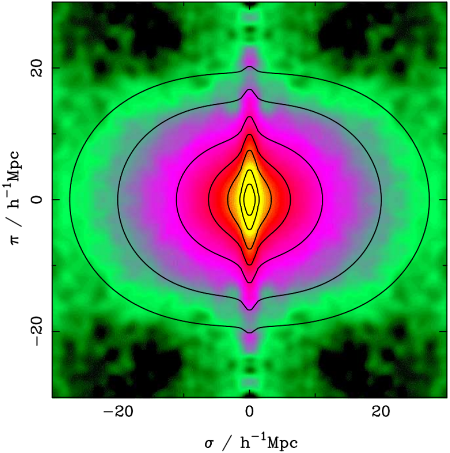
```
scipy.signal.fftconvolve(grid,grid[::-1,::-1,::-1])
```

See Slepian & Eisenstein 2016 for details

Outline

Anisotropic correlations

- In general, function of vector separation: $\xi(\vec{r})$
- Separate correlations in different directions



Peacock et al 2001

Cross correlations

Definition:

$$\xi_{ab}(\vec{x}_1 - \vec{x}_2) = \langle \delta_a(\vec{x}_1)\delta_b(\vec{x}_2) \rangle$$

Same idea: average overlap between field δ_a and field δ_b with some displacement.

Landy & Szalay estimator:

$$\hat{\xi}_{ab}(r) = \frac{D_aD_b(r) - D_aR_b(r) - R_aD_b(r) + R_aR_b(r)}{R_aR_b(r)}$$

Example usage

- Correlations between subset and larger survey
- Correlations between different types of galaxies

3-point correlations

$$\zeta(\vec{x}_1, \vec{x}_2, \vec{x}_3) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2)\delta(\vec{x}_3) \rangle$$

Gives information about the shape of structure.

Useful to write as function of triangle:

- $\zeta(r, q, \psi)$
- $\zeta(r_1, r_2, r_3)$

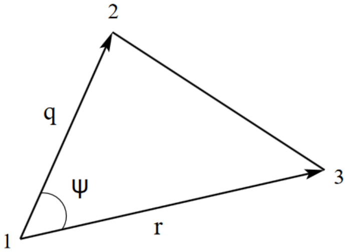


Image source: Takada & Jain 2003

Szapudi & Szalay estimator:

$$\hat{\zeta} = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}$$

3-point correlations

$$\zeta(\vec{x}_1, \vec{x}_2, \vec{x}_3) = \langle \delta(\vec{x}_1)\delta(\vec{x}_2)\delta(\vec{x}_3) \rangle$$

Gives information about the shape of structure.

Useful to write as function of triangle:

- $\zeta(r, q, \psi)$
- $\zeta(r_1, r_2, r_3)$

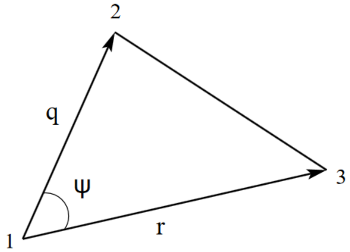


Image source: Takada & Jain 2003

Szapudi & Szalay estimator:

$$\hat{\zeta} = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}$$

$$\hat{\xi}_N = \frac{(D_1 - R_1)(D_2 - R_2)...(D_N - R_N)}{R_1 R_2 ... R_N}$$

Summary

- 2-point (auto)correlation function describes scale of structure
- Typically compute with pairwise estimator
- Computationally intensive, optimizations possible
- Generalization: cross correlations and N-point correlations

A. J. Connolly et al (SDSS Collaboration)
The Angular Correlation Function of Galaxies from Early SDSS Data
Astrophysical Journal, 579(1):42–47, 2002.

D. J. Eisenstein et al
Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies
Astrophysical Journal, 633(2):560–574, 2005.

M. Kerscher, I. Szapudi, and A. S. Szalay
A Comparison of Estimators for the Two-Point Correlation Function
Astrophysical Journal, 535(1):L13–L16, 2000.

P. Norberg, C. M. Baugh, E. Gaztañaga, and D. J. Croton
Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics
MNRAS, 396(1):19–38, 2009.

Z. Ivezic, A. J. Connolly, J. T. VanderPlas, and A. Gray
Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data
Princeton University Press, 2014.

Z. Slepian and D. J. Eisenstein
Accelerating the two-point and three-point galaxy correlation functions using Fourier transforms
MNRAS, 455(1):L31–L35, 2016.

J. A. Peacock et al
A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey
Nature, 410(6825)169–173, 2001.

M. Takada and B. Jain
The three-point correlation function in cosmology
MNRAS, 340(2):580-608, 2003.