Hermes - Toward an Ultimate Algorithm for Cosmic Statistics in Extremely Large Data Set

Feng Longlong

School of Physics and Astronomy
Sun Yat-Sen University

Yanchuan Cai, Feng, Fang, Wenjie Ju, Zhiqi Huang,

Zuoyang Li, Shiyu Yue, Jun Pan, Weishan Zhu

2024.06.04 – Higgs Center @ University of Edinburgh

Outline

- Introduction
- Multi-Resolution Analysis - MRACS Algorithm
- Unified Framework for Correlation Functions
- Quantifying the Binning Effect in Correlation Analysis
- Fast Algorithm for 2PCF & 3PCF -> NPCF
- Summary

Introduction

Motivation: demanding a fast algorithm of cosmic statistics to tackle extremely large data sets from ongoing/upcoming galaxy surveys and numerical simulations.

Clustering Statistics in Cosmology

↓

Counting the Number of Objects 数 数

HOW Quickly ?

What’s Hermes ?

Hermes: HypER–speed MultirEsolution cosmic Statistics

- **An open-source, massively parallel & GPU accelerated Python toolkit for cosmic statistics**
- $N_g \log N_g$ **Algorithm, independent of number of sampling points (** N_g **: Grid Number)**
- **Making a unified scheme for all variants of clustering statistical measures**

THE ASTROPHYSICAL JOURNAL, 658:25–35, 2007 March 20

THE BEYLKIN-CRAMER SUMMATION RULE AND A NEW FAST ALGORITHM OF COSMIC STATISTICS FOR LARGE DATA SETS

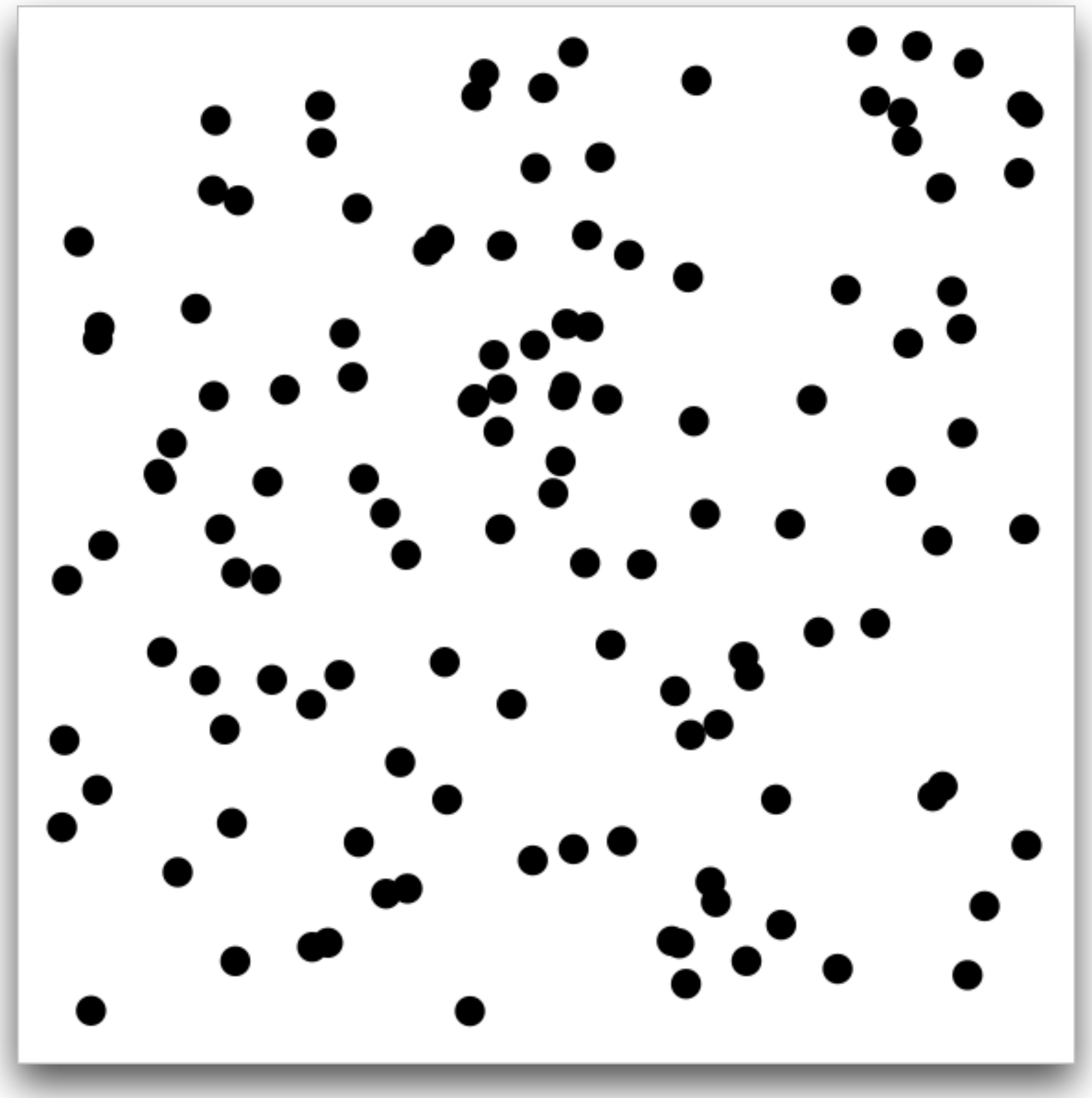
LONG-LONG FENG
Purple Mountain Observatory, Nanjing, China; and Joint Center for Particle Nuclear Physics and Cosmology, Nanjing, China; fengll@pmo.ac.cn
Received 2006 March 7; accepted 2006 November 15

Modeling Spatial Point Processes

$$n(\mathbf{x}) = \sum_{i=1}^{N} w_i \delta_D^3(\mathbf{x} - \mathbf{x_i})$$

Weight $\{w_i, i = 1....N\}$

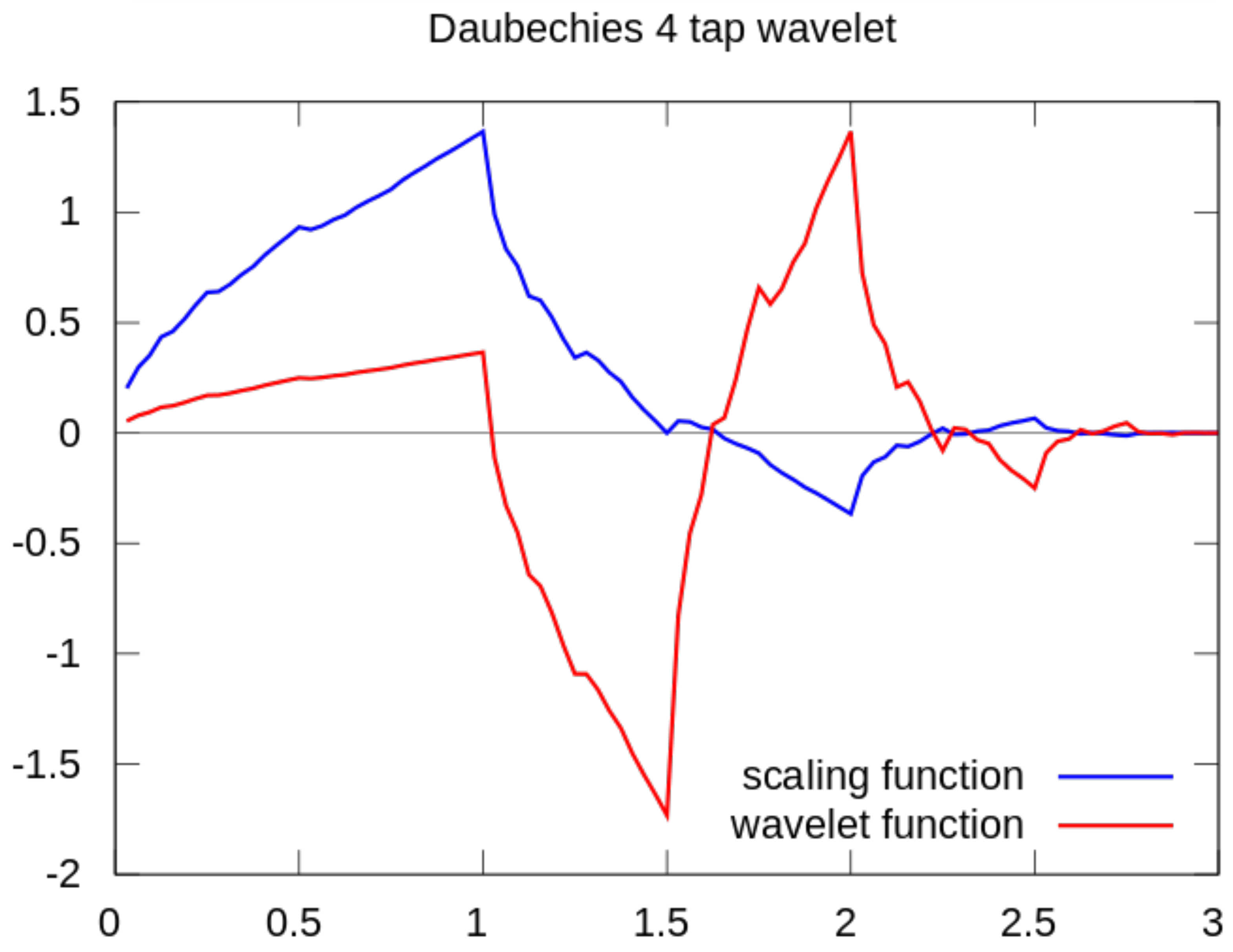
- **window (selection) function in galaxy surveys**
- **some markers related to the intrinsic properties of galaxies**
- **environment variables (local density, cosmic-web type …)**
- **velocity components, etc.**



$$\phi_{j\mathbf{l}}(\mathbf{x}) = \prod_{i=1}^{D} \phi_{jl_i}(x_i)$$

$$\left\{\phi_{j,k}(x) = 2^{j/2}\phi\left(2^j x - k\right) \quad \middle| \, k \in \mathbf{Z}\right\}$$

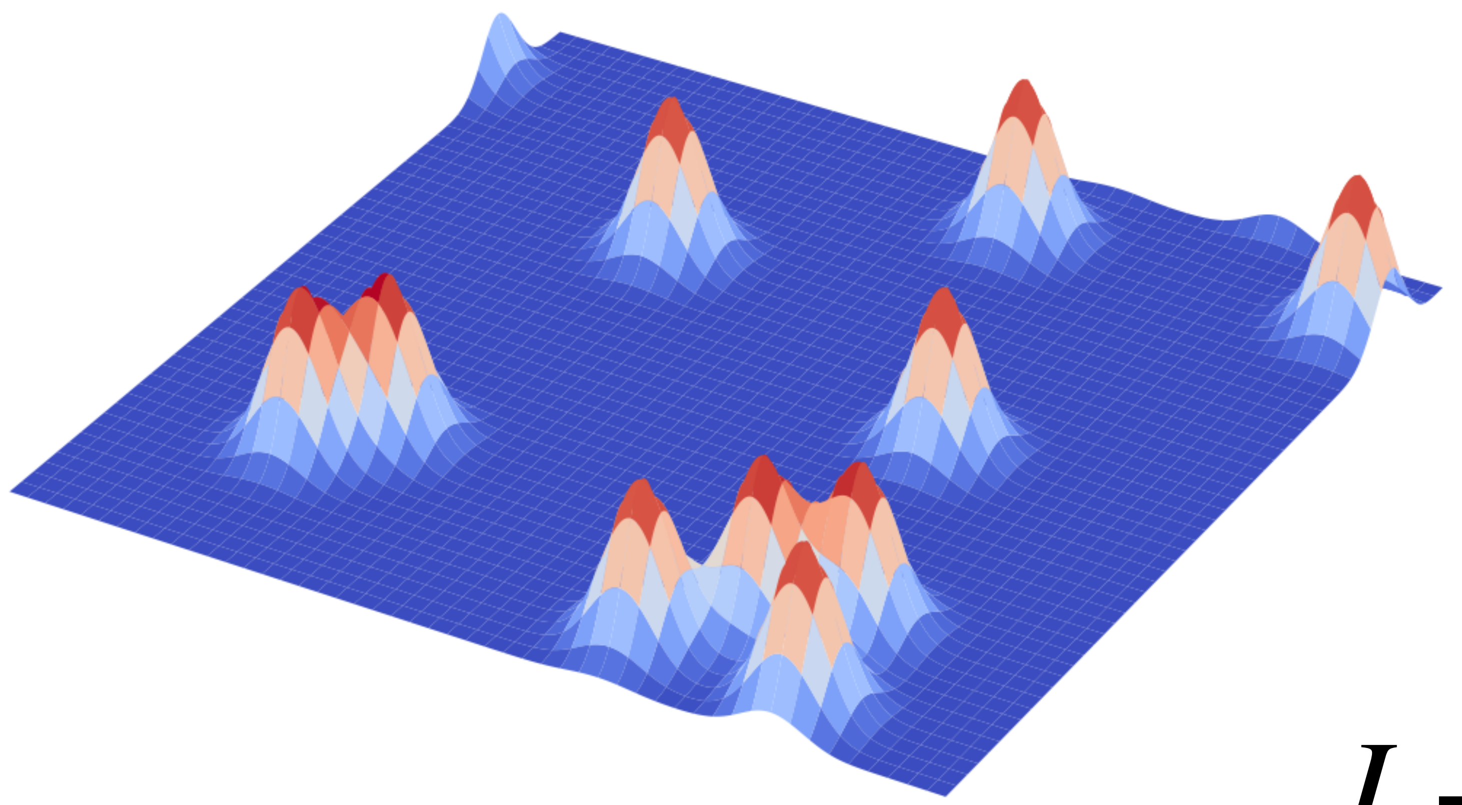
Completeness: $$\sum_{\mathbf{l}} \phi_{j\mathbf{l}}(\mathbf{x})\phi_{j\mathbf{l}}(\mathbf{x}') = \Delta_j(\mathbf{x}, \mathbf{x}') \to \delta_D(\mathbf{x} - \mathbf{x}') \quad j \to \infty$$



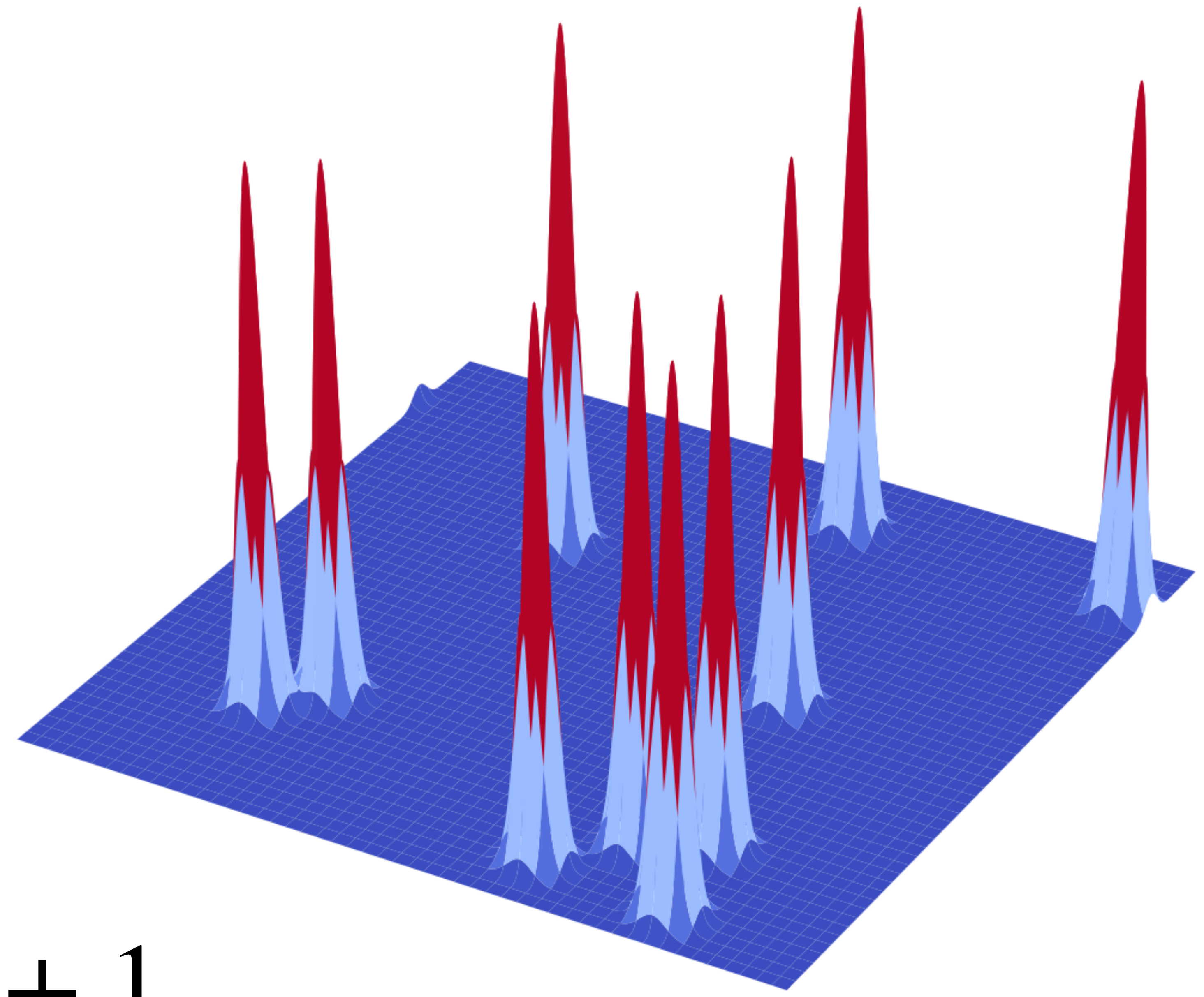
Reconstruction - From Point Processes To Continuous Fields

$$n(\mathbf{x}) \rightarrow n_j(\mathbf{x}) = \sum_{\mathbf{1}} \epsilon_{j\mathbf{1}} \phi_{j\mathbf{1}}(\mathbf{x})$$

$$\epsilon_{j\mathbf{1}} = \int n(\mathbf{x}) \phi_{j\mathbf{1}}(\mathbf{x}) d\mathbf{x} = \sum_{i}^{N} w_i \phi_{j\mathbf{1}}(\mathbf{x}_i)$$



$$J \rightarrow J + 1$$



The Galaxy-Reckoner

Arithmetic：Count-in-Cell → **Algebra：Window Function**

$$n_W(\mathbf{x}) = \sum_{i=1}^{N} w_i W(\mathbf{x} - \mathbf{x}_i) = \int W(\mathbf{x} - \mathbf{x}')n(\mathbf{x}')d^3\mathbf{x}'$$

Normalized Condition:

Low-Pass Filters

$$\int W(\mathbf{x})d\mathbf{x} = 1$$

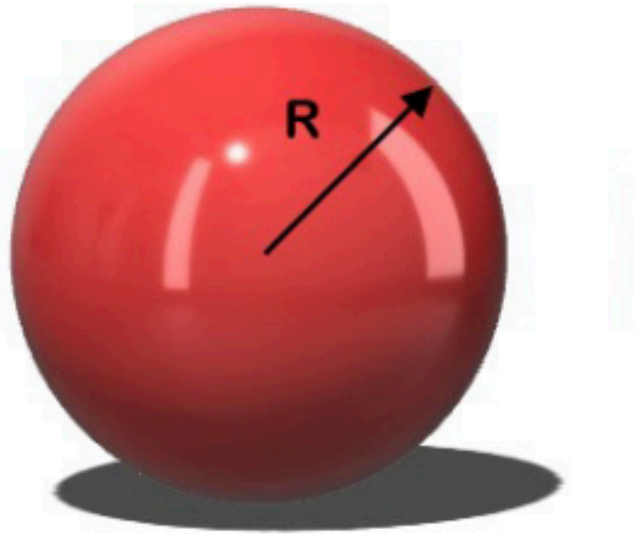
High-Pass Filters

$$\int W^2(\mathbf{x})d\mathbf{x} = 1$$

$$W_{\text{sphere}}(r;R) = \frac{3}{4\pi R^3}\theta(R-r)$$

,

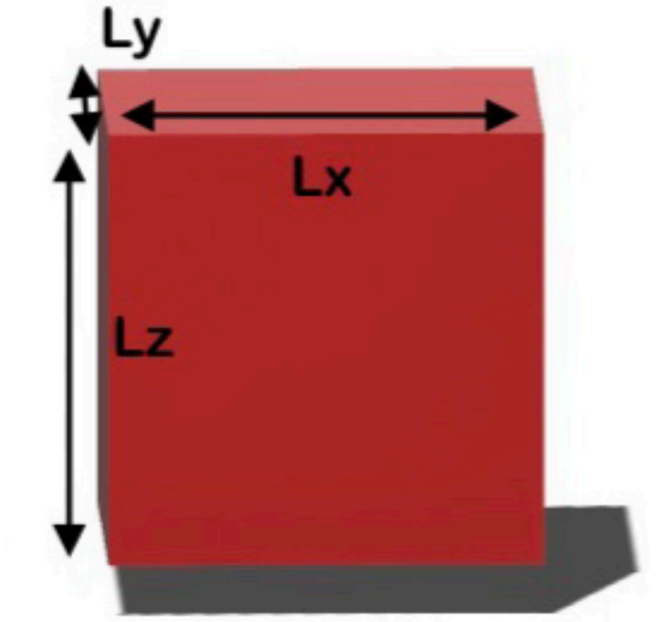
$$\hat{W}_{\text{sphere}}(k;R) = \frac{3(\sin(kR) - kR\cos(kR))}{k^3R^3}$$



$$W_{\text{cubic}}(\mathbf{x};\mathbf{L}) = \Pi_{i\in\{x,y,z\}}\frac{1}{L_i}\theta(L_i/2 - |x_i|)$$

,

$$\hat{W}_{\text{cubic}}(\mathbf{k};\mathbf{L}) = \Pi_{i\in\{x,y,z\}}\frac{\sin(k_iL_i/2)}{k_iL_i/2}$$



The Galaxy Reckoner - A Fast Algorithm

Count-in-Cell $n_W(\mathbf{x}) = \sum_{i=1}^{N} w_i W(\mathbf{x} - \mathbf{x}_i) = \int W(\mathbf{x} - \mathbf{x}') n(\mathbf{x}') d^3\mathbf{x}'$

$$n(\mathbf{x}) \to n_j(\mathbf{x}) = \sum_{\mathbf{l}} \epsilon_{j\mathbf{l}} \phi_{j\mathbf{l}}(\mathbf{x})$$

$$W(\mathbf{x}, \mathbf{y}) \to W_j(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{l},\mathbf{m}} w^j_{\mathbf{l},\mathbf{m}} \phi_{j,\mathbf{l}}(\mathbf{x}) \phi_{j,\mathbf{m}}(\mathbf{y})$$

$$n_W(\mathbf{x}) \to n^j_W(\mathbf{x}) = \sum_{\mathbf{l}} \tilde{\epsilon}_{j\mathbf{l}} \phi_{j,\mathbf{l}}(\mathbf{x})$$

$$\tilde{\epsilon}_{j\mathbf{l}} = \sum_{\mathbf{m}} w^j_{\mathbf{l},\mathbf{m}} \epsilon_{j\mathbf{m}}$$

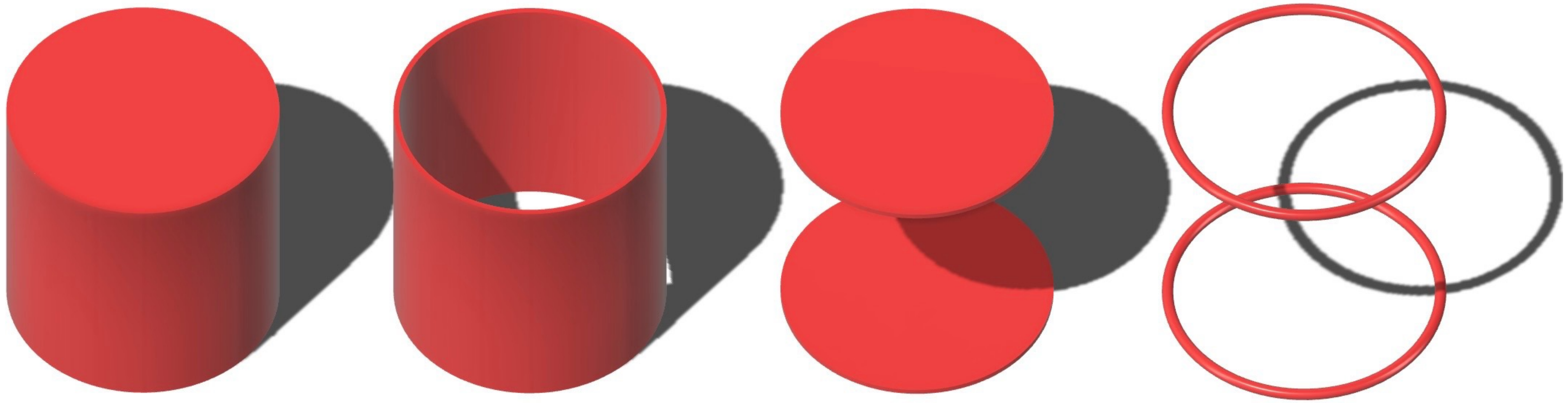
W: Homogeneous Kernel $w^j_{\mathbf{lm}} = w^j_{\mathbf{l-m}}$ → **Toeplitz Matrix**

→ **FFT Technique**

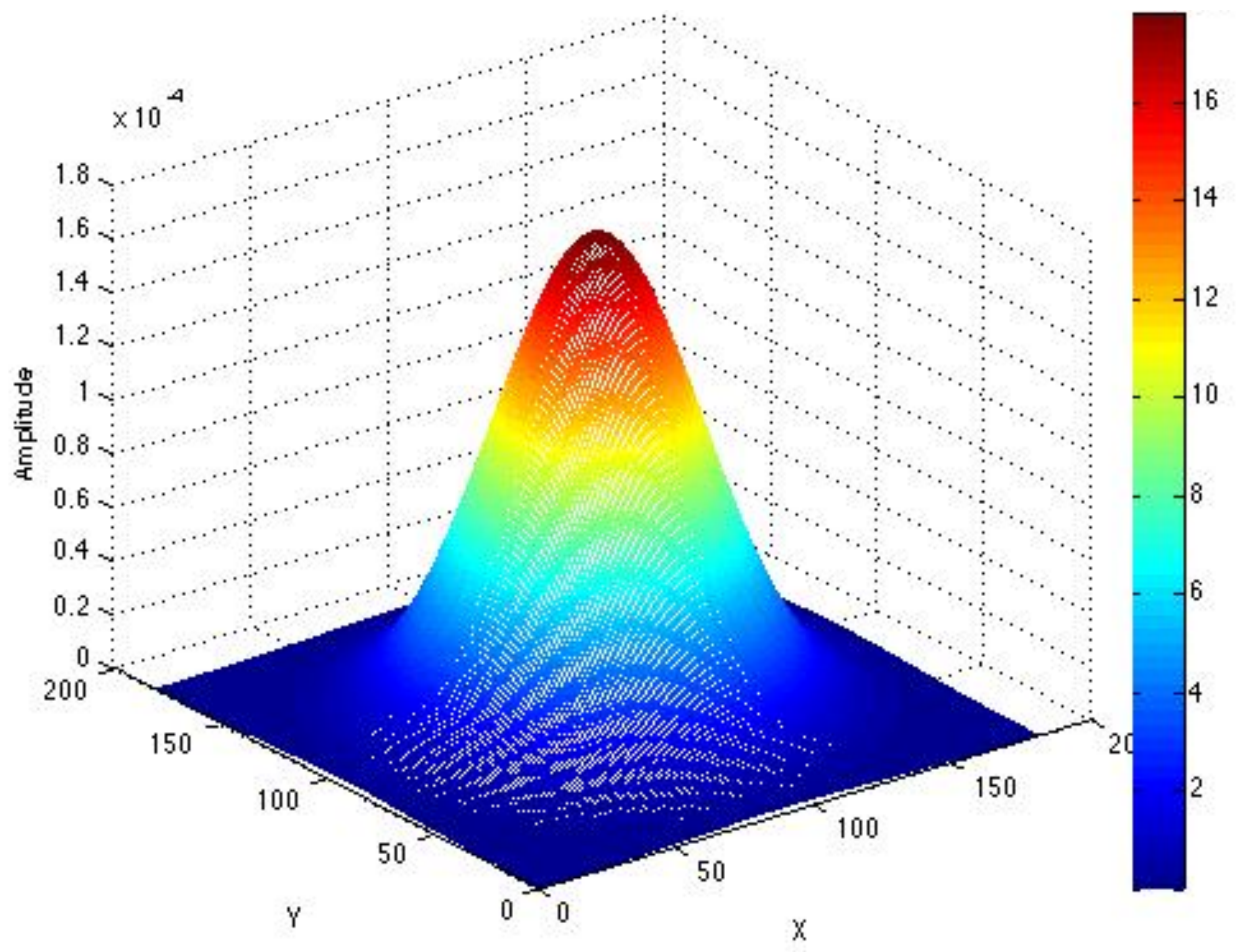
$$\hat{W}_{\text{tophat}}(\mathbf{k}) = \frac{1}{V} \int_V e^{i\mathbf{k}\cdot\mathbf{x}} d^3\mathbf{x}$$

Independent of Number of Particles & Geometry of Count-in-Cell

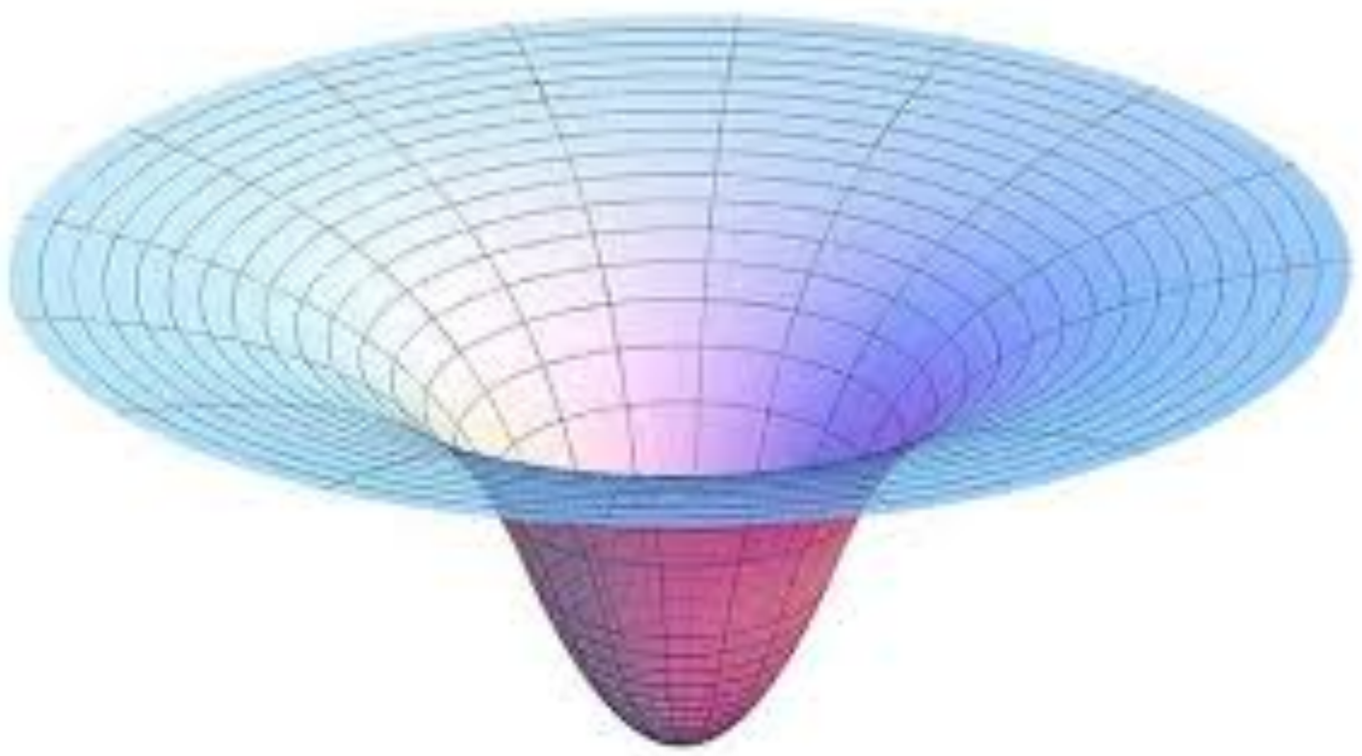
Axial Symmetric Filters



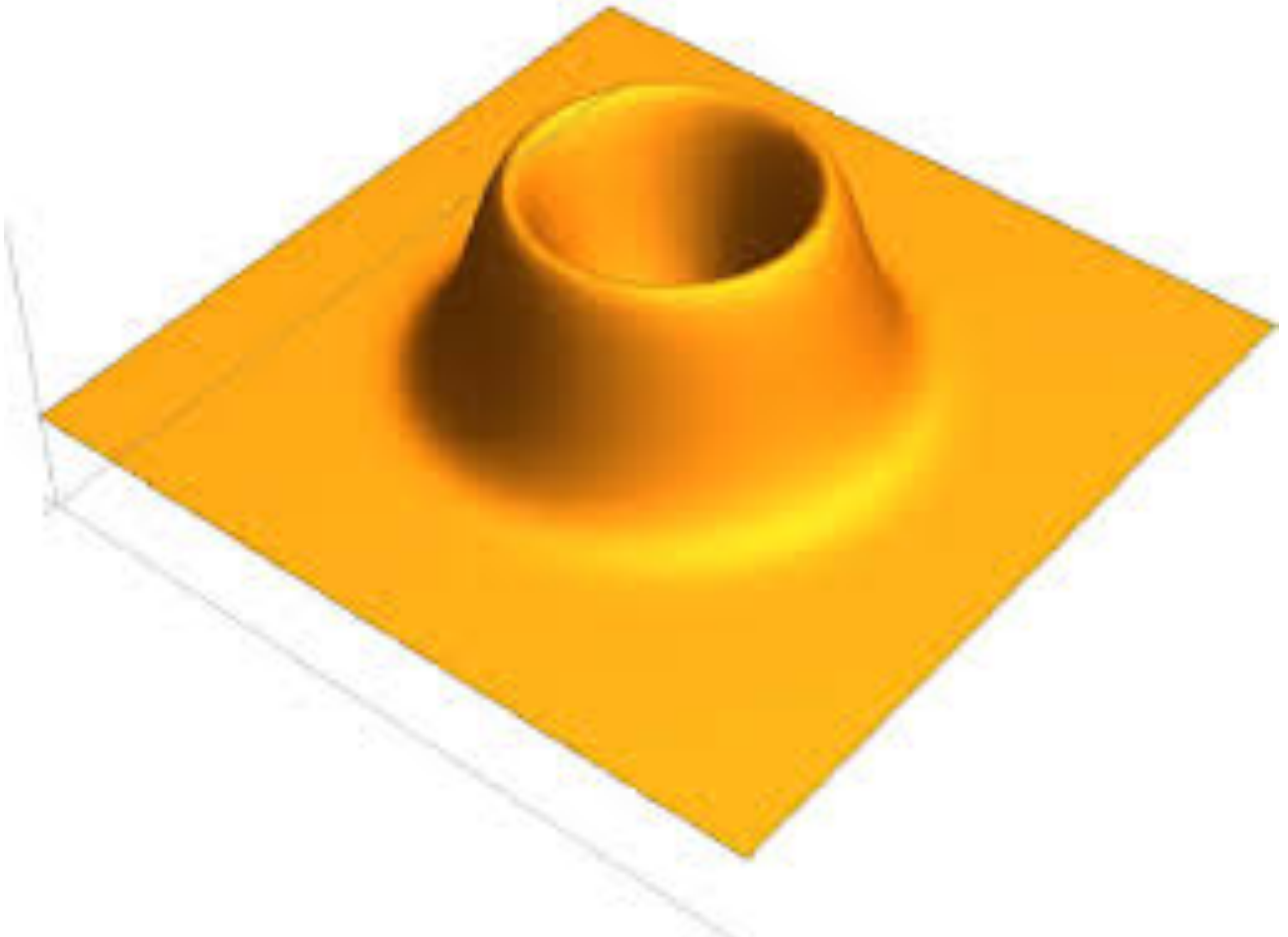
Gaussian Filter



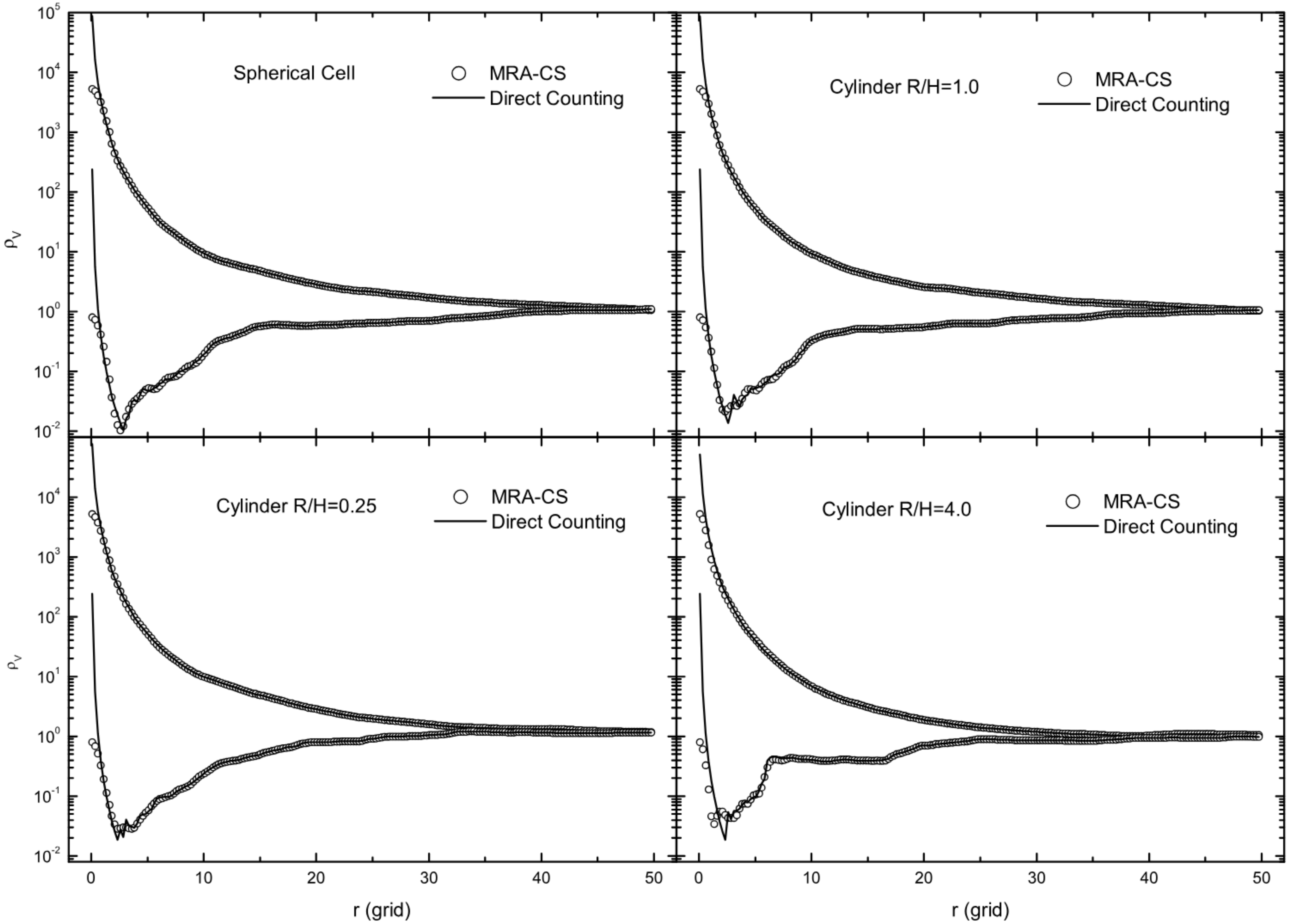
Gravitational Potential

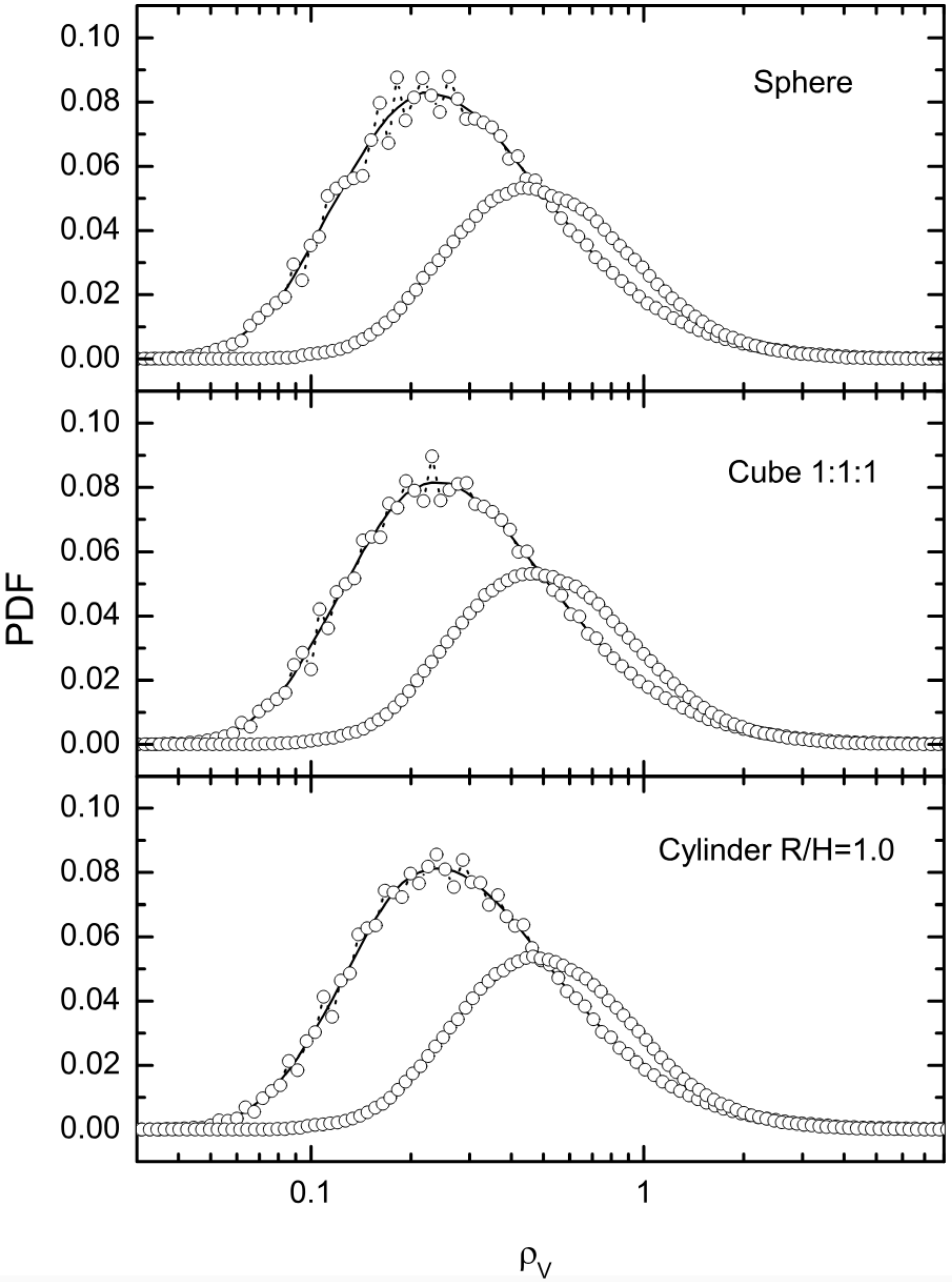


Gaussian High-Pass

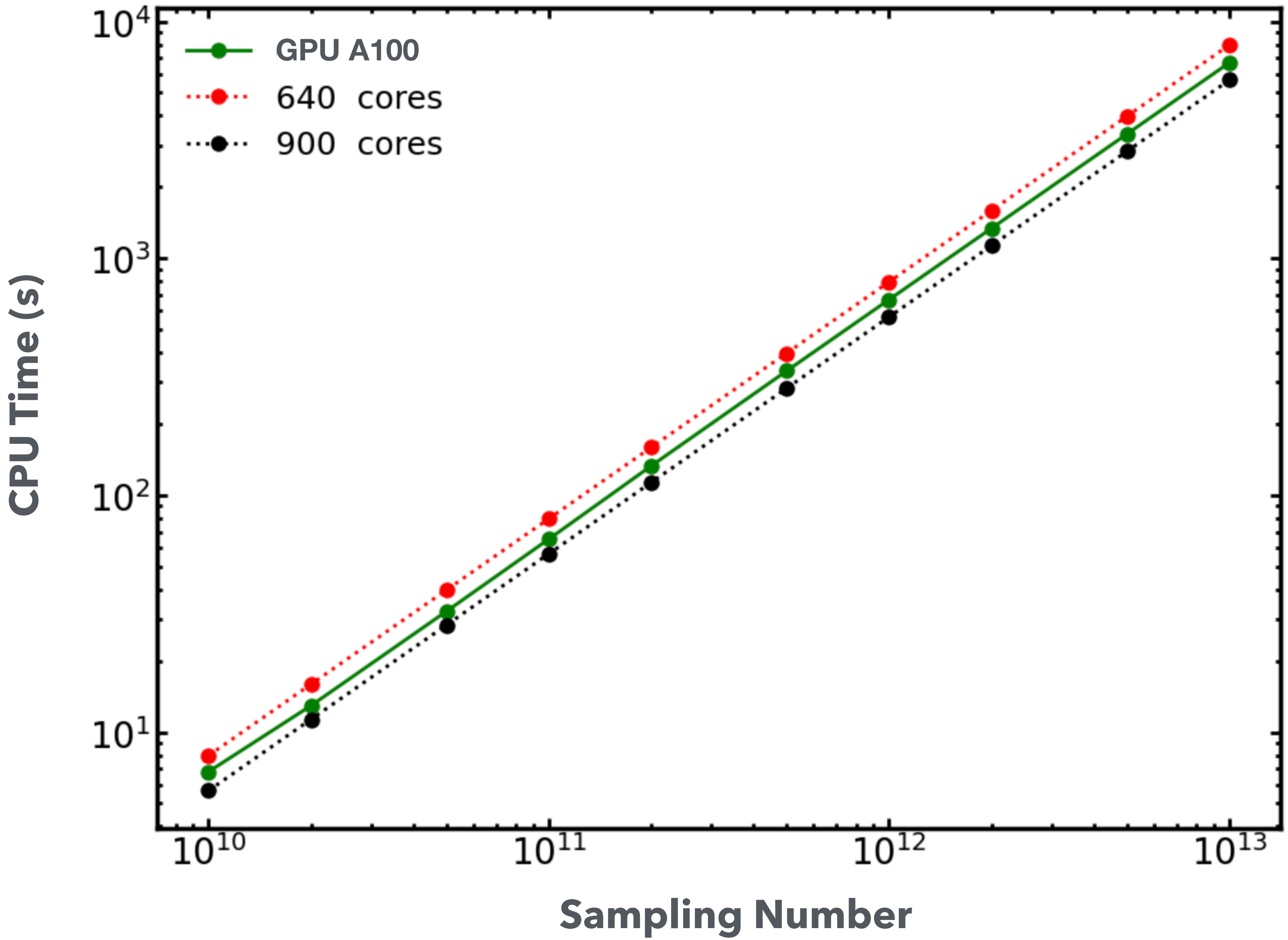


Count-in-Cell: Numerical Tests





Hermes - MRACS Algorithm



Two-Point Correlation Function
- An Alternative View From Ex-situ to In-situ

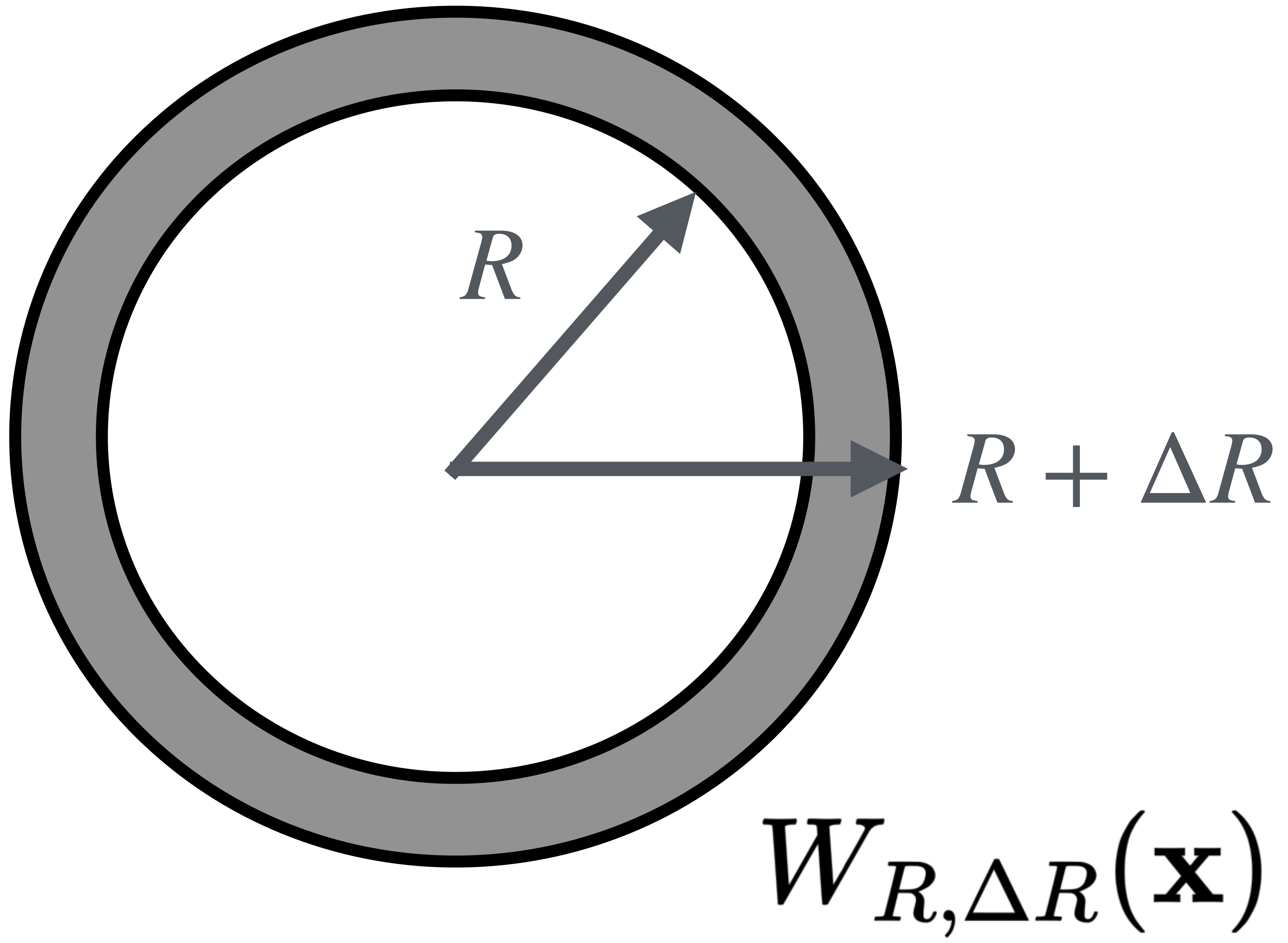
Yue et.al 2024 in prep.

Ex-Situ View: $$\xi(R) = \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{R})\rangle_{\Omega_{\mathbf{R}},\mathbf{x}}$$

Translational Field

$$\delta_R(\mathbf{x}) = \langle \delta(\mathbf{x}+\mathbf{R})\rangle_{\Omega_{\mathbf{R}}} = W(\mathbf{x}, R) \circ \delta(\mathbf{x})$$

In-Situ View: $$\xi(R) = \langle \delta(\mathbf{x})\delta_R(\mathbf{x})\rangle$$



Binning via Finite Spherical Shell

$$W(\mathbf{x}, R) = W_{R,\Delta R}(\mathbf{x}) = \frac{1}{V_{R,\Delta R}}(\theta(r - R - \Delta R) - \theta(r - R))$$

$$= \frac{1}{V_{R,\Delta R}}\left[V_{R+\Delta R}W_{\text{sphere}}(r, R+\Delta R) - V_R W_{\text{sphere}}(r, R)\right]$$

Preliminary

Quantifying the Binning Effect

$$\xi_{\Delta R}(R) = \langle \delta, \delta \circ W_{R,\Delta R} \rangle = \int_0^\infty P(k) W_{R,\Delta R}(k) \frac{k^2 dk}{2\pi^2}$$

$\Delta R \to 0$ →

$$\tilde{W}_{\text{shell}}(k, R) = \frac{\sin(kR)}{kR}$$

$$W_{\text{shell}}(r, R) = \frac{1}{4\pi r^2} \delta_D(r - R)$$

The Relation between 2PCF with and without Binning

$$\xi_{\Delta R}(R) = \frac{1}{V_{R,\Delta R}} \int_{V_{R,\Delta R}} \xi(R) dV_R$$

Fast Pair-Counting Algorithm from In-situ View

$$DD = \langle n(\mathbf{x}), n_W(\mathbf{x}) \rangle = \frac{1}{V} \int n(\mathbf{x}) n_W(\mathbf{x}) d^3\mathbf{x}$$

$$n(\mathbf{x}) = \sum_{\mathbf{l}} \epsilon_{j\mathbf{l}} \phi_{j,\mathbf{l}}(\mathbf{x}) \qquad n_W(\mathbf{x}) = \sum_{\mathbf{l}} \tilde{\epsilon}_{j\mathbf{l}} \phi_{j,\mathbf{l}}(\mathbf{x})$$

The orthogonality of basis functions

$$\tilde{\epsilon} = W_R \cdot \epsilon$$

Real Space

$$DD = \sum_{\mathbf{l}} \epsilon_{j\mathbf{l}} \tilde{\epsilon}_{j\mathbf{l}} = \epsilon \cdot \tilde{\epsilon} \qquad O(N_g \log N_g)$$

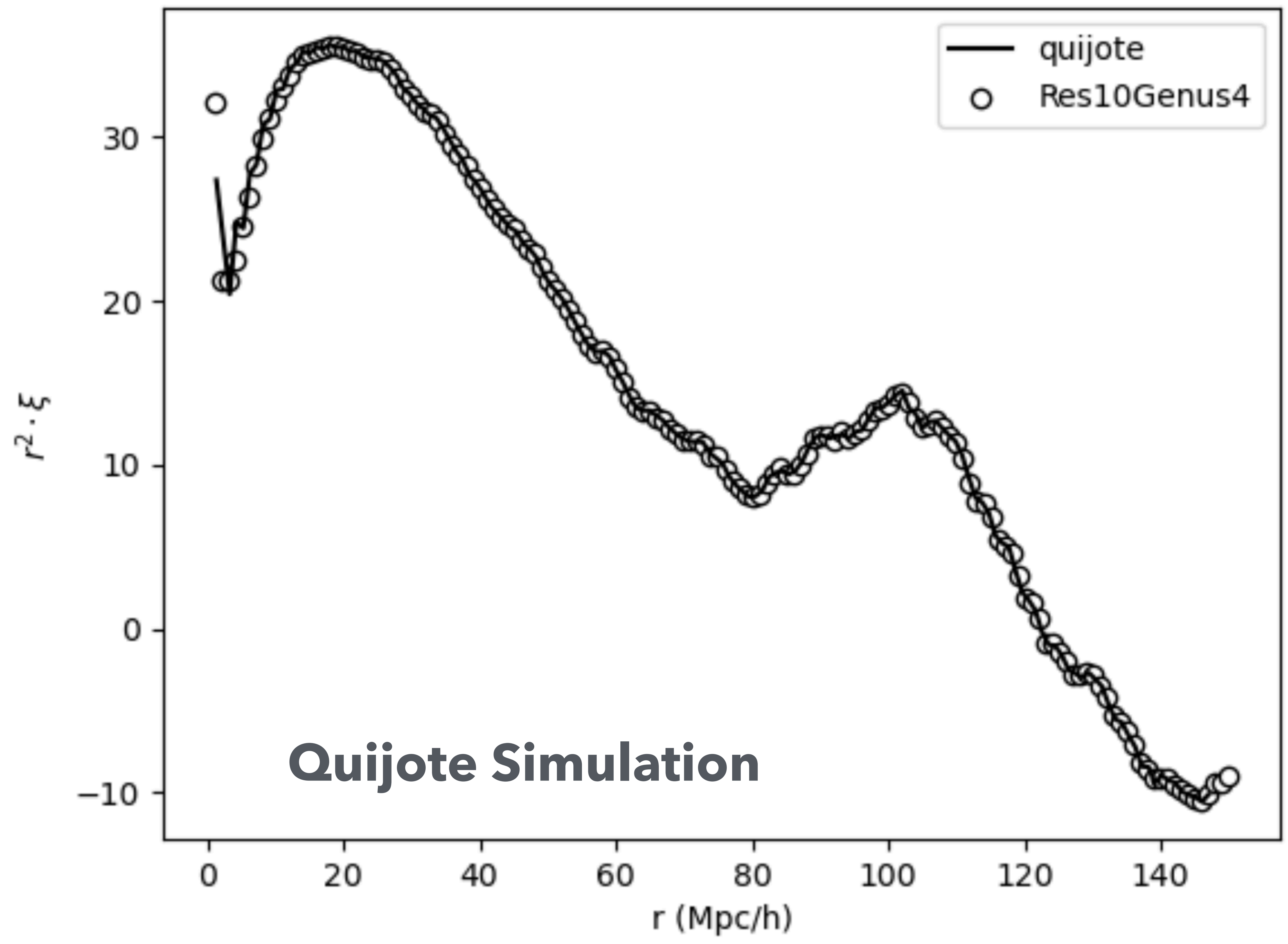
Parseval's theorem

Wavenumber Space

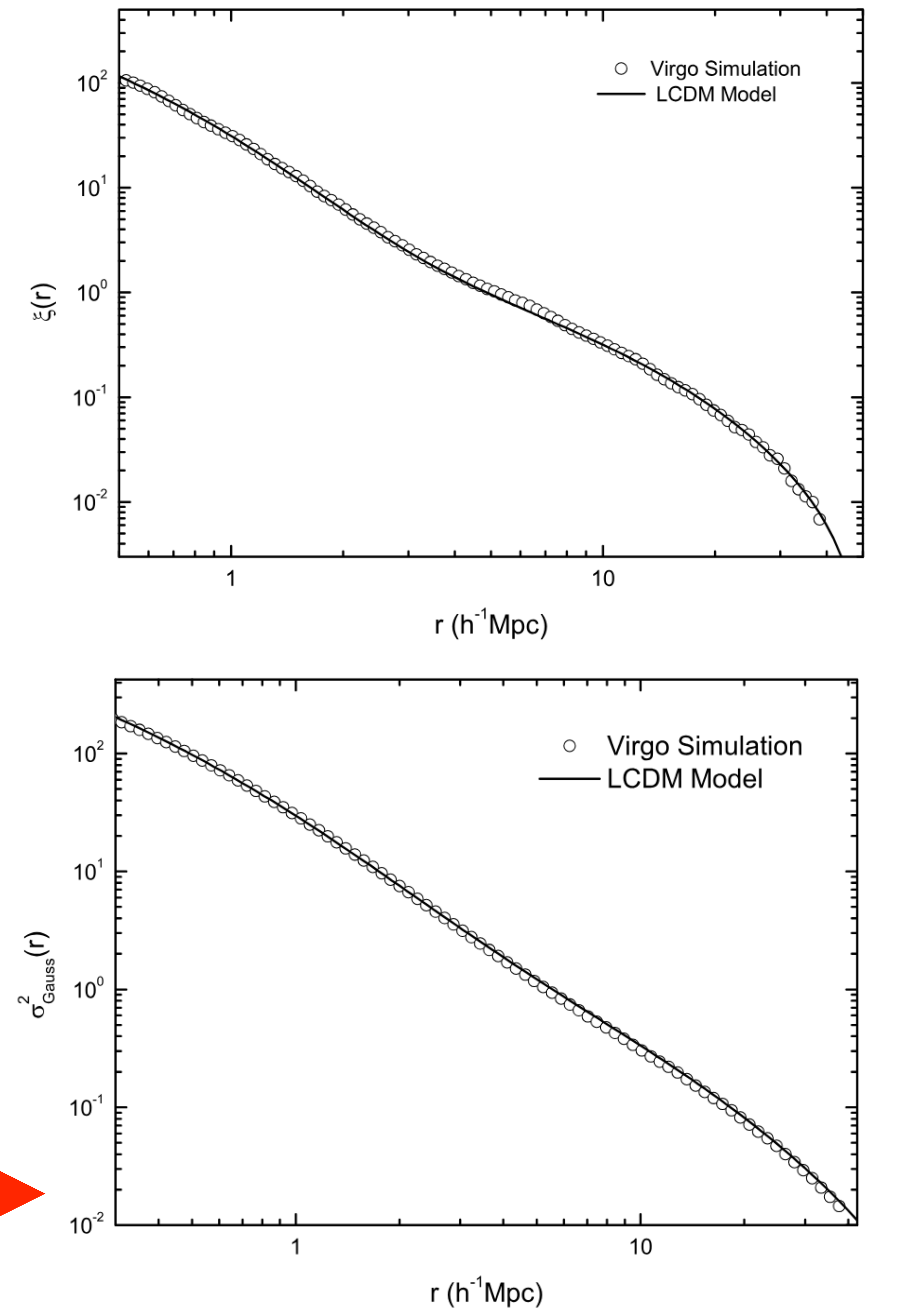
$$DD = \epsilon \cdot \tilde{\epsilon} = \sum_{\mathbf{k}} W_{\mathbf{k}} |\epsilon_{\mathbf{k}}|^2 \qquad O(N_g)$$

Pair-Counting without Counting

Numerical Tests

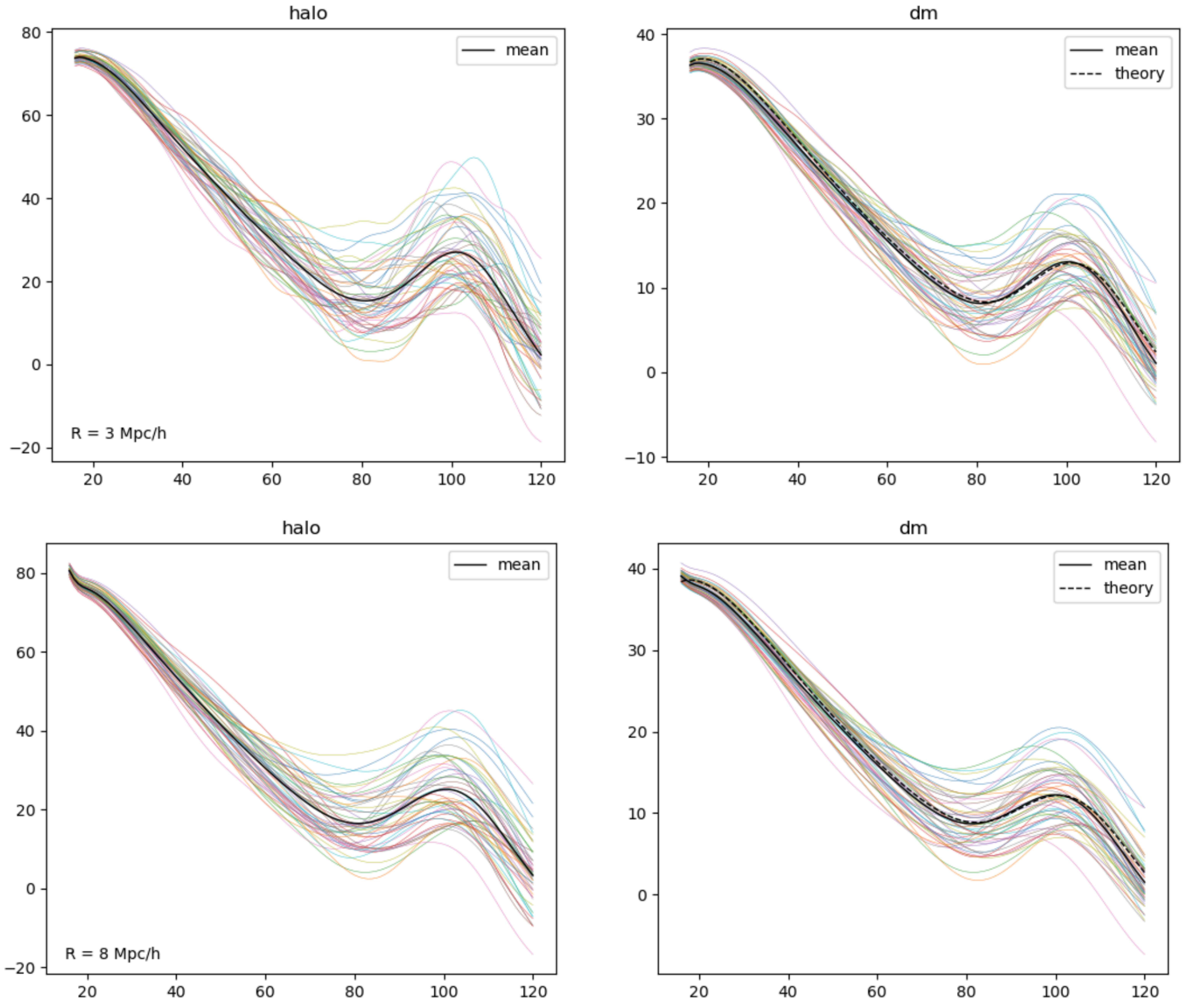


$$\sigma^2(\cdot) = \langle \delta_W^2(\cdot) \rangle = \frac{1}{(2\pi)^3} \int |W_{\text{filter}}(\boldsymbol{k}, \cdot)|^2 P(k) d^3\boldsymbol{k}$$

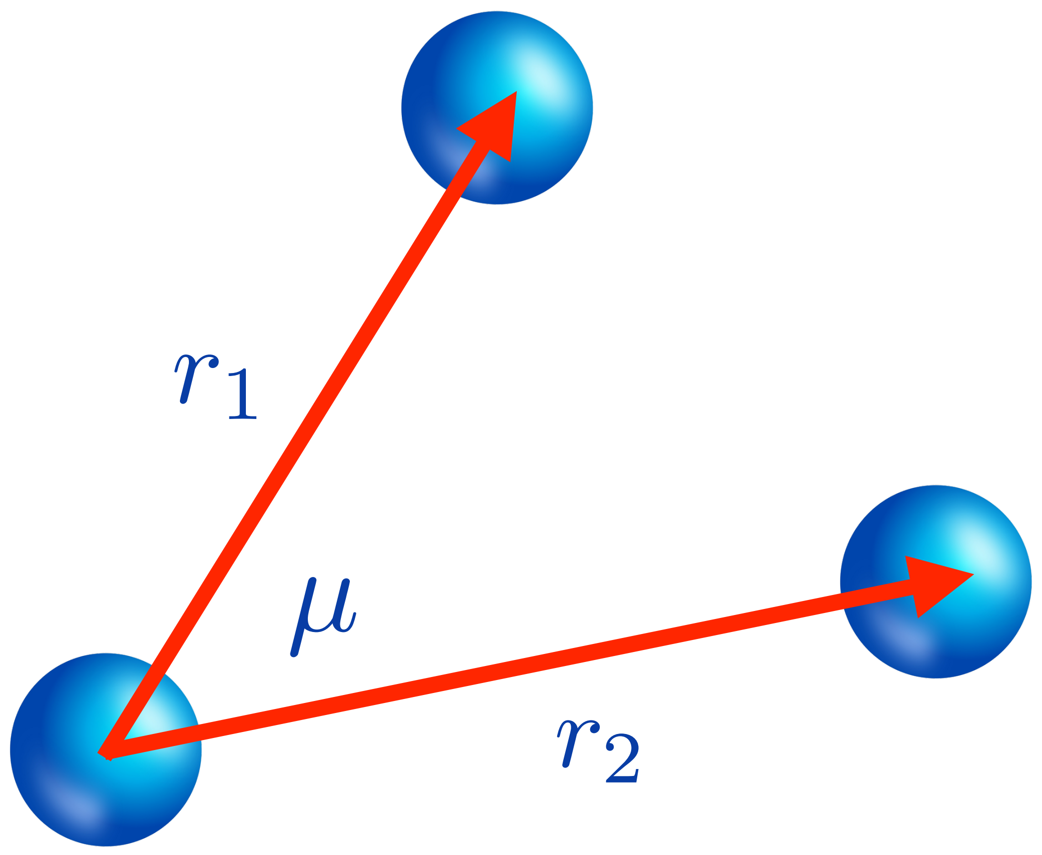


Numerical Tests

Quijote Simulation



Filtered Three-Point Correlation Function



$$\zeta(r_1, r_2, \mu) = \sum_{lmn} (-1)^{m+n} C_{lmn} P_n(\hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2) \int \frac{k_1^2 dk_1}{2\pi^2} \frac{k_2^2 dk_2}{2\pi^2} W(k_1, R) W(k_2, R) j_n(k_1 r_1) j_n(k_2 r_2) G_m(k_1, k_2, R) B_l(k_1, k_2)$$

Bi-Spectrum

$$B(k_1, k_2, \mu) = \sum_l B_l(k_1, k_2) P_l\left(\hat{\mathrm{k}}_1 \cdot \hat{\mathrm{k}}_2\right)$$

$$C_{lmn} = (2m+1)(2n+1) \begin{pmatrix} l & m & n \\ 0 & 0 & 0 \end{pmatrix}^2$$

Top-hat

$$G_m(k_1, k_2, R) = 2\pi^2 \frac{k_2 R J_{m-\frac{1}{2}}(k_2 R) J_{m+\frac{1}{2}}(k_1 R) - k_1 R J_{m-\frac{1}{2}}(k_1 R) J_{m+\frac{1}{2}}(k_2 R)}{\left(k_1^2 - k_2^2\right)(k_1 k_2)^{\frac{1}{2}}}$$

Gaussian

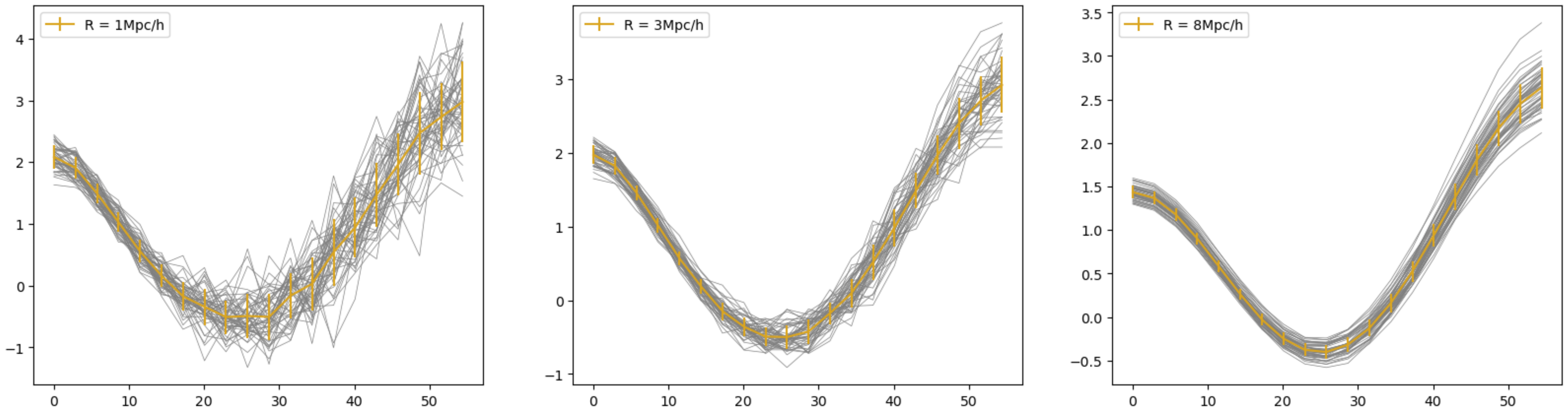
$$G_m(k_1, k_2) = e^{-\frac{1}{2}(k_1^2 + k_2^2) R^2} \sqrt{\frac{\pi}{2 k_1 k_2 R^2}} \mathrm{I}_{m+\frac{1}{2}}\left(k_1 k_2 R^2\right)$$

Yue et.al 2024 in prep.

Measuring Three-Point Correlation Function in Halos

- **Quijote Halo: 406793 halos**
- **Hermes Brute Force Calculation: triangles on halos + $4\pi r_s^2 R\bar{n}$ spatial rotation + 256^3 grid (J=8) + Daubechies 4 scaling function**
- **Szapudi & Szalay Estimator:** $\hat{\xi}_{\rm N} = \frac{\prod_{i=1}^{N}(D_i - R_i)}{\prod_{i=1}^{N} R_i} = \frac{(D-R)^N}{R^N}$ $\quad (r_{12}, r_{13}) = (20, 40)h^{-1}Mpc$
- **Computing Server: Intel 128 Cores**

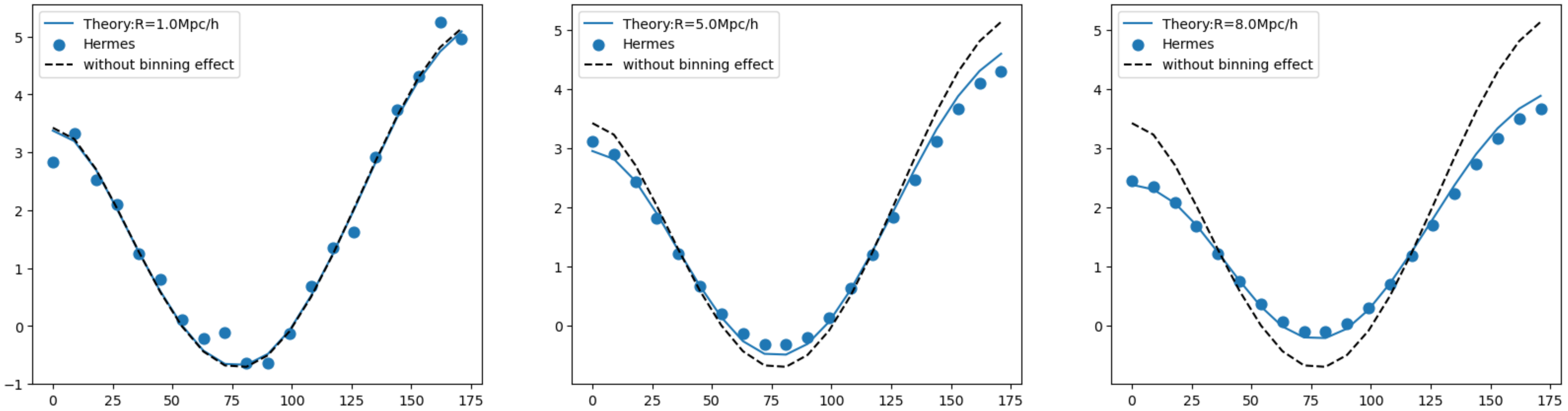
1.8 seconds for one data point



Measuring Three-Point Correlation Function in DM

Preliminary

- **MDPL2 DM :** 2.8×10^8 **(0.05%) Particles**
- **Hermes Brute Force Calculation:** 10^9 **randomly placed triangles +** $\sim 10^3$ **spatial rotation +** 1024^3 **grid (J=10) + Daubechies 4 scaling function**
- **Szapudi & Szalay Estimator:** $\hat{\xi}_{\mathrm{N}} = \dfrac{\prod_{i=1}^{N}(D_i - R_i)}{\prod_{i=1}^{N} R_i} = \dfrac{(D-R)^N}{R^N}$ $\quad (r_{12}, r_{13}) = (20, 40)h^{-1}Mpc$
- **Computing Server: Intel 128 Cores** **5 minutes for one data point**



Summary

Hermes: HypER–speed MultirEsolution cosmic Statistics

- An open-source, massively parallel & GPU accelerated Python toolkit for cosmic statistics
- $N_g \log N_g$ Algorithm, independent of number of sampling points
- Making a unified scheme for all variants of clustering statistical measures

Hermes v1.0 will be publicly available on July 2024

Thank You!