How to use statistics to describe the large-scale structure of the universe

David Kirkby, UC Irvine

2 February 2015, SLAC

Statistics in Cosmology

David Kirkby - UC Irvine

LSST Tutorials for Experimental Particle Physicists

2 Feb 2006

"Large-Scale Structure"?

How large is "large"?

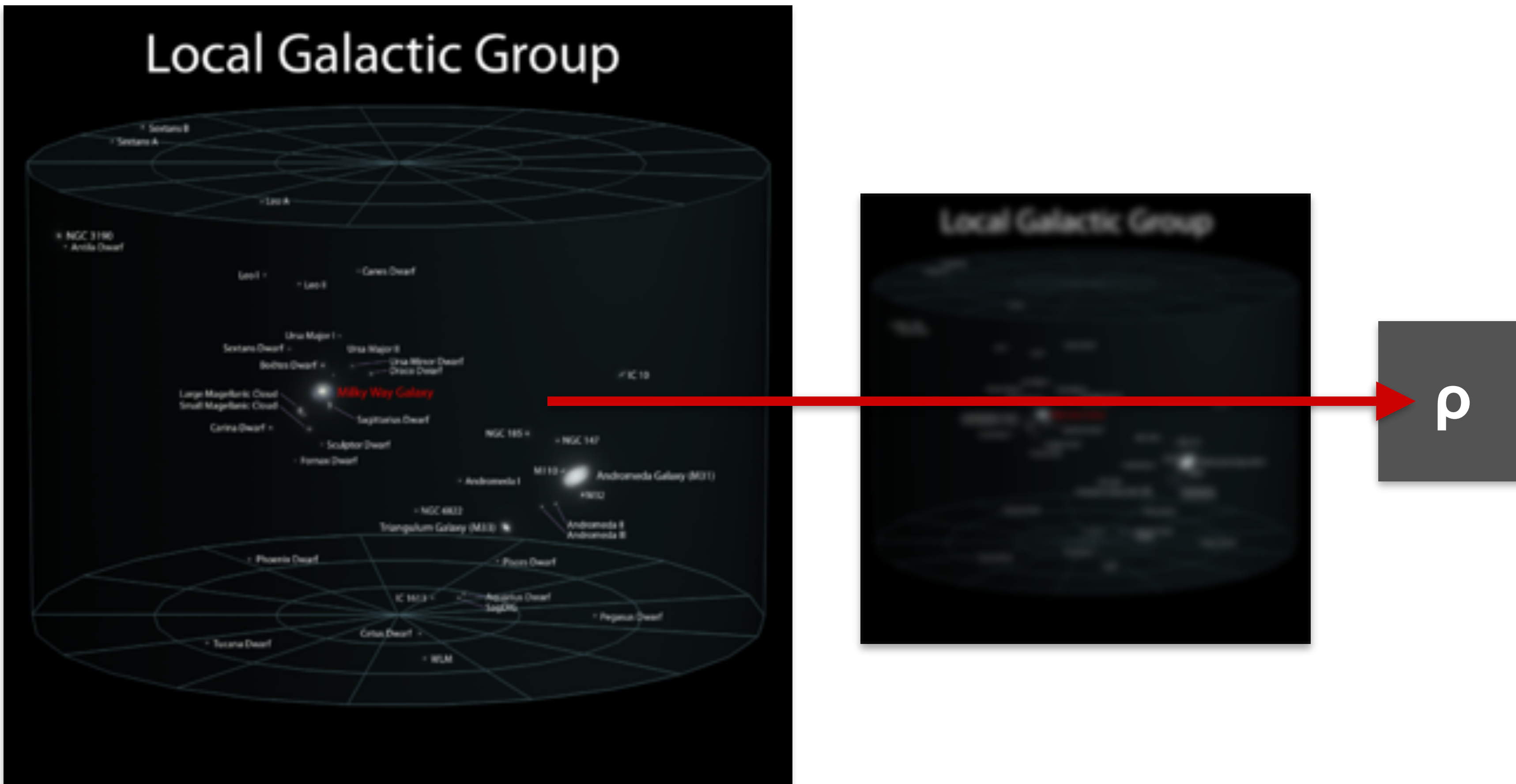
At least 1 Mpc:

- ~3.3 x 10^6 light years
- a galaxy 1 Mpc away is receding at ~70 km/s (corresponding to h = 0.7)

I will always use comoving distances, but will mix Mpc and (the smaller) Mpc/h units.

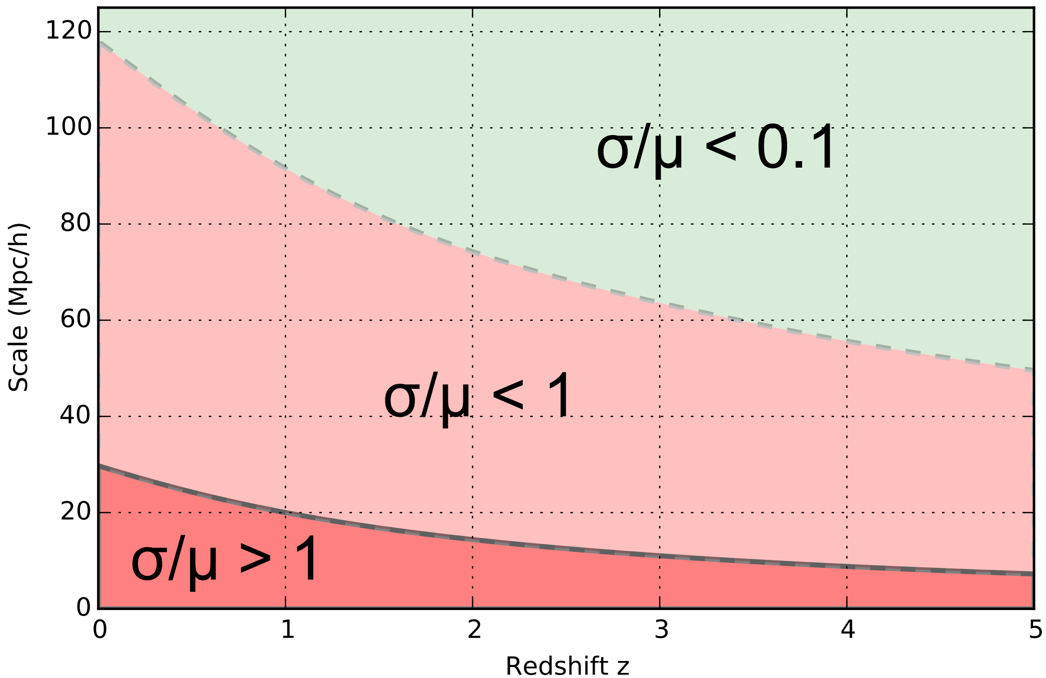
“Large-Scale Structure”?

Our local group is ~1 Mpc across:



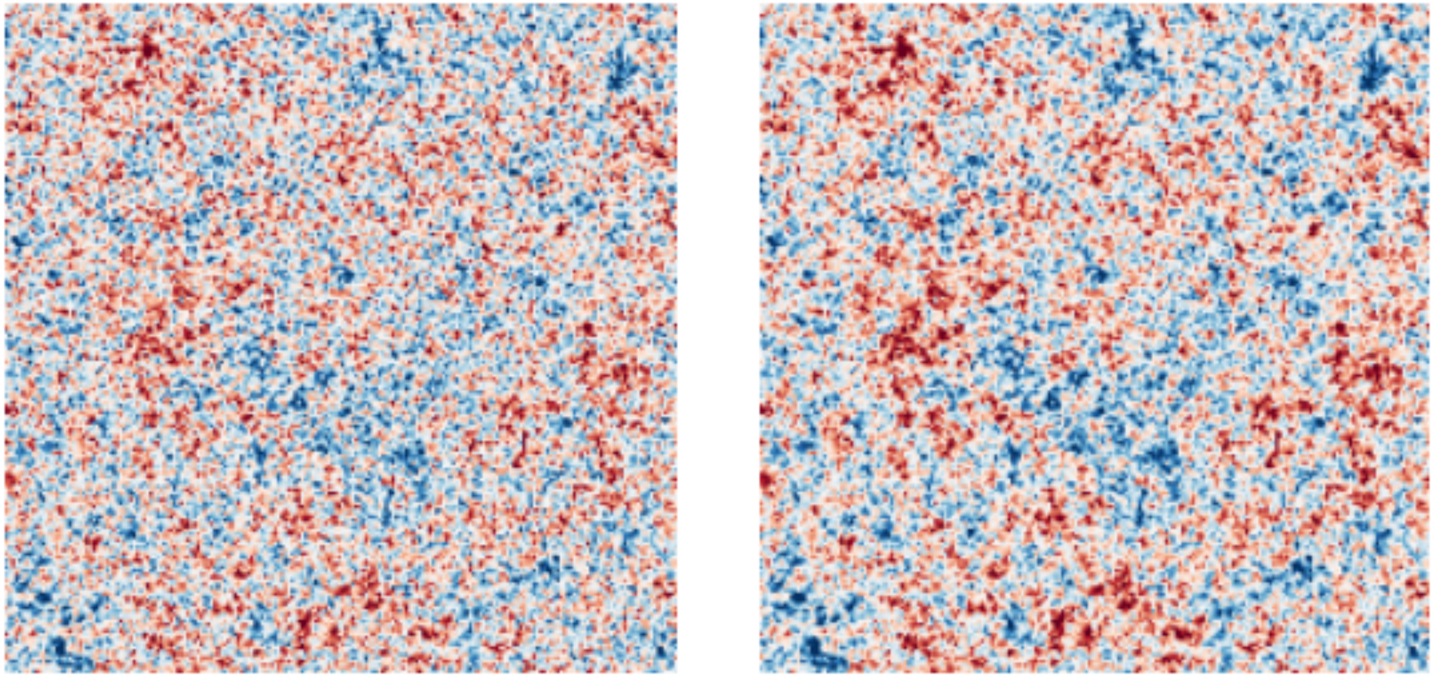
“Large-Scale Structure”?

Statistics of LSS are much easier to interpret on scales where growth of structure can be described with linear perturbation theory.

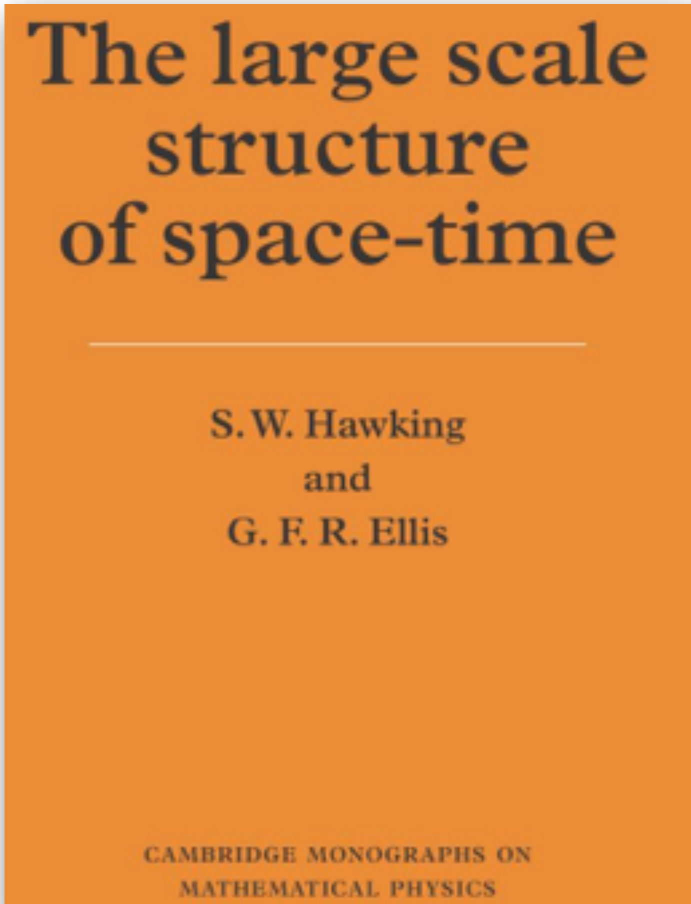


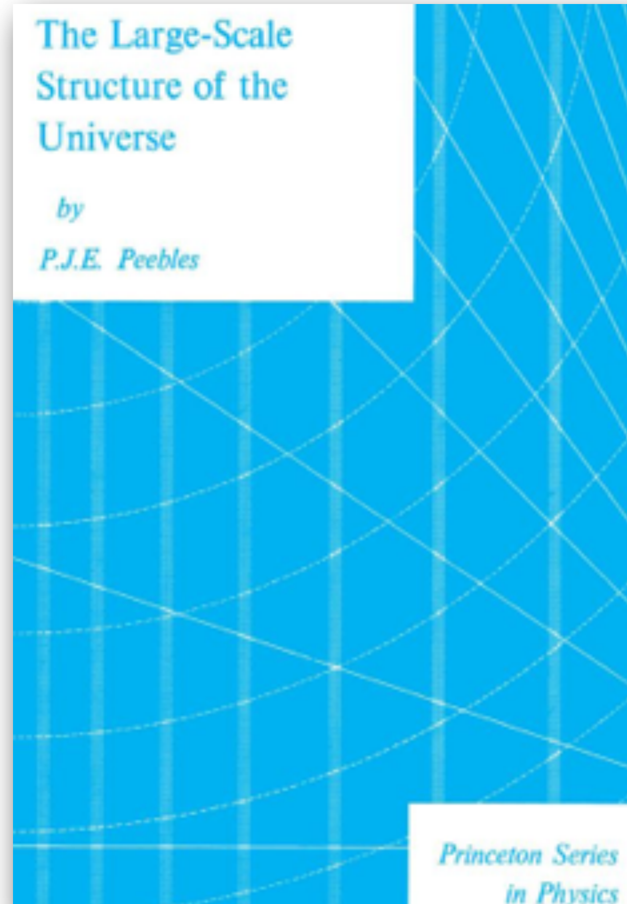
“Large-Scale Structure”?

Where’s the “structure”?



“Large-Scale Structure”?



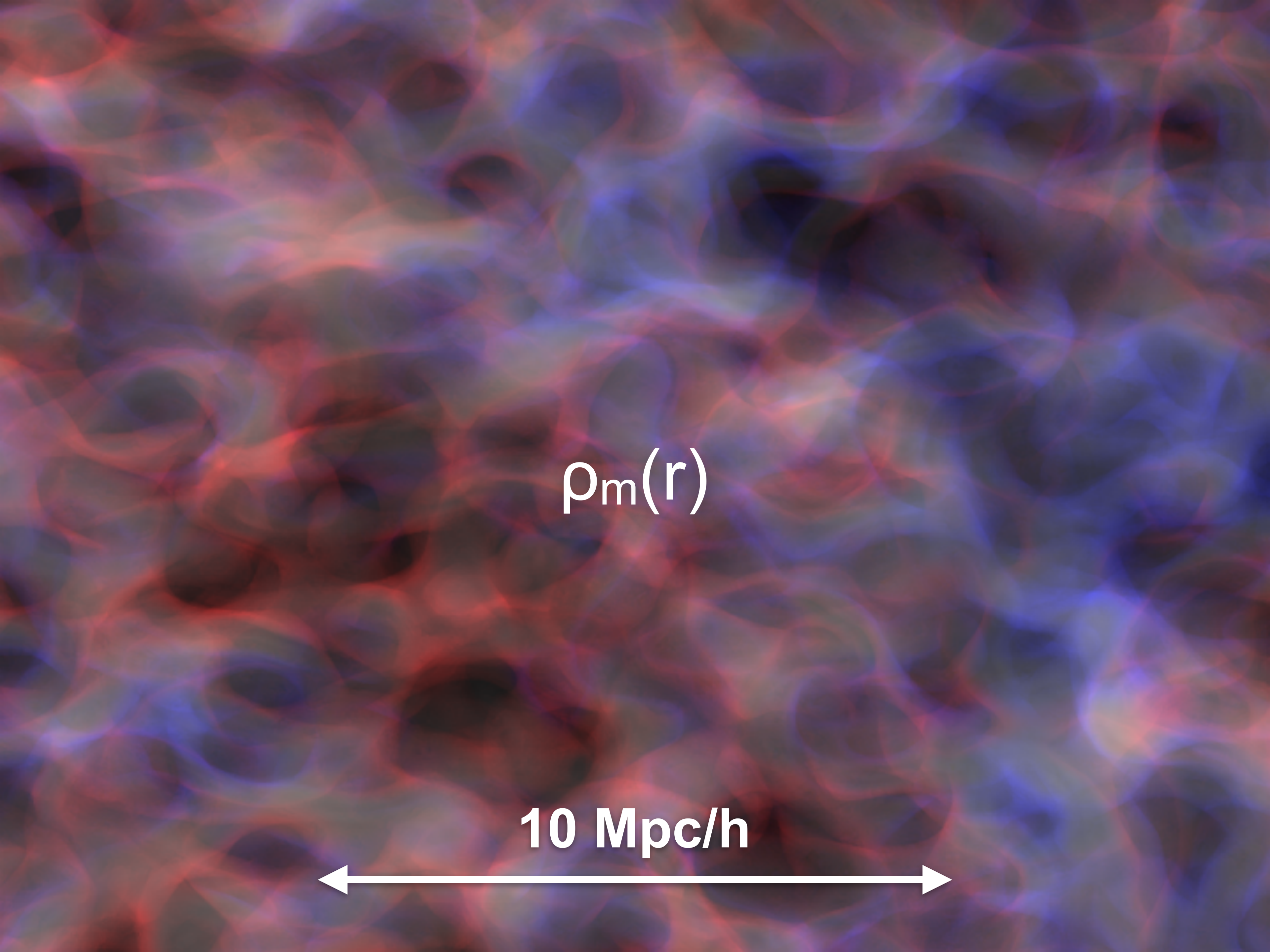


LSS in LSST

LSST will map the locations of ~20B galaxies over 10 years. We will use <u>statistics</u> to analyze the:

- structure of coherent shape distortions (WL)
- imprint of baryon acoustic oscillations (BAO)
- effects of neutrino mass on clustering
- …

The fundamental object in LSS analysis is the (dark + baryonic) matter density field $\rho_m(r)$.


$\rho_m(r)$
10 Mpc/h

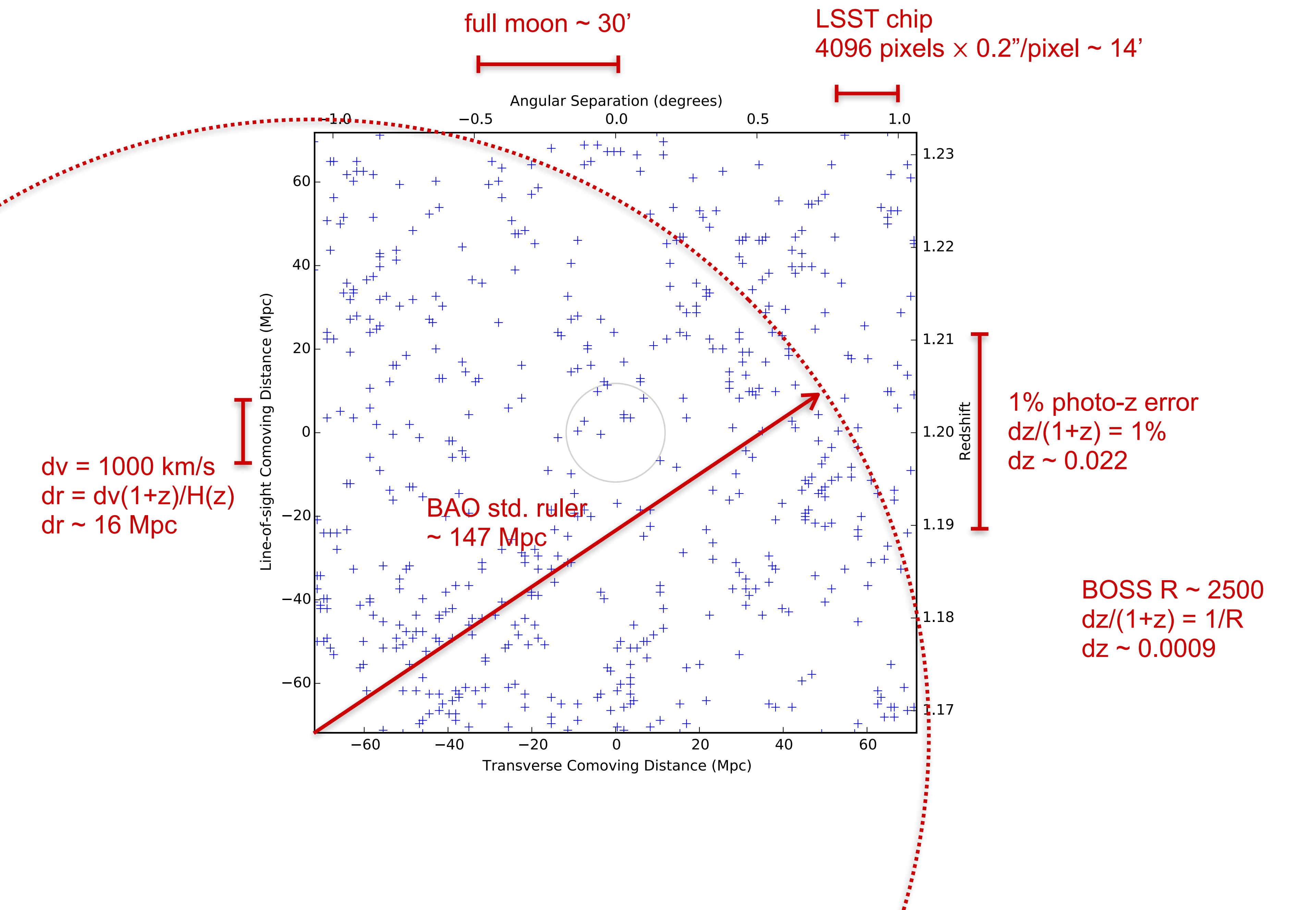
LSS in LSST

2D slice handout:

- how should it be oriented relative to observer?
- how far away should it be held to be in scale?

Draw lines outside axes to indicate sizes of:

- full moon
- LSST chip
- 2% photo-z error
- BOSS spectroscopic resolution (R~2500)
- 1000 km/s relative line-of-sight velocity
- BAO standard ruler



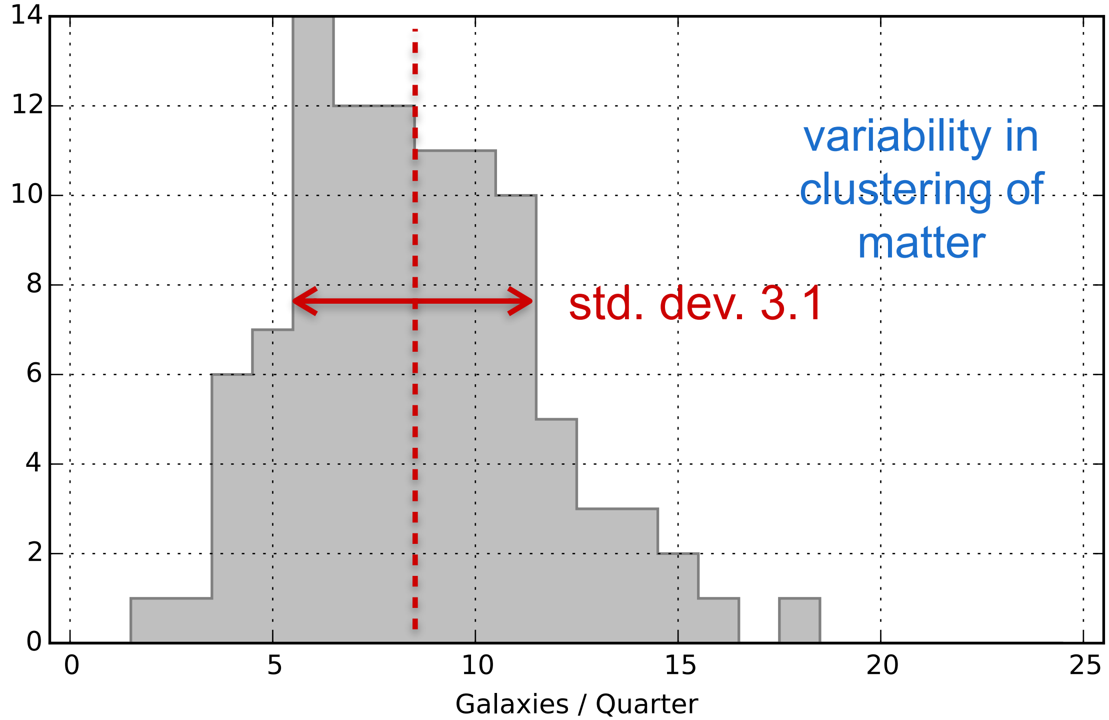
full moon ~ 30'
LSST chip
4096 pixels × 0.2"/pixel ~ 14'
Angular Separation (degrees)
−1.0
−0.5
0.0
0.5
1.0
Line-of-sight Comoving Distance (Mpc)
60
40
20
0
−20
−40
−60
Transverse Comoving Distance (Mpc)
−60
−40
−20
0
20
40
60
Redshift
1.23
1.22
1.21
1.20
1.19
1.18
1.17
dv = 1000 km/s
dr = dv(1+z)/H(z)
dr ~ 16 Mpc
BAO std. ruler
~ 147 Mpc
1% photo-z error
dz/(1+z) = 1%
dz ~ 0.022
BOSS R ~ 2500
dz/(1+z) = 1/R
dz ~ 0.0009

Team Exercise

Use a quarter to sample this toy universe. Record the number of galaxies (+,x) covered by a quarter in different positions.

- Enter your team's data at http://goo.gl/I6tnl1
- Use the column assigned to your team number.
- Is your survey strategy unbiased?

Team Exercise



mean number density 8.5

mean density of (dark + baryonic) matter

Delta Fields

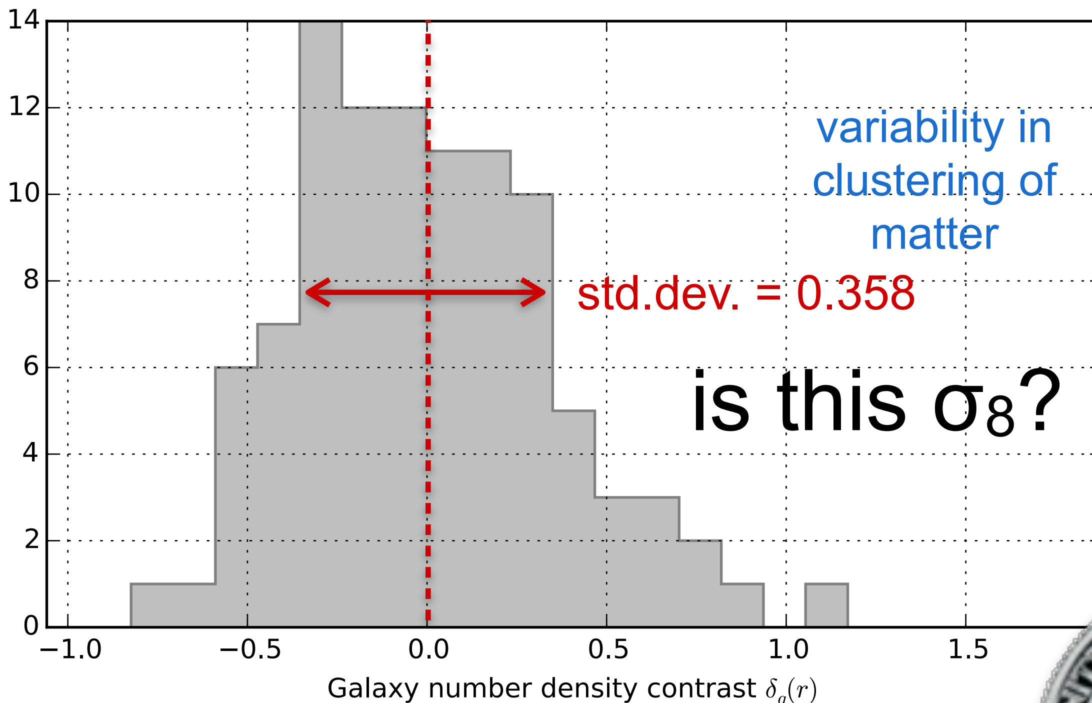
It is convenient to use dimensionless quantities with zero mean (by construction):

Galaxy number density
$$\delta_g(\boldsymbol{r}) = \frac{n_g(\boldsymbol{r}) - \langle n_g \rangle}{\langle n_g \rangle}$$

ensemble averages (over what?)

Matter density
$$\delta_m(\boldsymbol{r}) = \frac{\rho_m(\boldsymbol{r}) - \langle \rho_m \rangle}{\langle \rho_m \rangle}$$

Did we Measure σ_8 ?



mean = 0

~~mean density of (dark + baryonic) matter~~

r = 8 Mpc/h

Sigma8

The quantity σ_8 is defined as the std. deviation of fluctuations in the total matter contained within random hard spheres with radius
$r = 8$ Mpc/h.

Usually calculated assuming linear theory.

The 8 Mpc/h originally chosen to give $\sigma_8 \sim 1$ (not very linear!)

Tracers of Matter Density

Galaxies are "biased" tracers of the underlying matter density delta field (via a ~Poisson process):

$$\delta_g(\boldsymbol{r}) = b_g \delta_m(\boldsymbol{r})$$

Note that we estimate δ_m with number densities, i.e. assuming that all galaxies represent the same amount of (dark + baryonic) matter.

Linear Bias Approximation

A linear bias relationship is simplest possible model, but widely used.

This is wrong in detail, but not too bad on large scales.

Galaxies and quasars generally have bias 1-4 (depending on how they are selected).
Lyman-α forest has bias ~ -0.15 (!)

Sigma8 and bias !

One more wrinkle: delta fields (and therefore b_g and σ_8) evolve with redshift.

Therefore, we actually measured the product:

$$b_g(z)\sigma_8(z) \sim 0.36 \text{ at } z = 1.2$$

Quantities measured from LSS often have this bias normalization uncertainty and so are quoted like this.

Tracer Smoothing

LSS statistics derived from tracers (galaxies, quasars, Lyman-α,…) always involve some smoothing to “large” scales.

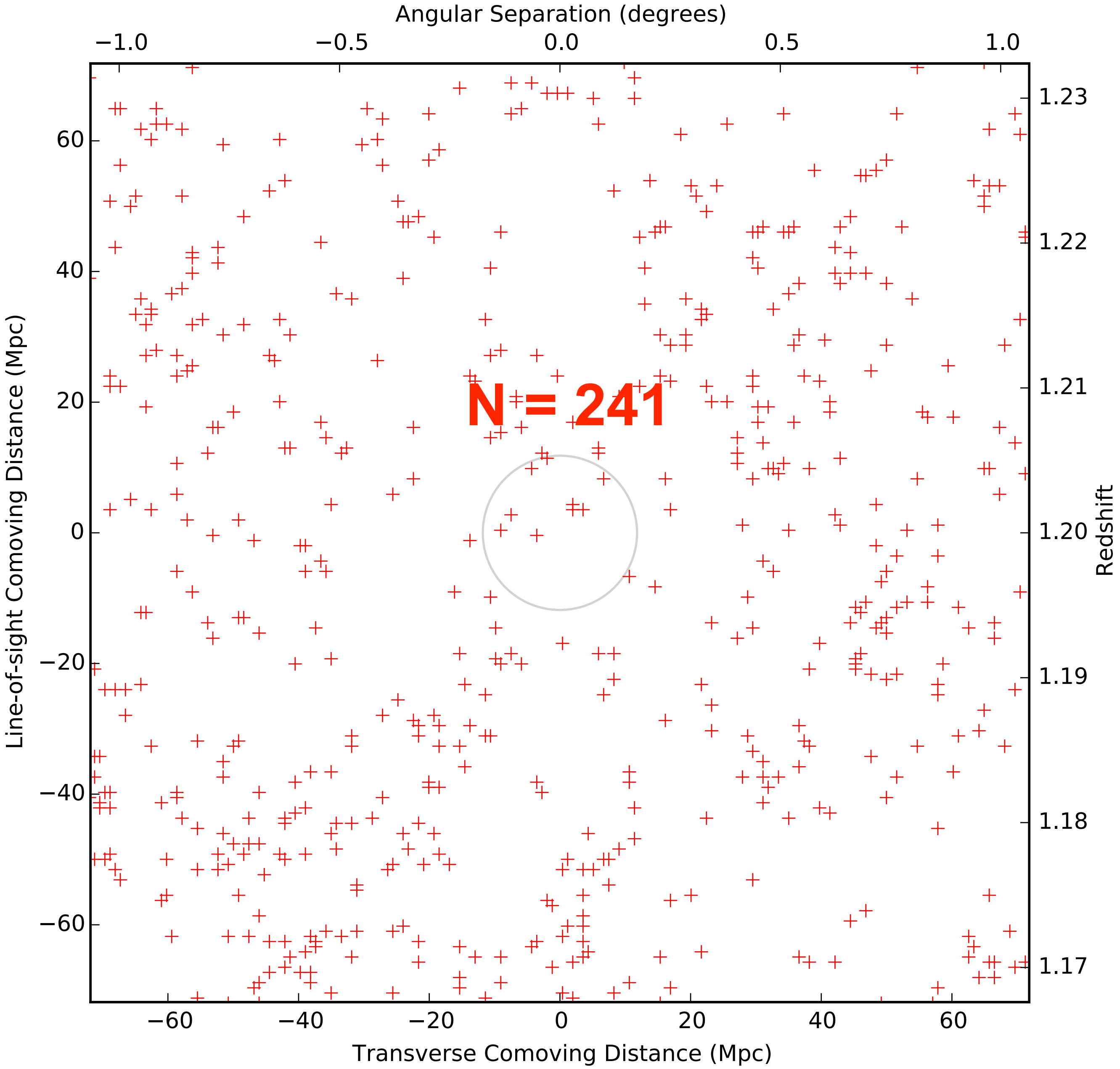
Not always obvious how the smoothing is implemented.

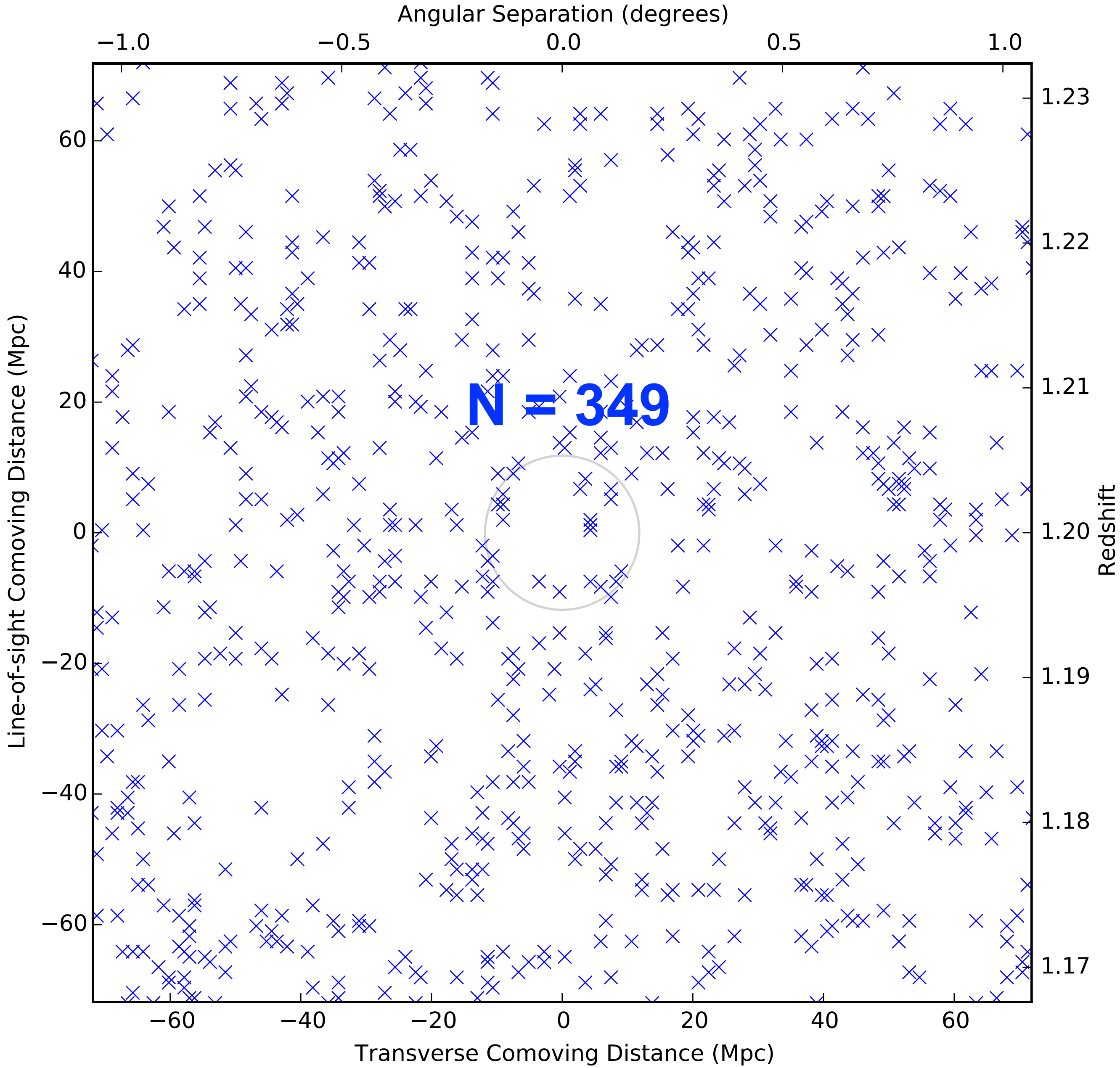
In our exercise, we smoothed over the area of a quarter ($r = 8$ Mpc/h).

Very Large-Scale Modes

There are 2 versions (red/blue) of the handout.

Are they drawn from the same underlying matter density distribution?


Angular Separation (degrees)
−1.0
−0.5
0.0
0.5
1.0
60
40
20
0
−20
−40
−60
Line-of-sight Comoving Distance (Mpc)
1.23
1.22
1.21
1.20
1.19
1.18
1.17
Redshift
N = 241
−60
−40
−20
0
20
40
60
Transverse Comoving Distance (Mpc)

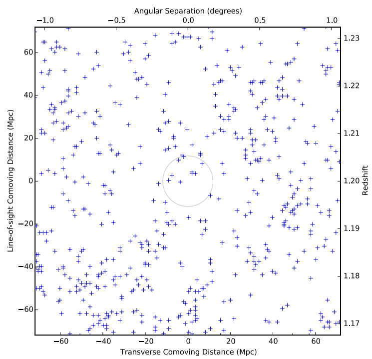


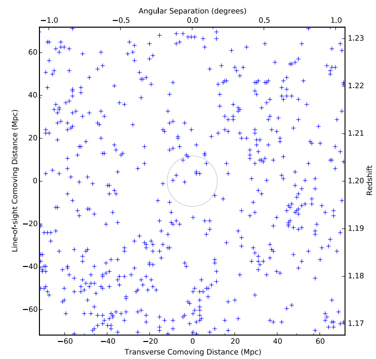
Angular Separation (degrees)
−1.0
−0.5
0.0
0.5
1.0
N = 349
60
40
20
0
−20
−40
−60
Line-of-sight Comoving Distance (Mpc)
1.23
1.22
1.21
1.20
1.19
1.18
1.17
Redshift
−60
−40
−20
0
20
40
60
Transverse Comoving Distance (Mpc)

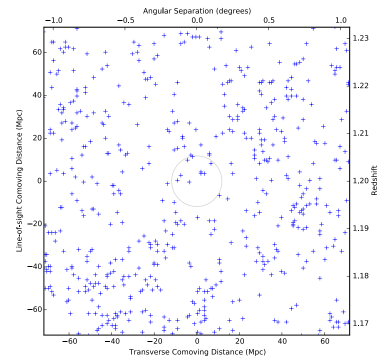
Cosmological Principle

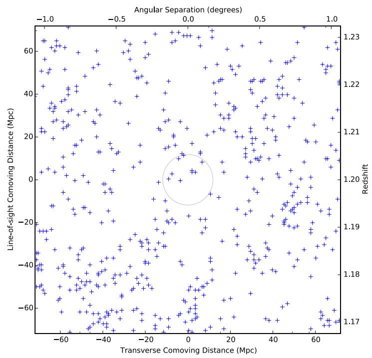
Different locations in the same universe

Same location in different universes



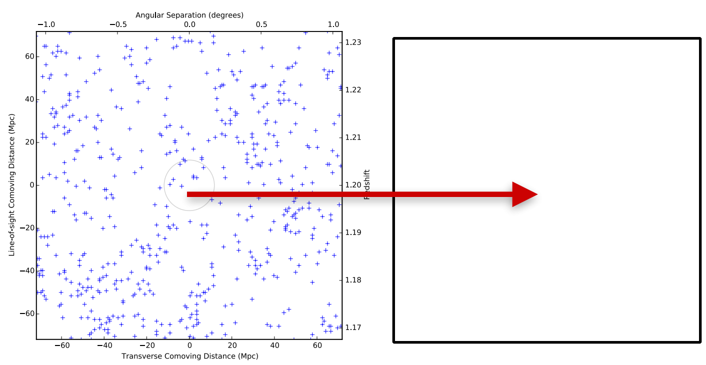


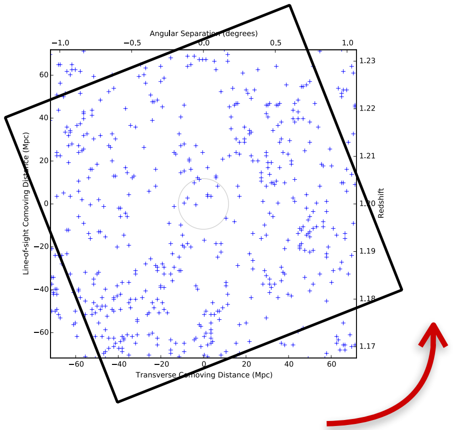




Cosmological Principle

Given just one universe, we assume that the <u>statistics</u> of the matter density field are invariant under translation & rotation.



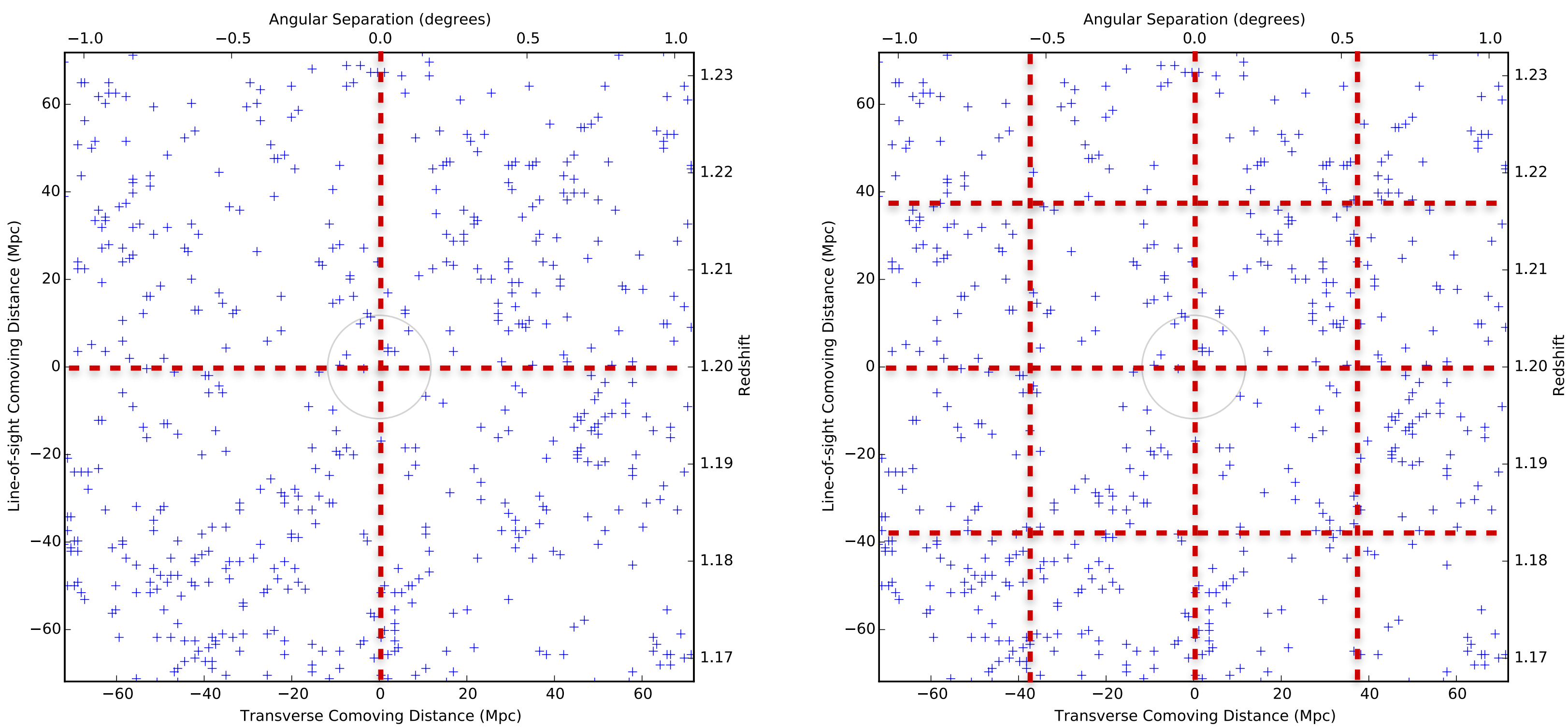


statistically homogeneous universe

statistically isotropic universe

Cosmic Variance

A very large smoothing scale does not allow many independent measurements.



Cosmic Variance

A very large smoothing scale does not allow many independent measurements.

Cosmic-variance limit for measuring $b_g(z)\sigma_8(z)$ in the exercise:

- fractional error on std.dev. ~ $(2(n-1))^{-1/2}$
- max. number of independent samples ~ 50
- cosmic-variance limit is ~10% error

Co-variance

σ_8 and friends are measures of variance. Because we assume homogeneity and isotropy, variances do not depend on position.

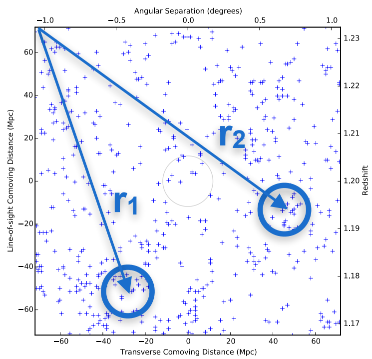
More generally, we are interested in covariances $\xi(r_1,r_2)$.

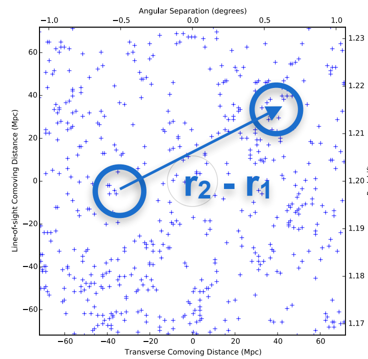
$\xi(r_1,r_2)$ answers the question:
given $\delta(r_1)$, what is your best guess for $\delta(r_2)$?

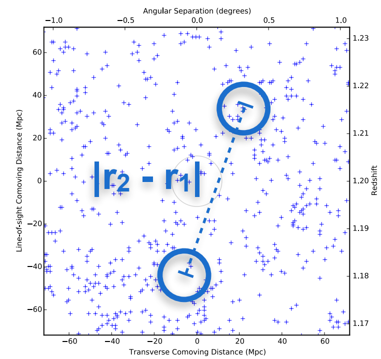
Co-variance

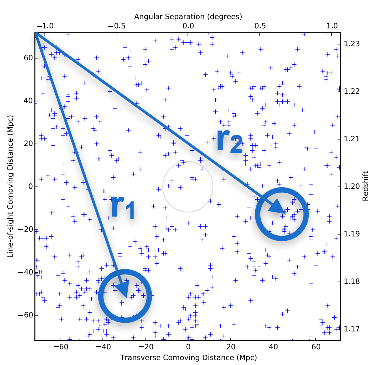
Different locations in the same universe

Same location in different universes









$\xi(r_1,r_2)$

$\xi(r_2-r_1)$

statistically homogeneous universe

AND

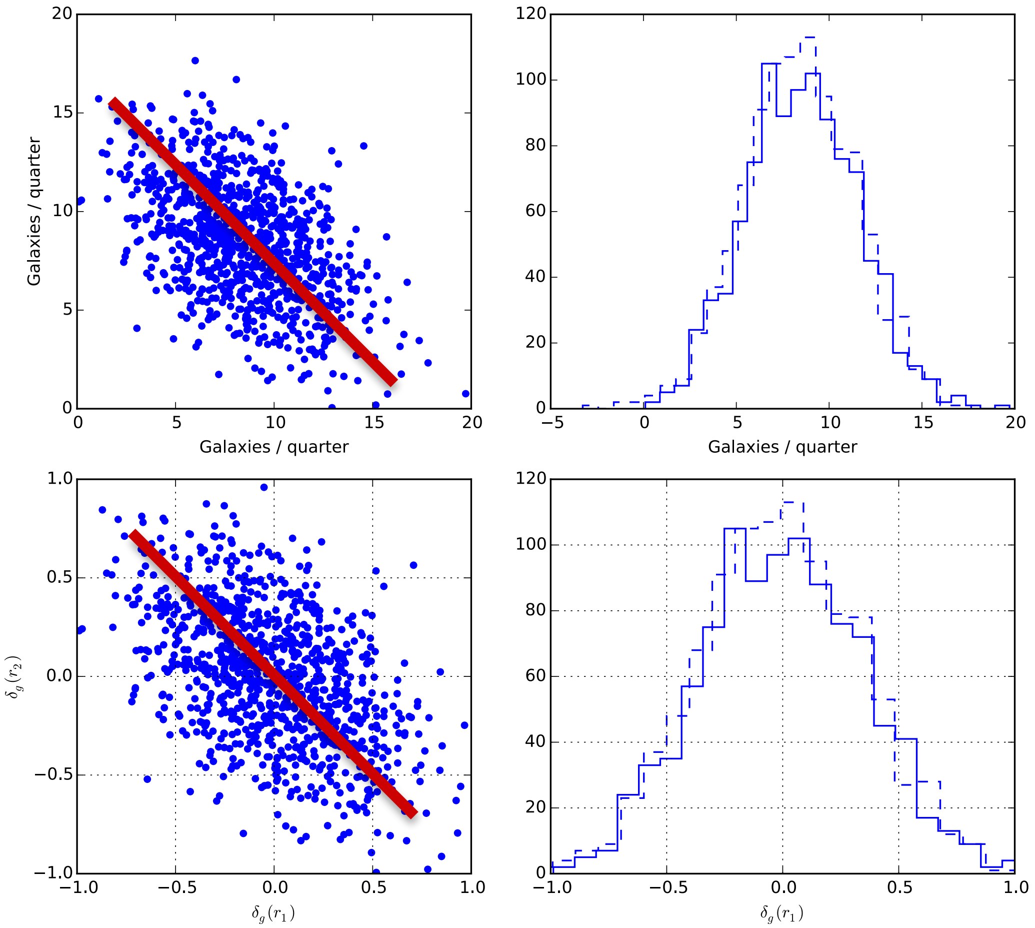
$\xi(|r_2-r_1|)$

statistically isotropic universe

Correlation Coefficient

The 2D scatter plot introduces a new piece of information that could not be deduced from the 1D histograms: the dimensionless correlation coefficient $-1 \leq \rho \leq +1$.

(another ρ !)



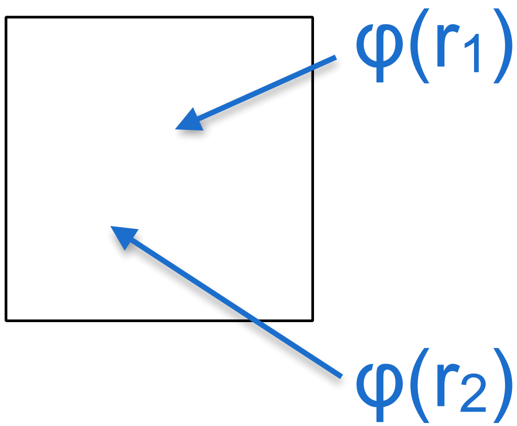
Random Fields

The cosmological matter density in our universe is a single realization of some underlying random field.

What does this mean??

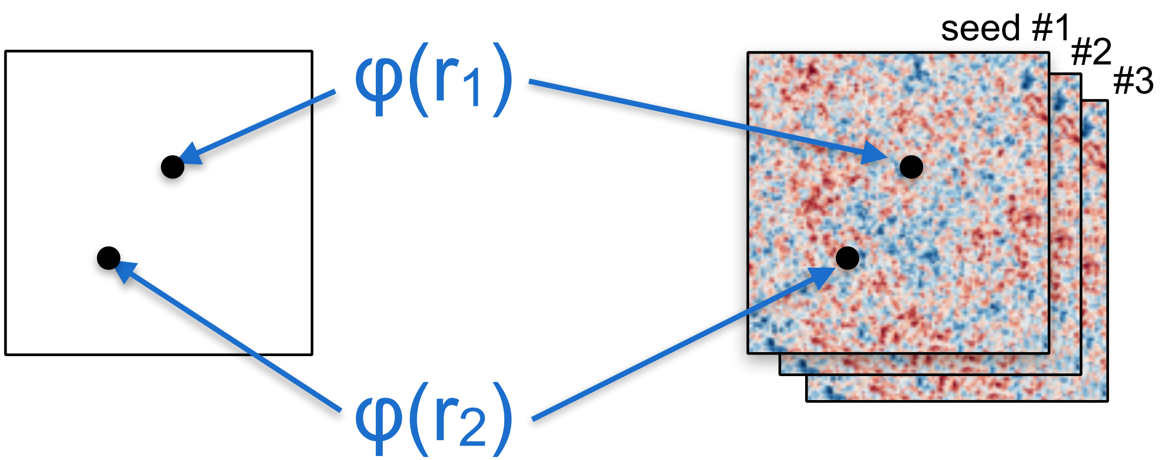
Random Fields

A "random field" is an infinite set of random-number generators $\varphi(r)$ indexed by their position in space r.



Random Fields

When each generator produces a single random number, the result is a realization. Many realizations are possible.



Random Fields

The infinite set of 1D generator distributions does not completely specify the field.

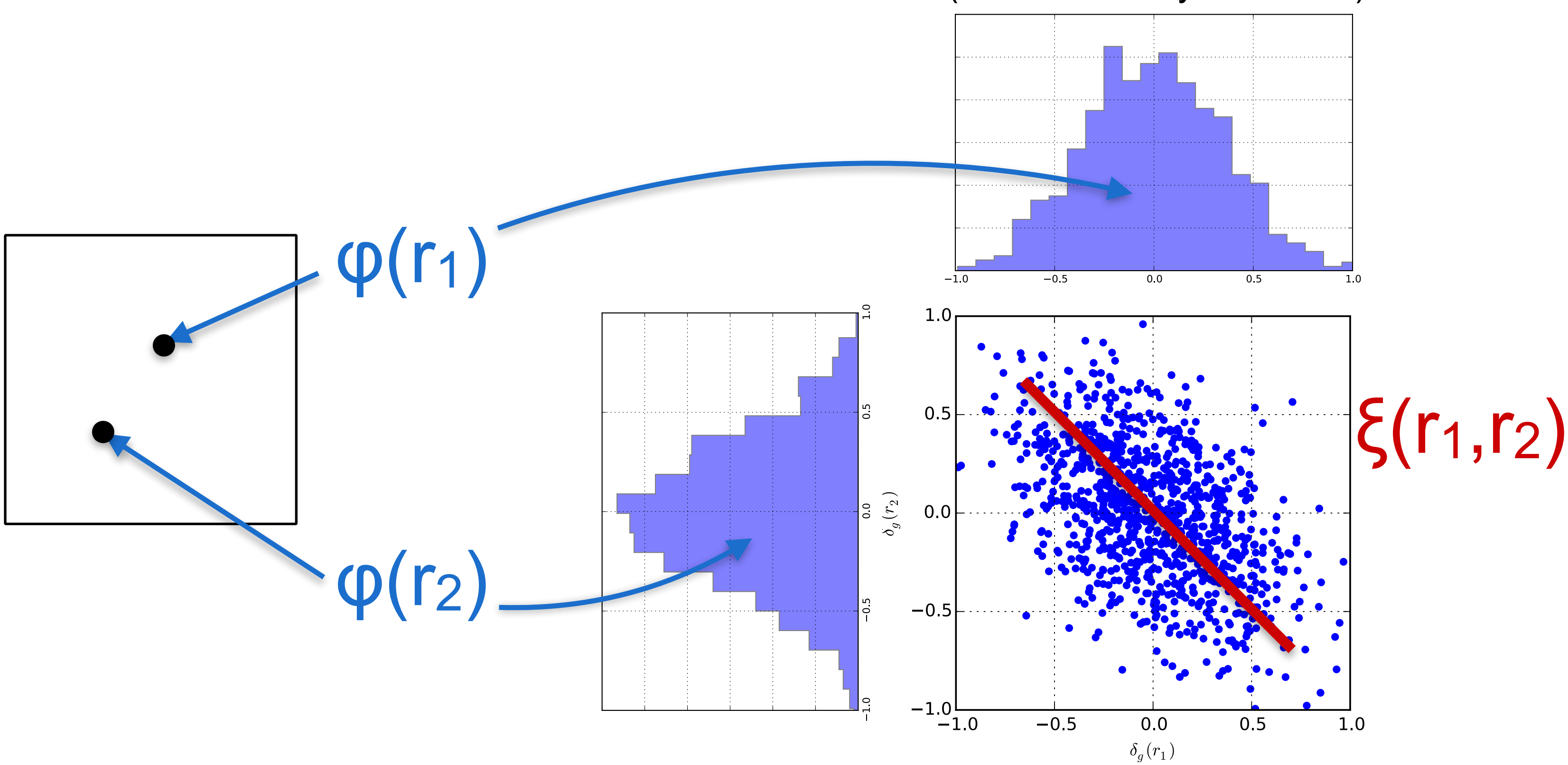
The covariance $\xi(r_1,r_2)$ is an arbitrary (positive-definite) function that specifies the correlated behavior of pairs of generators.

$\xi(r_1,r_2)$ is often (confusingly) called the correlation function.

Random Fields

Generators must talk to each other to ensure that correlations are included:

(not necessarily Gaussian)



Random Fields

A general random field requires that all n-point joint probability distributions be specified also!

- requires an infinite number of additional functions to be specified
- knowing all n-point correlations does not allow you to predict (n+1)-point correlations

Gaussian Random Fields

We have already met two major simplifications of the matter density field:

- statistical homogeneity
- statistical isotropy

A third major simplification is that our universe is the realization of a Gaussian random field

- a good approximation on large scales
- this is due to physics: inflation initial conditions & linear growth of inhomogeneities

Gaussian Random Field

The n-point joint prob. distributions of a GRF are, by definition, all multivariate Gaussians.

All n-point statistics are fully specified by the mean $\mu(r)$ and covariance functions $\xi(r_1,r_2)$.

Homogeneous & isotropic simplifications:

- $\mu(r) \Rightarrow \mu$ equals 0 for $\delta(r)$
- $\xi(r_1,r_2) \Rightarrow \xi(|r_2-r_1|)$ dimensionless for $\delta(r)$

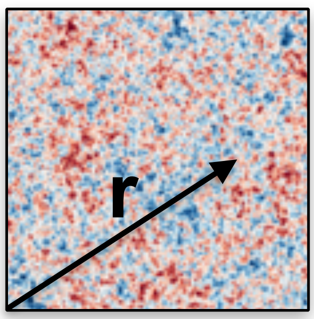
Hilbert Space

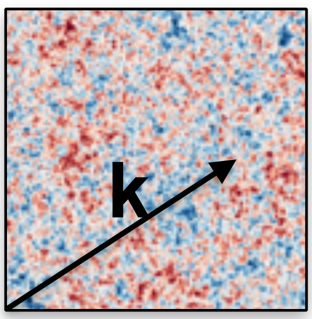
The delta field of a physical quantity varying in space is a physical thing, independent of any representation.

Useful to represent a delta field as an (infinite-dimensional) vector **$\boldsymbol{\delta}$** in a Hilbert space.

Hilbert Space

One possible representation is in (x,y,z) space:

δ = 

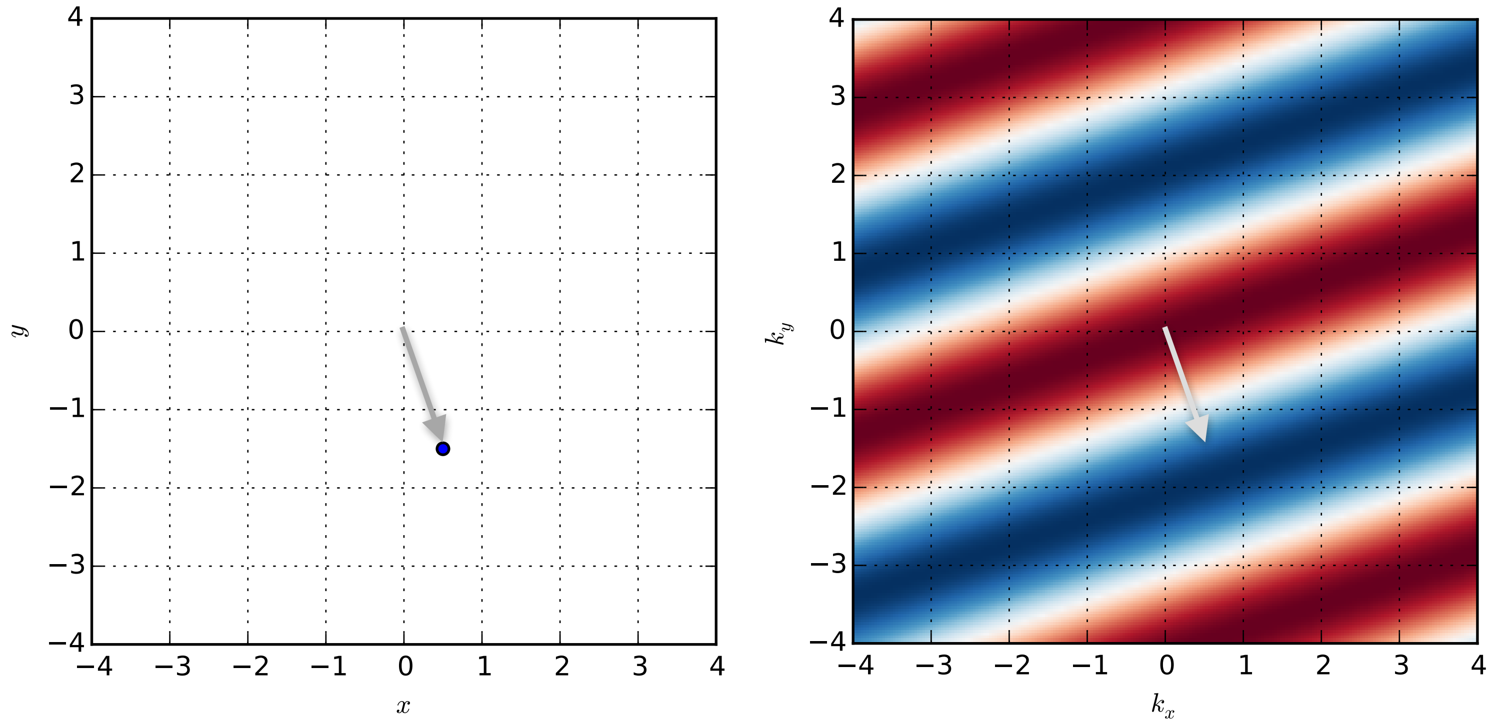
δ = 

An equally valid representation uses the conjugate Fourier transform space (k_x, k_y, k_z).

Basis Vectors

Basic vectors are related via Fourier transforms

$$\delta_D(\boldsymbol{r} - \boldsymbol{r}_0) \qquad \exp(i\boldsymbol{k} \cdot \boldsymbol{r}_0)$$



Covariance Matrix

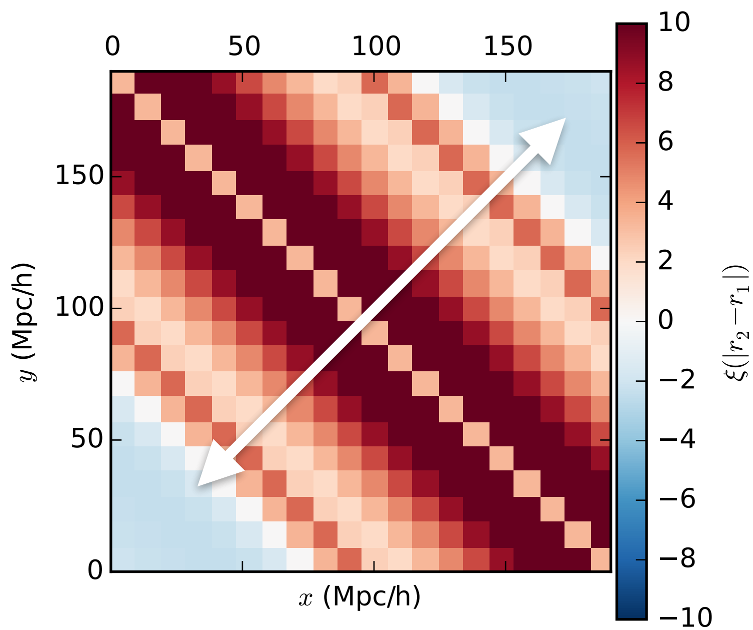
The probability (density) for observing *n* samples of a Gaussian random field **δ** is:

$$\mathrm{prob}(\boldsymbol{\delta}) = (2\pi)^{-n/2}|C|^{-1/2}\exp\left(\boldsymbol{\delta}^\dagger C^{-1}\boldsymbol{\delta}\right)$$

covariance matrix C:
can be represented in
any basis, like **δ**.

Covariance Matrix

The covariance matrix is fully specified by $\xi(|r_2-r_1|)$ in the r-space basis, e.g.



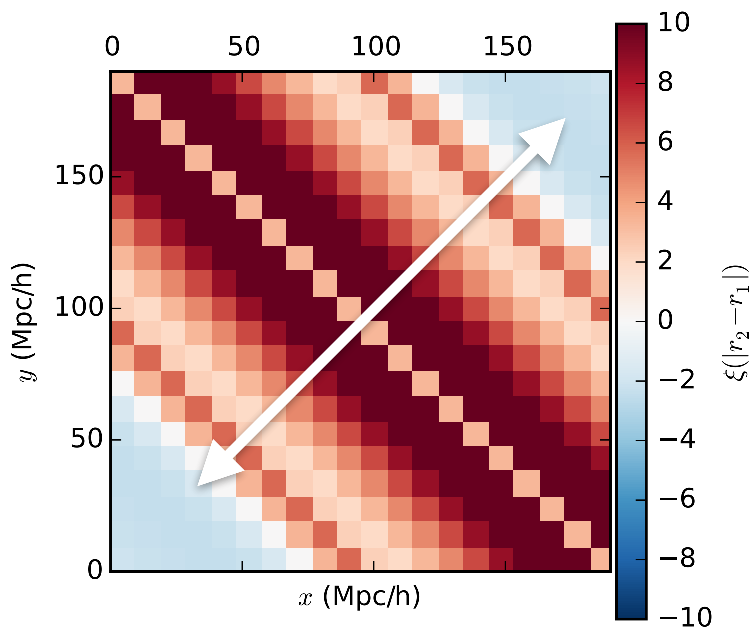
$\xi(|r_2-r_1|)$ values are dimensionless & can be negative.

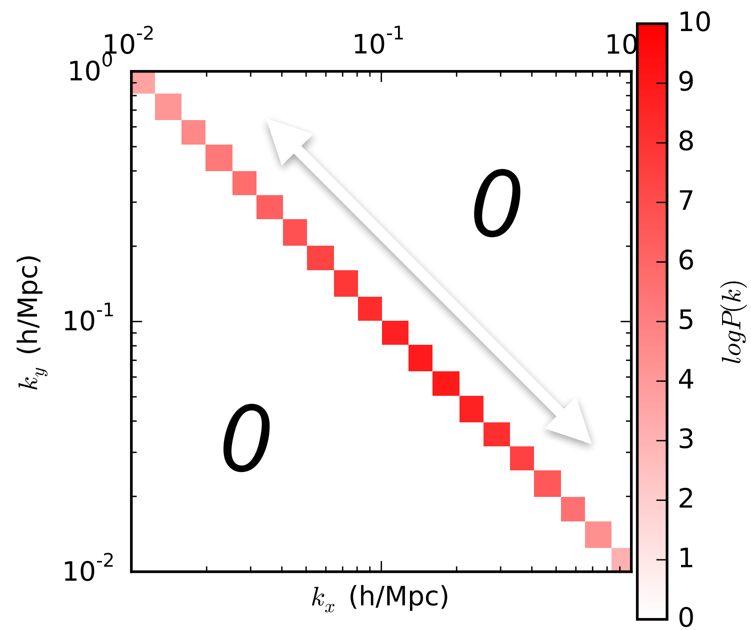
Often write (confusingly) $\xi(r)$ with $r = |r_2-r_1|$.

Matrix is dense but has lots of redundancy.

Covariance Matrix

The same covariance matrix in the k-space basis is diagonal!



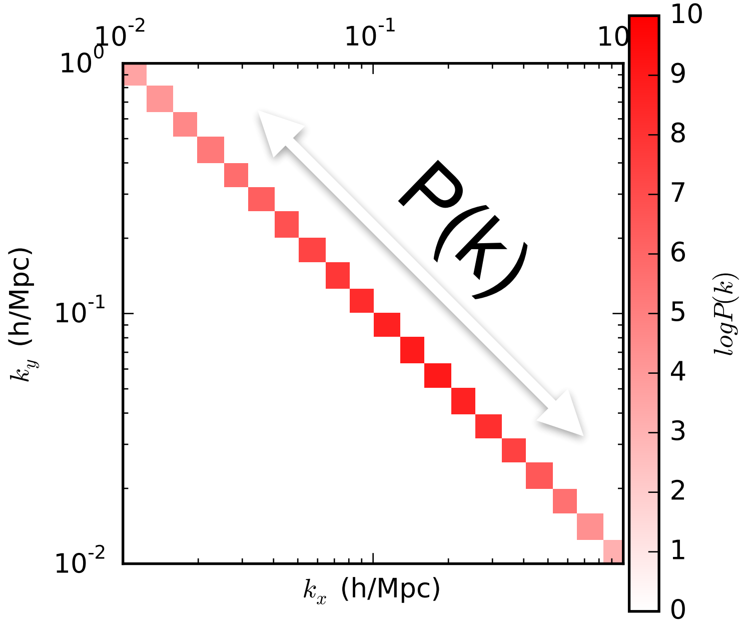


Variations are now along diagonal, instead of transverse.

Power Spectrum

The values along the diagonal are the power spectrum P(k).

P(k) values have dimensions of 1/length^3 and must be ≥ 0.



P(k) vs ξ(r)

Diagonal covariance in k-space basis means that different k-modes are statistically uncorrelated. Is P(k) therefore better than ξ(r)?

Yes, in simplest cases, but benefits are less clear in a real analysis due to:

- redshift space distortions (z vs. Δθ)
- survey boundaries
- non-linear growth of structure

Most recent LSS analyses primarily use ξ(r).

P(k) and ξ(r)

P(k) and ξ(r) are same mathematical object, expressed in different bases.

There is no more or less information in one representation than the other.

We generally want to know the errors on P(k) or ξ(r) (covariances of covariances!) This often drives practical choices between P(k) and ξ(r).

P(k) and ξ(r)

Elements of the covariance matrix are defined as averages ⟨…⟩ over a hypothetical ensemble of realizations (universes!) of the same GRF.

In practice, we average over positions and orientations instead:

$$C = \langle \boldsymbol{\delta}\boldsymbol{\delta}^\dagger \rangle$$

coordinate-free definition

r-space

$$\xi(r) = \langle \delta(\boldsymbol{r}_1)\delta(\boldsymbol{r}_2) \rangle_{|\boldsymbol{r}_2 - \boldsymbol{r}_1| = r}$$

$$(2\pi)^3 \delta_D(\boldsymbol{k}_1 + \boldsymbol{k}_2) P(k) = \langle \delta(\boldsymbol{k}_1)\delta(\boldsymbol{k}_2) \rangle_{|\boldsymbol{k}_i| = |\boldsymbol{k}_j| = k}$$

k-space

Fourier Transforms

Use Fourier transforms to convert between r- and k-space representations:

n-dims:

$$P(\boldsymbol{k}) = \int \xi(\boldsymbol{r}) e^{+i\boldsymbol{k}\cdot\boldsymbol{r}} d^n r$$

$$\xi(\boldsymbol{r}) = \int P(\boldsymbol{k}) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} \frac{d^n k}{(2\pi)^n}$$

These establish our Fourier conventions

Angular integrals can be performed explicitly:

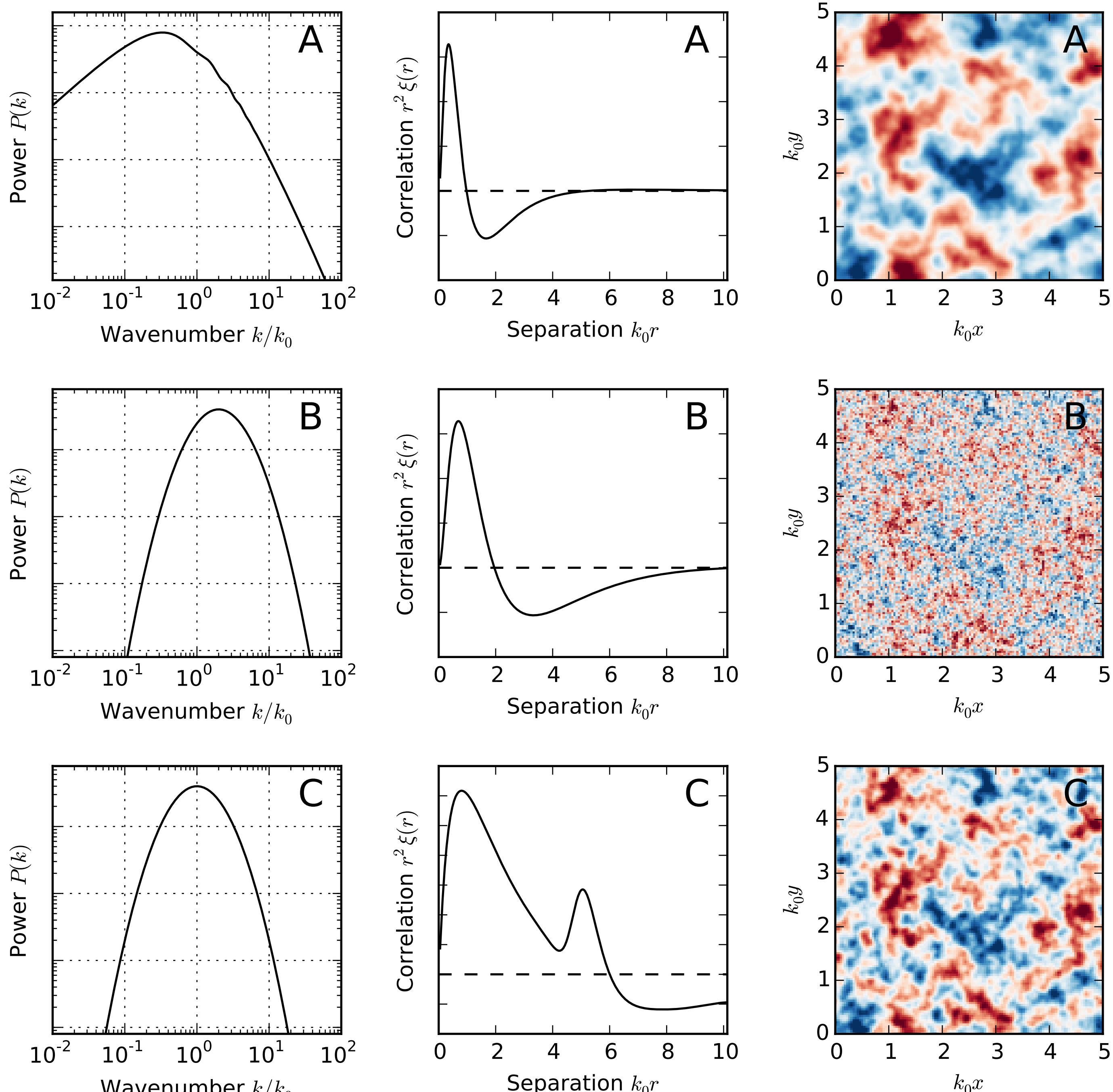
$$\xi(r) = \int_0^\infty \frac{dk}{(2\pi)^n} k^{n-1} P(k) \begin{cases} 2\cos(kr) & n=1 \\ 2\pi J_0(kr) & n=2 \\ 4\pi j_0(kr) & n=3 \end{cases}$$

$$j_0(kr) = \frac{\sin(kr)}{kr}$$

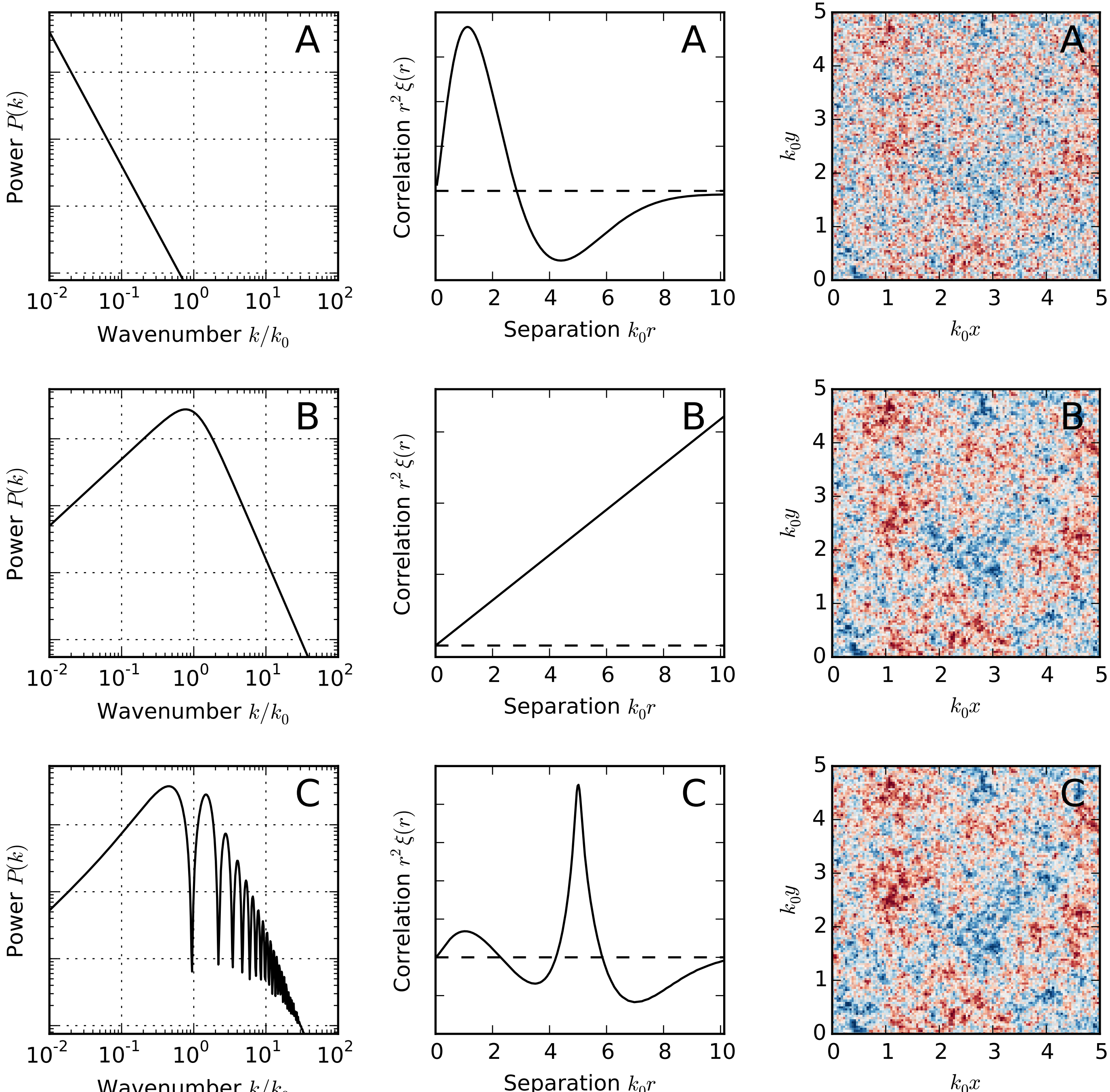
Think, Discuss…

Match each P(k) to its corresponding ξ(r) and r-space realization δ(r).

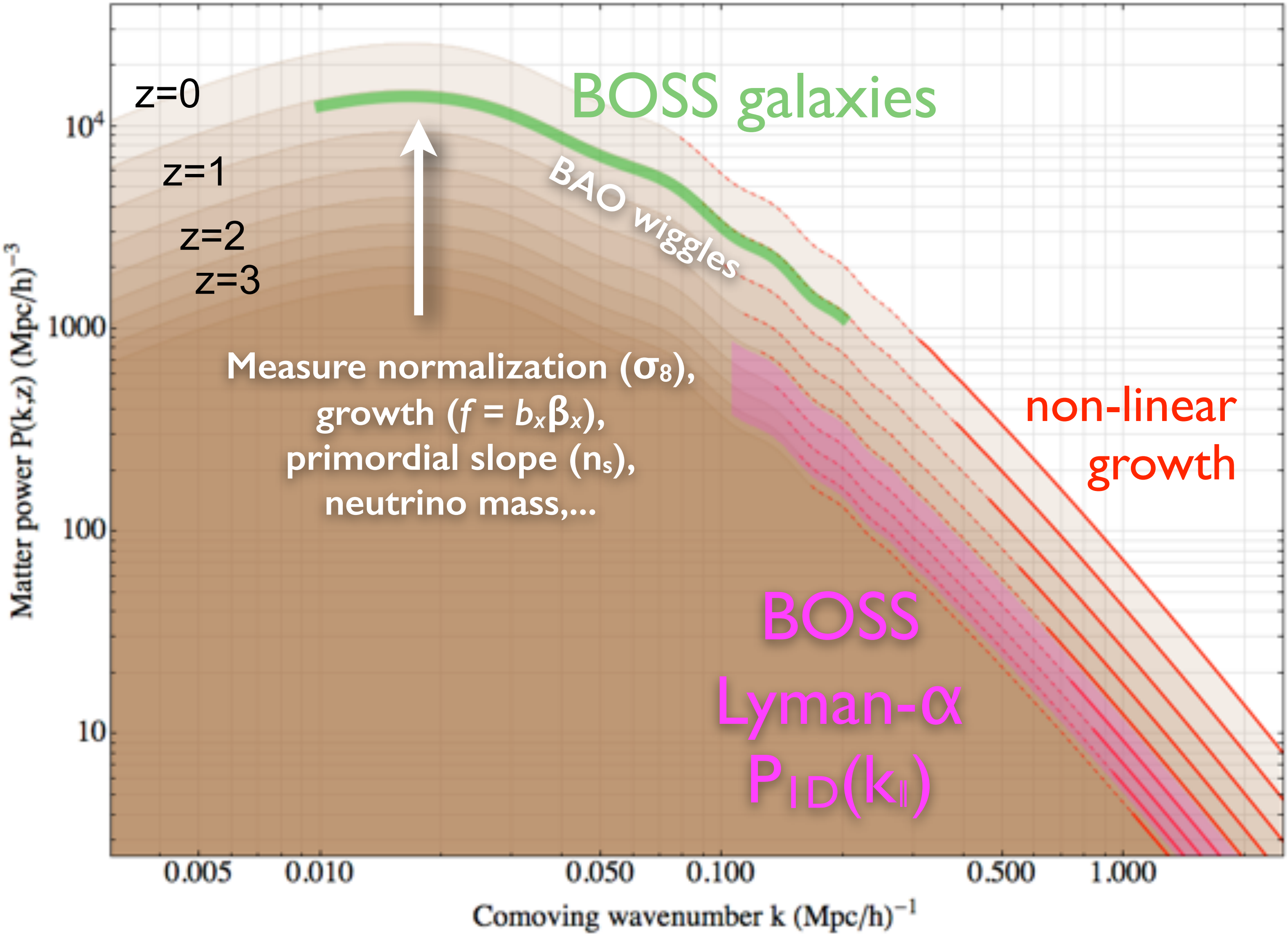
- All axes are dimensionless, in terms of k_0
- P(k) plots are log-log
- ξ(Δr) plots are linear, but r^2 weighted
- density fields are 2D slices of 3D realizations



A
B
C
Power $P(k)$
Wavenumber k/k_0
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Correlation $r^2\xi(r)$
Separation k_0r
0 2 4 6 8 10
k_0y
k_0x
0 1 2 3 4 5


A
B
C
Power $P(k)$
Wavenumber k/k_0
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Correlation $r^2\xi(r)$
Separation k_0r
0 2 4 6 8 10
k_0y
k_0x
0 1 2 3 4 5

The ΛCDM Power Spectrum



Sigma8 Revisited

LSST DESChool
Dark Energy Science Collaboration

P(k) has units of 1/length^3 in 3D, but:

$$\Delta^2(k) \equiv \frac{1}{2\pi^2} k^3 P(k) = \frac{d\sigma^2}{d\log k}$$

is dimensionless, and measures contributions to the variance from different scales:

$$\sigma^2 = \langle \delta(k)^2 \rangle$$

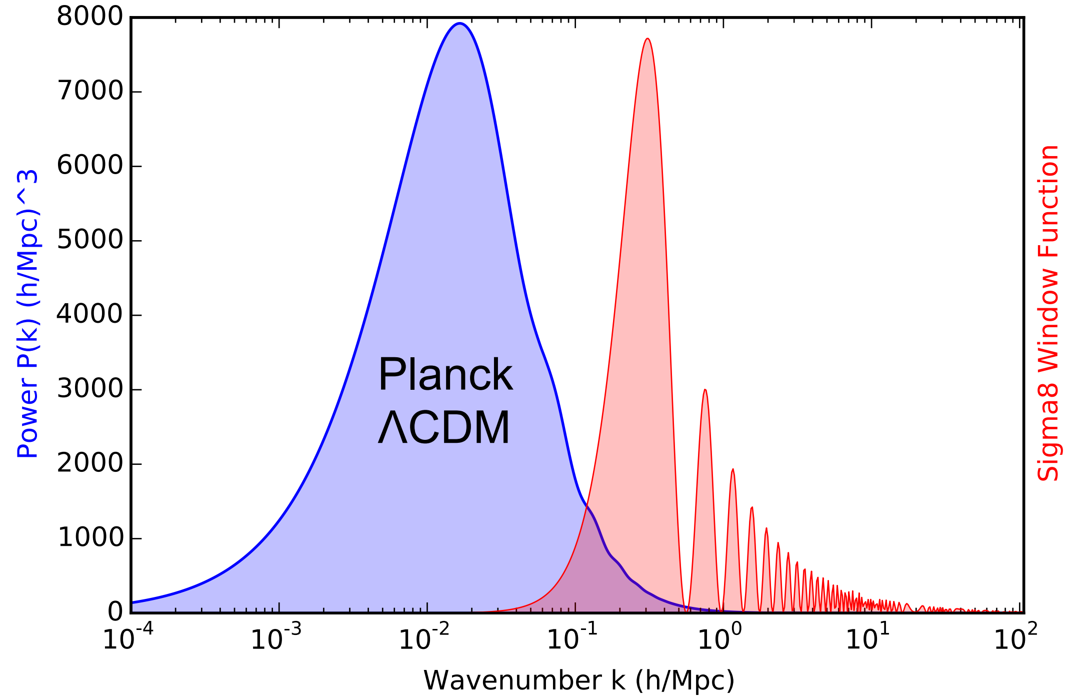
The total variance is just ξ(0) since:

$$\xi(r) = \frac{1}{4\pi^2} \int_0^\infty dk\, k^2 \frac{\sin(kr)}{kr} P(k)$$

Window Functions

LSST DESChool
Dark Energy Science Collaboration

Any smoothing scheme can be represented using a normalized (r- or k-space window function. For a hard sphere of radius 8 Mpc/h:



σ_8 measures power over a range of scales, but mostly 0.1 - 1.0 h/Mpc.

Glossary

- large-scale structure
- tracer
- bias
- sigma8
- power spectrum
- correlation function
- k-mode
- isotropic
- homogeneous
- cosmic variance
- r- and k-space
- baryon acoustic oscillations (BAO)
- covariance
- correlation coefficient
- (Gaussian) random field
- realization
- covariance matrix
- window function
- delta field
- Hilbert space
- cosmological principle

Recommended Reading

A.J.S. Hamilton, “Power Spectrum Estimation I. Basics”, arXiv:astro-ph/0503603 (2005).

A.J.S. Hamilton, “Power Spectrum Estimation II. Linear Maximum Likelihood”, arXiv:astro-ph/0503604 (2005).

M. Tegmark *et al*, “Karhunen-Loève Eigenvalue Problems in Cosmology: How should we Tackle Large Data Sets?”, arXiv:astro-ph/9603021 (1996).

Ž. Ivezić *et al*, “Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data”, ISBN: 9781400848911.

Last Words

Handouts and an accompanying iPython notebook (used to make plots) will be available from the DE school wiki page.

Corrections or suggestions for improving this talk are welcome.

Thanks for your attention!