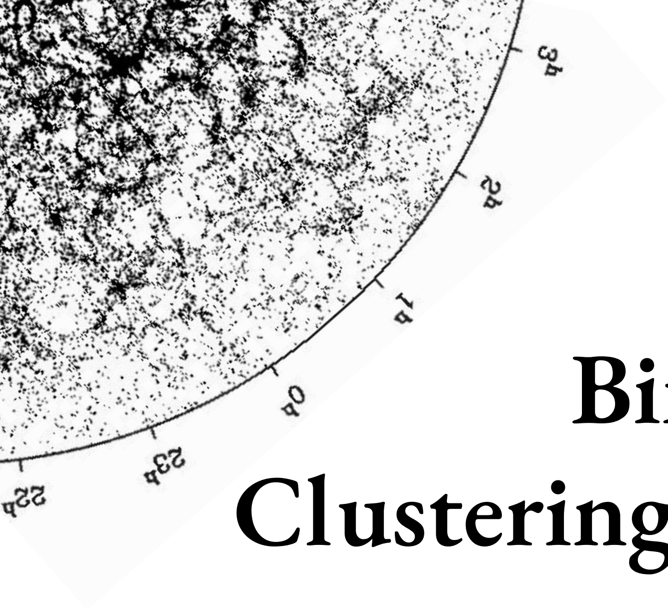
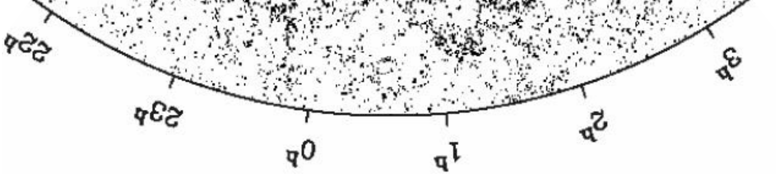
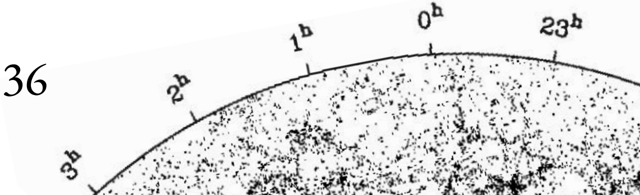




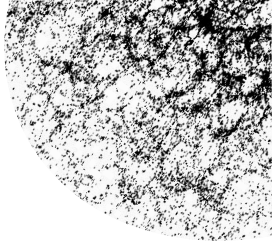
Binning is sinning: Clustering estimation without bins

Kate Storey-Fisher
New York University · NASA FINESST Fellow

German Center for Cosmological Lensing (GCCL) Seminar
January 29, 2021
with David W Hogg · arXiv:2011.01836



Background and research context



statistical and data science methods for
galaxy surveys and large-scale structure analyses

making better cosmological measurements

➔ generalized estimator for the two-point correlation function

➔ emulation of clustering statistics for cosmology & galaxy formation

detecting weirdness in data

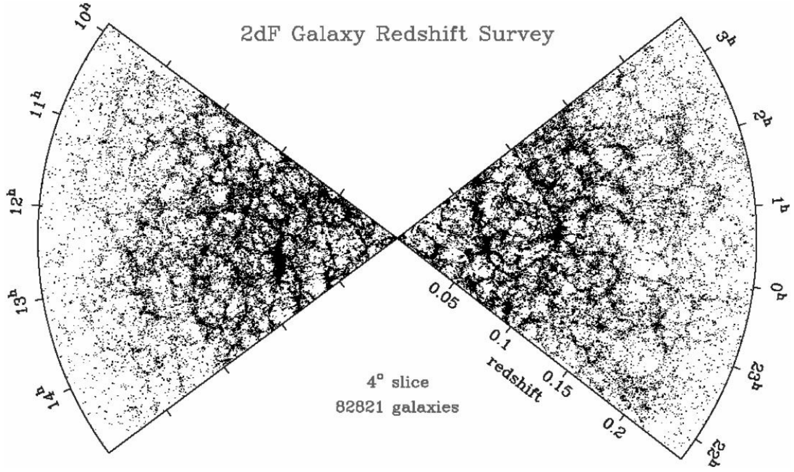
➔ anomaly detection in galaxy surveys with generative models

➔ systematics & anomalies in large-scale structure with probabilistic ML

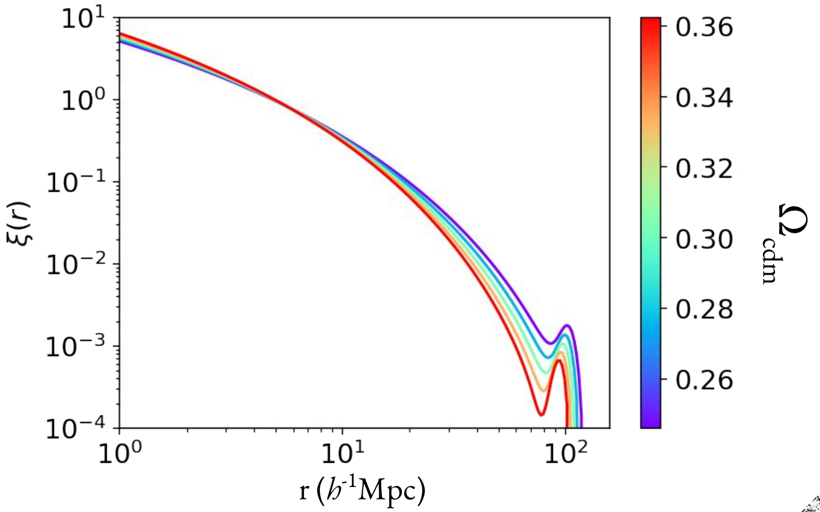
Cosmology from large-scale structure

Galaxies trace the underlying density field → extract cosmological information from clustering

Surveys increasing in volume, complexity, # galaxies:
SDSS: 2M (2019), DESI: 18M (2025) Euclid: 50M (2030)



Infer cosmological parameters from summary statistics:
e.g. 2-point correlation function (2pcf), *ξ(r)*:

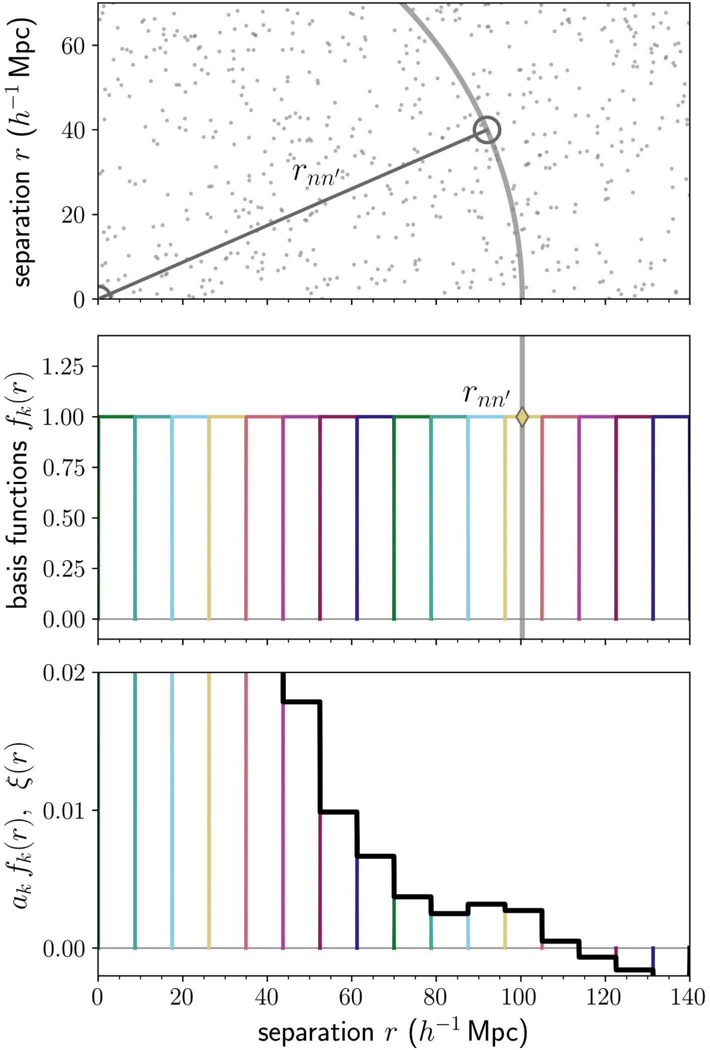


2-point function estimation

The 2pcf in practice:
Excess probability of finding a galaxy at a distance *r* from another galaxy, compared to a uniform distribution

$$dP = \bar{n}[1+\xi(r)]dV$$

binning = projection onto tophat functions



Count up pairs in bin *k*:

$$DD_k \equiv \frac{2}{N_D(N_D-1)} \sum_n \sum_{n'} i(r_{\mathrm{min},k} < |r_{nn'}| < r_{\mathrm{max},k})$$

Naïve estimator, taking into account window function:

$$\hat{\xi}_{\mathrm{PH},k} = \frac{DD_k}{\boxed{RR_k}} - 1$$

pair counts in a uniform random catalog

(Peebles & Hauser 1974, PH)

2-point function estimation

Standard Estimator:
(Landy & Szalay 1993, LS)

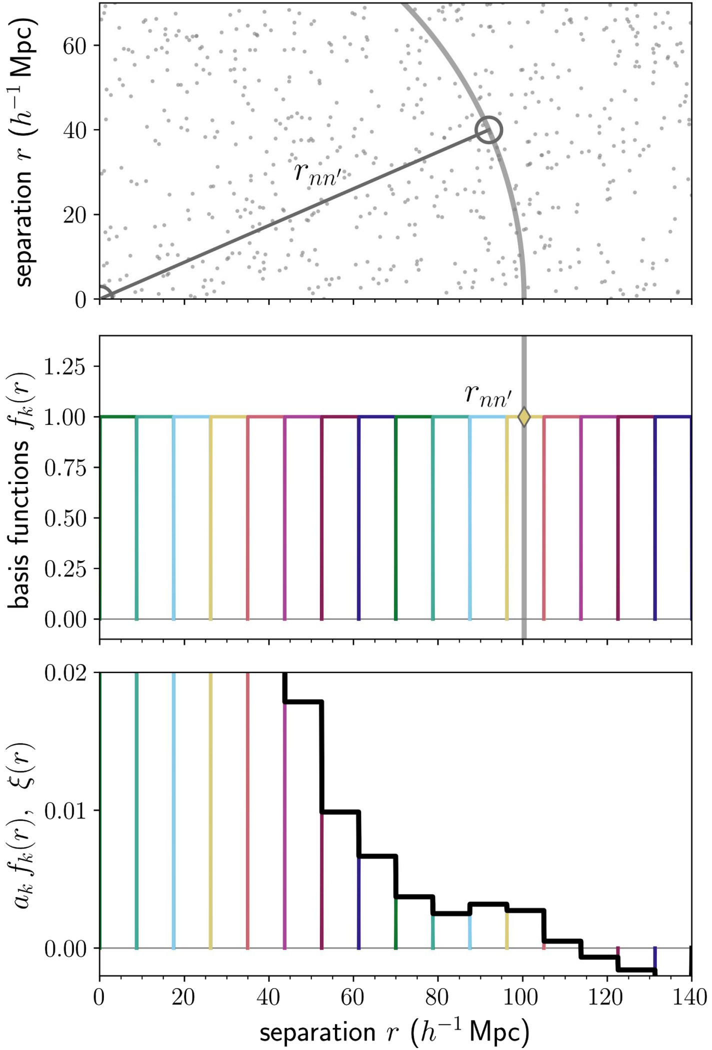
$$\hat{\xi}_{\mathrm{LS},k} = \frac{DD_k - 2DR_k + RR_k}{RR_k}$$

Limitations of Landy-Szalay estimator:

- Suboptimal variance properties + bias at large scales
- Requires choice of bins: tradeoff between bias & variance
- Must bin along another axis to look at dependence on other properties
- Need many mocks to estimate covariance - limiting factor in cosmological analyses

$$\text{error} \sim (1 + N_{\boxed{\text{components}}} / N_{\text{mocks}})$$

a.k.a. bins in standard formulation (e.g. Percival+2015)



The Continuous-Function Estimator

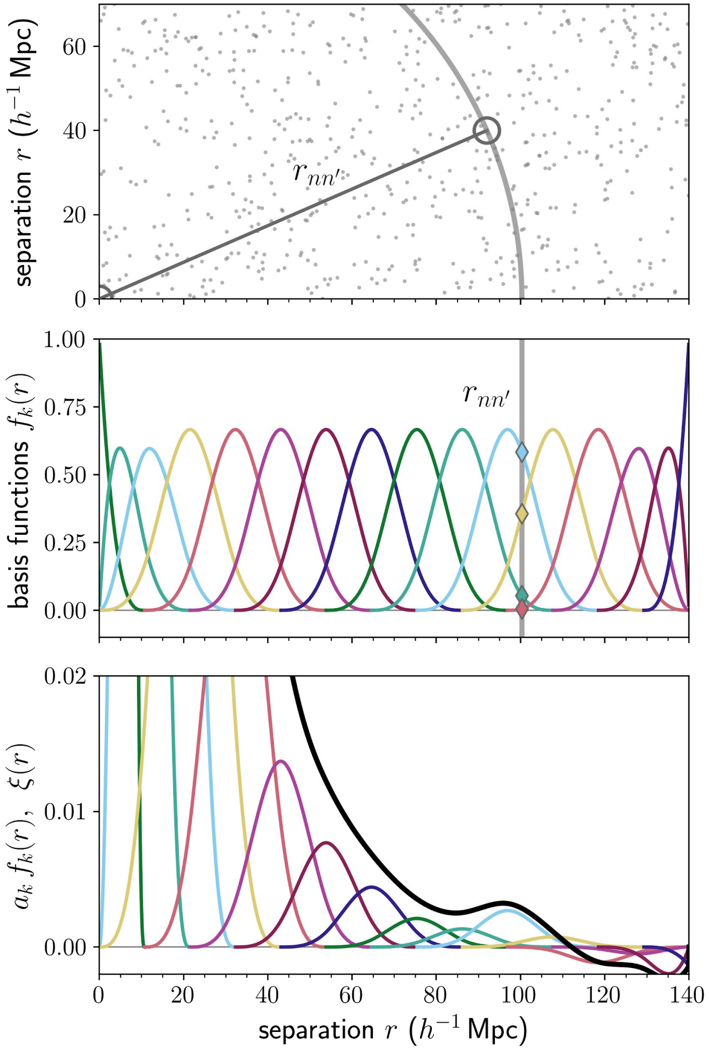
Motivation: connection to linear least-squares fitting

$$\hat{\theta} = [\mathbf{X}^{\mathsf{T}}\,\mathbf{C}^{-1}\,\mathbf{X}]^{-1}\,[\mathbf{X}^{\mathsf{T}}\,\mathbf{C}^{-1}\,\mathbf{y}]$$

rescales features into space of parameters [RR^{-1}]

projects data onto features [DD]

project onto any basis functions!



Reformulate pair counts as **projections of data onto basis functions;** apply a normalization based on the random catalog.

The Continuous-Function Estimator finds the **best-fit linear combination of basis functions** to estimate the 2pcf.

The Continuous-Function Estimator

Project pairs onto continuous basis functions:

$$\mathbf{v}_{\mathrm{DD}} \equiv \frac{2}{N_{\mathrm{D}}\,(N_{\mathrm{D}}-1)} \sum_n \sum_{n'<n} \boxed{\mathbf{f}(\mathsf{G}_{nn'})}$$

any continuous function of pair

$$\mathbf{v}_{\mathrm{DR}} \equiv \frac{1}{N_{\mathrm{D}}\;N_{\mathrm{R}}} \sum_n \sum_m \mathbf{f}(\mathsf{G}_{nm})$$

$$\mathbf{v}_{\mathrm{RR}} \equiv \frac{2}{N_{\mathrm{R}}\,(N_{\mathrm{R}}-1)} \sum_m \sum_{m'<m} \mathbf{f}(\mathsf{G}_{mm'})$$

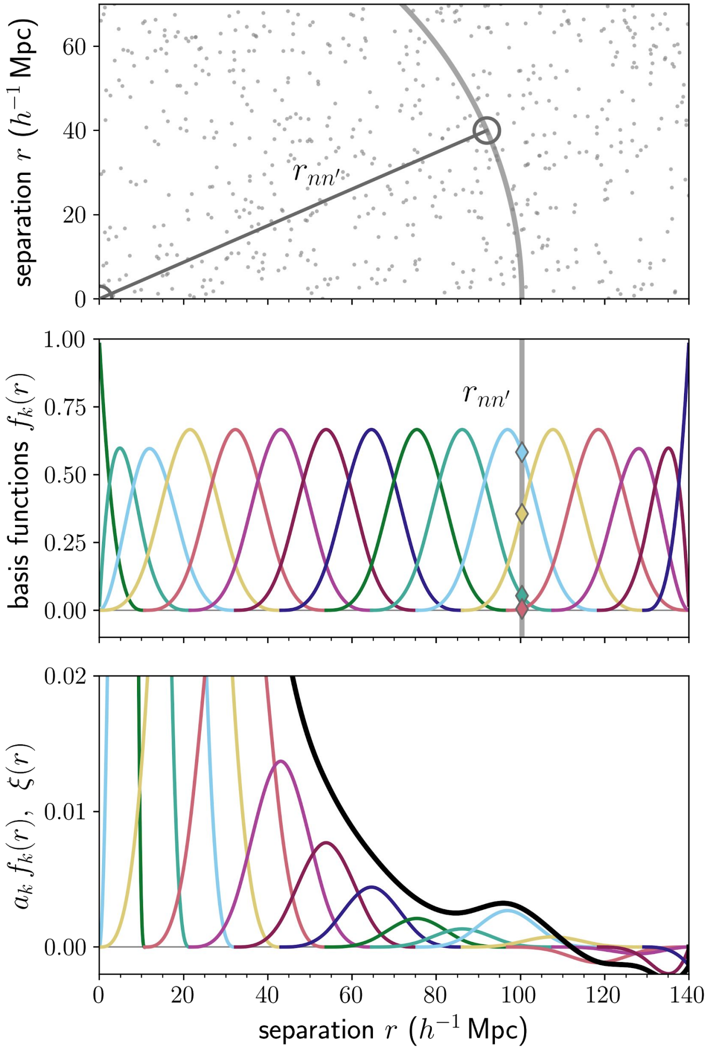
tensor term to rescale features!

$$\mathbf{T}_{\mathrm{RR}} \equiv \frac{2}{N_{\mathrm{R}}\,(N_{\mathrm{R}}-1)} \sum_m \sum_{m'<m} \mathbf{f}(\mathsf{G}_{mm'}) \cdot \mathbf{f}^{\mathsf{T}}(\mathsf{G}_{mm'})$$

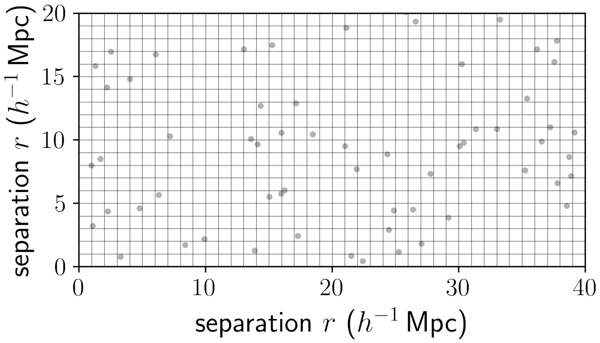
The 2pcf becomes:

$$\mathbf{a} \equiv \mathbf{T}_{RR}^{-1} \cdot (\mathbf{v}_{DD} - 2\,\mathbf{v}_{DR} + \mathbf{v}_{RR})$$

$$\hat{\xi}_{\mathsf{CFE}}(\mathsf{G}_{ij}) \equiv \mathbf{a}^{\mathsf{T}} \cdot \mathbf{f}(\mathsf{G}_{ij})$$



Connection of the Estimator to Least-Squares Fitting



Divide volume into N_{CC} cells containing 1 or 0 galaxies, "**cell occupation number**" **N**:

Construct **data vectors** of whether a cell pair contains a galaxy pair, and a **design matrix** of features, a.k.a basis function component values:

$$\mathbf{y}_{DD} = \begin{bmatrix} \mathcal{N}_{00} \\ \mathcal{N}_{01} \\ \vdots \\ \mathcal{N}_{N_C N_C} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} f_0(\mathsf{G}_{00}) & \cdots & \boxed{f_K(\mathsf{G}_{00})} \\ f_0(\mathsf{G}_{01}) & \cdots & f_K(\mathsf{G}_{01}) \\ \vdots & \ddots & \\ f_0(\mathsf{G}_{N_C N_C}) & & f_K(\mathsf{G}_{N_C N_C}) \end{bmatrix}$$

the basis functions from the last slide, evaluated at cell 00

The amplitudes become:

$$\hat{\mathbf{a}} \approx \boxed{\left[\frac{1}{N_{CC}}\,\mathbf{X}^\mathsf{T}\,\mathbf{X}\right]^{-1}\left[\frac{1}{N_{DD}}\,\mathbf{X}^\mathsf{T}\,\mathbf{y}_{DD}\right.} - 2\,\frac{1}{N_{DR}}\,\mathbf{X}^\mathsf{T}\,\mathbf{y}_{DR} + \frac{1}{N_{RR}}\,\mathbf{X}^\mathsf{T}\,\mathbf{y}_{RR}\Big]$$

The terms look like a least-squares fit!
$\hat{\mathbf{a}} = [\mathbf{X}^\mathsf{T}\,\mathbf{C}^{-1}\,\mathbf{X}]^{-1}\,[\mathbf{X}^\mathsf{T}\,\mathbf{C}^{-1}\,\mathbf{y}]$

Each term is the least-squares estimate for the 2pcf of the occupation number of the catalog pair. For a cell pair *ll*, this is:

$$\langle \mathcal{N}_{D,\ell}\,\mathcal{N}_{D,\ell'} \rangle = \frac{N_{DD}}{N_{CC}}(1 + \xi_{\ell\ell'})$$

Estimate the 2pcf at cell pair *ll*:

$$\boxed{\mathbf{X}_{\ell\ell'}\,\hat{\mathbf{a}} \simeq \xi_{\ell\ell'}}$$

This is the Continuous-Function estimator we wrote on the last slide.

Demo: Tophat basis

Artificial dataset: 1000 lognormal mock catalogs

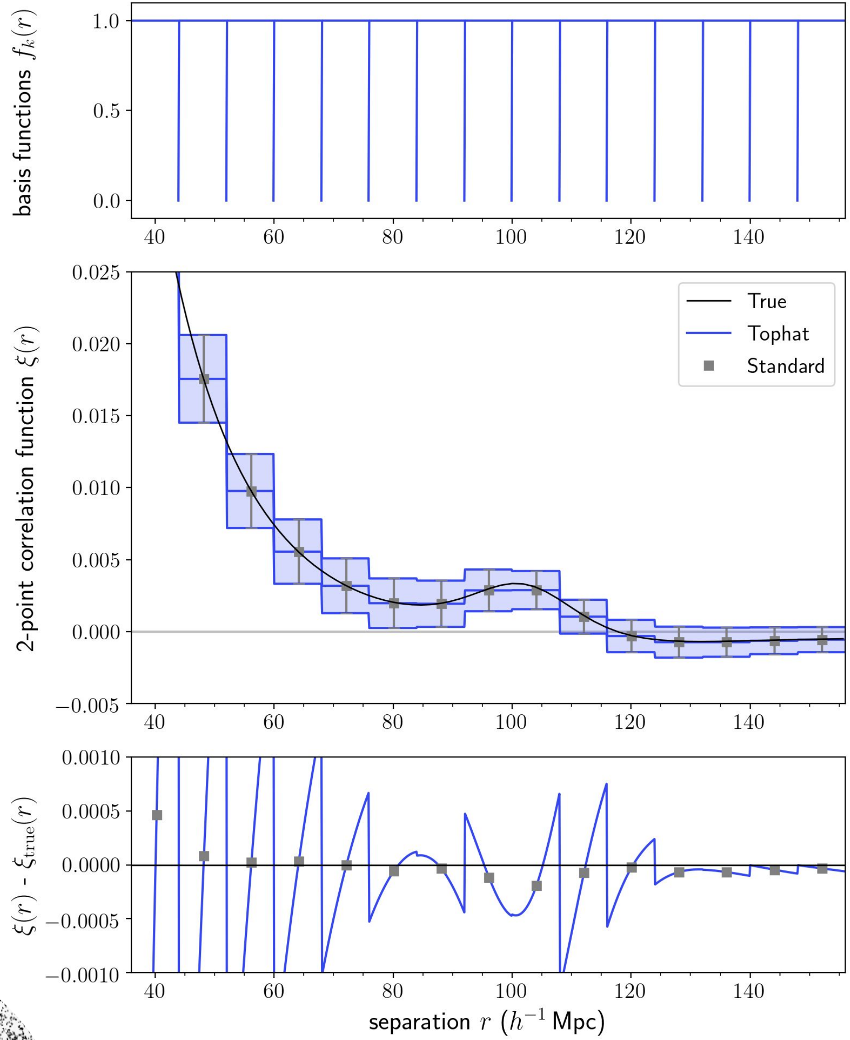
- Planck cosmology
- Periodic boxes with size (750 h^{-1}Mpc)^3
- Galaxy number density 2 x 10^{-4} h^3Mpc^{-3}

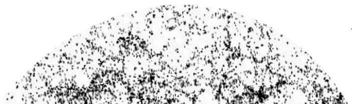
Choose basis functions $\mathbf{f}$ to be tophats:

$$\mathbf{f}(\mathsf{G}_{ij}) = \mathbf{f}_{\text{tophat}}(r_{ij})$$

The Continuous-Function Estimator then depends only on pair separation, and reduces to:

$$\hat{\xi}_{\text{CFE}}(\mathsf{G}_{ij}) = \hat{\xi}_{\text{LS}}(r)$$





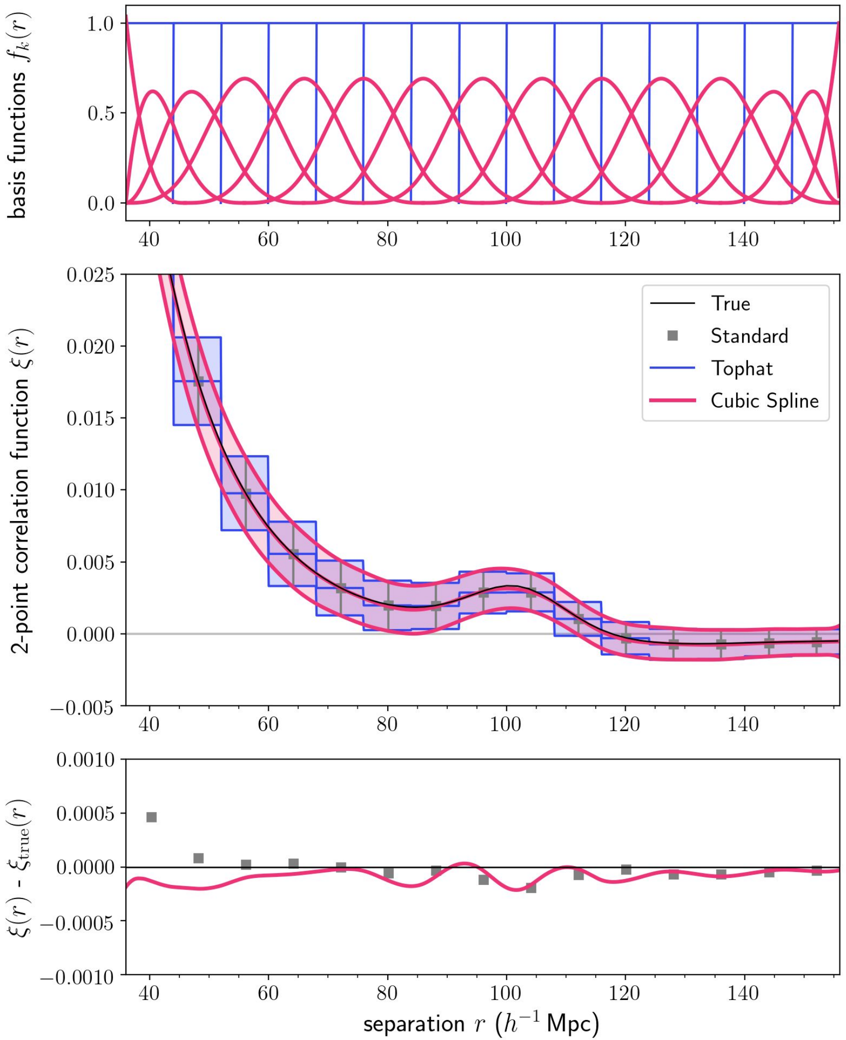
Demo: Cubic spline basis

Choose cubic splines as basis functions

- continuous functions, & continuous first derivative
- relatively well-localized

Advantages

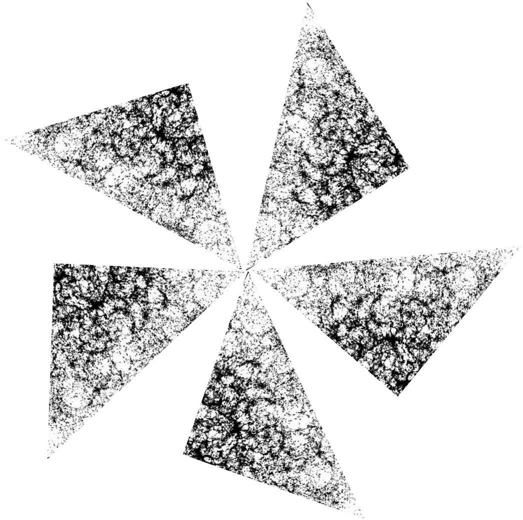
- Smooth estimate is more representative of the true 2pcf
- Continuous derivatives useful for certain applications
- Preserves information and has well-defined bias properties, unlike kernel density methods



Advantages and limitations of the estimator

Advantages

- No need for bins; more representative of true 2pcf (mixture of tophats is a poor representation; expect continuity)
- More expressive: space of basis functions >> space of bins
- Can include dependence on pair properties other than separation
- *Same accuracy with fewer components: impt. for covariance estimation*
- *Can tailor basis functions to science goals*



Limitations

- Must represent 2pcf form as linear combination
- Can increase computational cost; evaluating **f** for every pair
- Inherits many limitations of Landy-Szalay, incl. non-optimal bias & variance properties and window function estimation

Application: Direct BAO scale estimation

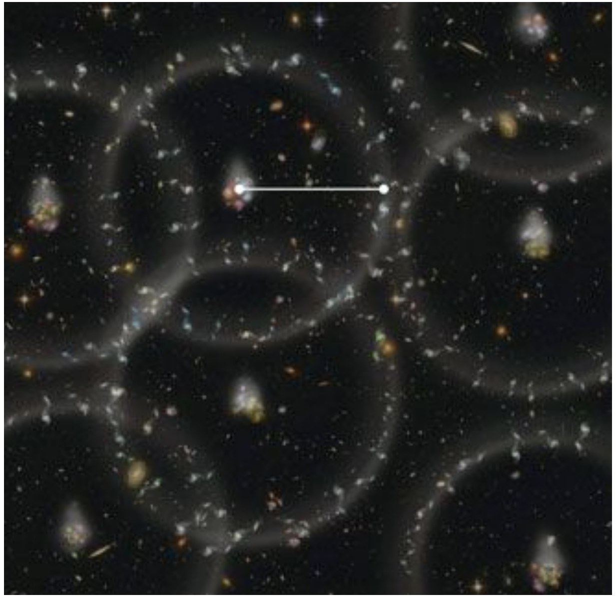
Baryon acoustic oscillations (BAO) in early U imprint higher density regions at a certain scale; measure of the distance-redshift relation.

Standard approach: Use fiducial model (mod) with some scale dilation parameter α that combines distance information:

$$\xi^{\text{fit}}(r) = \underbrace{B^2 \xi^{\text{mod}}(\alpha r)}_{\text{fiducial model}} + \underbrace{\frac{a_1}{r^2} + \frac{a_2}{r} + a_3}_{\text{nuisance parameters}}$$

$$\alpha = \underbrace{\left(\frac{D_A(z)}{D_A^{\text{mod}}(z)}\right)^{2/3}}_{\text{angular diameter distance}} \underbrace{\left(\frac{H^{\text{mod}}(z)}{H(z)}\right)^{1/3}}_{\text{Hubble parameter}} \underbrace{\left(\frac{r_s^{\text{mod}}}{r_s}\right)}_{\text{radius of sound horizon}}$$

Use binned estimator, then fit for best α.



Application: Direct BAO scale estimation

Continuous-Function Estimator approach: Linearize around α:

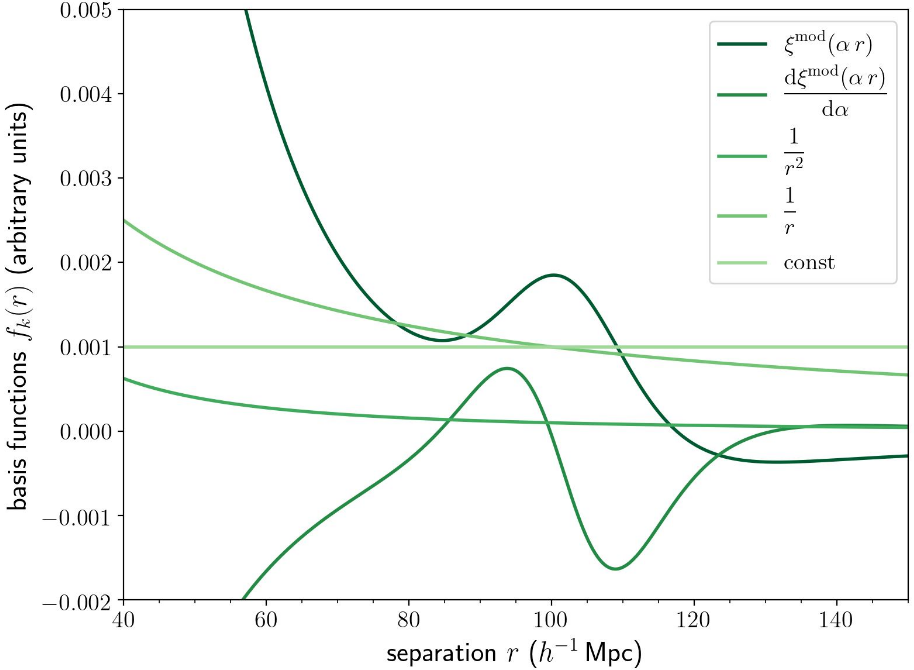
$$\xi^{\text{fit}}(r) = B^2 \xi^{\text{mod}}(\alpha_{\text{guess}}\, r) + \boxed{C\, k_0 \frac{\mathrm{d}\xi^{\text{mod}}(\alpha_{\text{guess}}\, r)}{\mathrm{d}\alpha}}$$

$$+ a_1 \frac{k_1}{r^2} + a_2 \frac{k_2}{r} + a_3\, k_3$$

derivative wrt α

Directly project data onto this 5-component model (no bins!). Gives us C, which gives an estimate of α:

$$\hat{\alpha} = \alpha_{\text{guess}} + C\, k_0$$

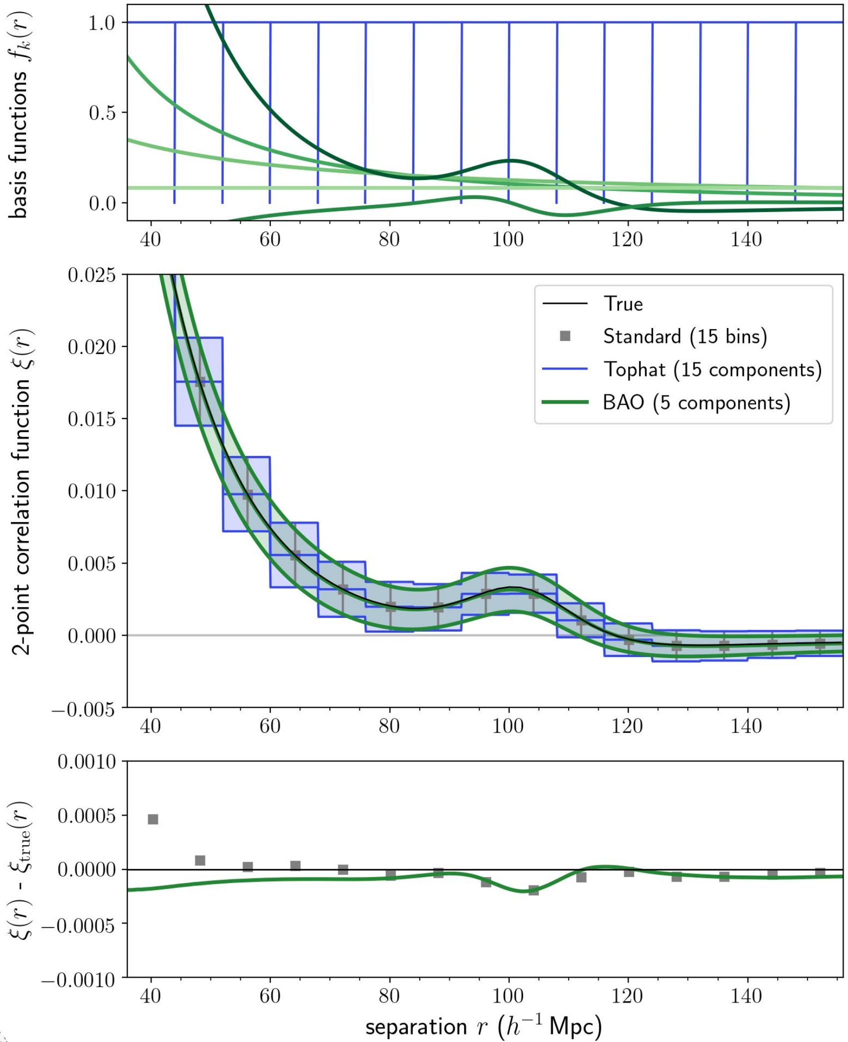


A continuous BAO 2pcf

- Produces a smooth correlation function
- Directly estimates parameter of interest; recovered scale dilation parameter:

  $$\hat{\alpha} = 0.9975 \pm 0.0290$$
  $$(\alpha_{\text{true}} = 0.9987)$$

- Uses only 5 components, compared to 15+ for tophat / standard estimator (fewer mocks required for covariance estimation!)



Other applications of the Continuous-Function Estimator

Reformulations of standard analyses

- Anisotropic BAO analysis $\hat{\xi}(s, \mu)$
- 2D correlation function $\hat{\xi}(r_p, \pi)$, for computing $w_p(r_p)$
- Power spectrum-adjacent estimator by projecting onto Fourier modes

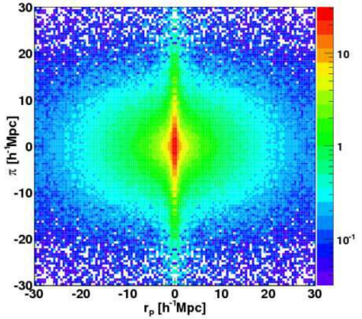
Direct estimation of cosmological quantities

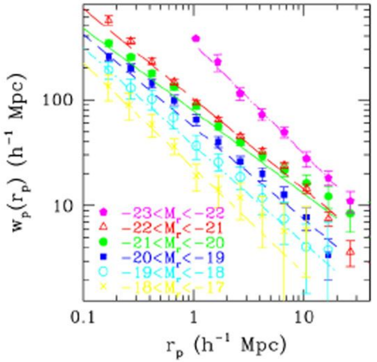
- Growth rate of structure *f(a)* or local primordial non-Gaussianity f_{NL}
- Cosmological and HOD parameters using derivatives of model wrt parameters

Dependence of 2pcf on tracer properties

- Redshift dependence / Alcock-Paczynski effect
- Luminosity/color/mass dependence of galaxy clustering

Signals indicating new physics, e.g. anisotropies or inhomogeneities





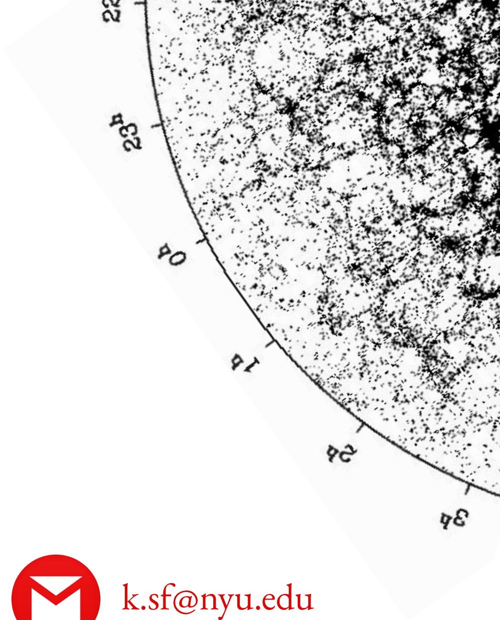
Summary & future research

The Continuous-Function Estimator: a new estimator for the 2pcf that projects pairs onto continuous basis functions. No bins required.

- Produces smooth correlation functions that better represent the true 2pcf
- Can incorporate dependence on pair properties other than separation
- Directly estimates parameters of interest using specialized basis functions
- Achieves same accuracy with fewer components: impt. for covariance estimation

Upcoming work and directions

- Apply the CFE to improve measurements & investigate new physics
- Revisit window function estimation, optimal bias & variance estimators
- Combine with new approaches to systematics mitigation



k.sf@nyu.edu

@katestoreyfish

@kstoreyf