The Fast Algorithm of Clustering Statistics for a Large Data Set

Long-Long Feng
Purple Mountain Observatory

ICTP, Trieste, June 4 2005

Outline

- Discrete Wavelet Transformation (DWT)
- DWT Power Spectrum
- Constructing the Power Spectrum
- Unequally Spaced FFT
- Applications

An Example: Haar Wavelet

One-dimension density fluctuation in [0,L]:

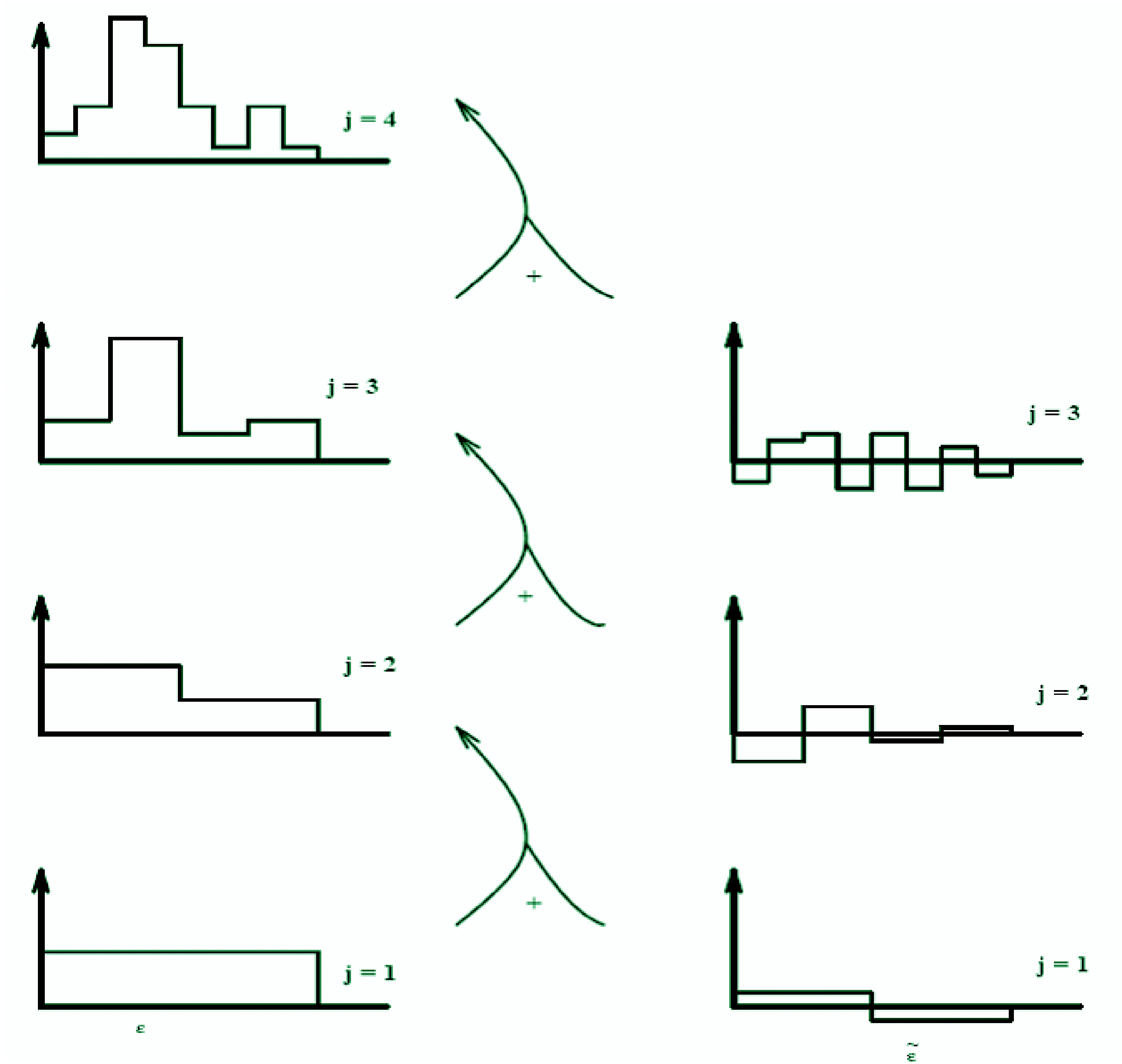
$$\epsilon(x) = \frac{\rho(x) - \bar{\rho}}{\bar{\rho}}$$

1 ϵ can be approximated by a histogram of 2^J bins, $L = 2^J \Delta x$.

2 The histogram is labeled by l running from 0 to $2^J - 1$. The bin covers a interval l to $(l+1)\Delta x$

3 The sample is described by 2^J $\epsilon_{J,l}$ defined by

$$\epsilon_{J,l} = \frac{1}{\Delta x} \int_{l\Delta x}^{(l+1)\Delta x} \epsilon(x) dx$$



j = 4
+
j = 3
j = 3
+
j = 2
j = 2
+
j = 1
j = 1
ε
ε̃

FT & DWT

A periodic function f

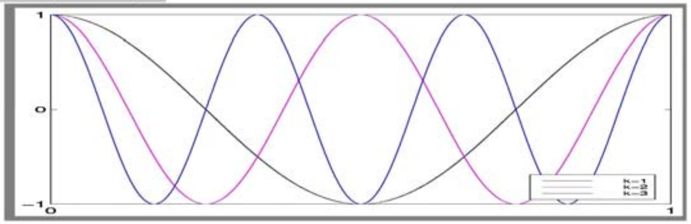
$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{i2\pi kx}$$

Fourier coefficients

$$c_k = \int_0^1 f(x) e^{-i2\pi kx}\, dx$$

$$e^{i2\pi kx} = \cos(2\pi kx) + i \sin(2\pi kx)$$

$\cos(2\pi kx)$



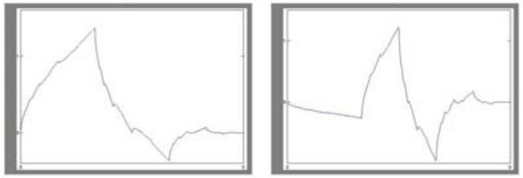
$$f(x) = \sum_{k=0}^{2^{J_0}-1} c_{J_0,k}\phi_{J_0,k}(x) + \sum_{j=J_0}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(x)$$

Two kinds of basis functions

Scaling functions: $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$

Wavelets: $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$

ϕ ψ



Scaling function coefficients

$$c_{J_0,k} = \int_{-\infty}^{\infty} f(x)\phi_{J_0,k}(x)\, dx$$ Averages

Scaling function coefficients

$$d_{j,k} = \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x)\, dx$$ Details

FT & DWT Power Spectrum Estimators

Fourier Representation:

$$\varepsilon(x) = \sum_n \varepsilon_n e^{i2\pi nx/L}$$

$$\varepsilon_n = \frac{1}{L}\int_0^L \varepsilon(x) e^{-i2\pi nx/L} dx$$

DWT Decomposition:

$$\varepsilon(x) = \sum_{j=0}^{\infty} \sum_{l=1}^{2^j} \varepsilon_{jl} \psi_{jl}(x)$$

$$\varepsilon_{jl} = \int_0^L \varepsilon(x) \psi_{jl}(x) dx$$

FT Power Spectrum:

$$P(n) = \left| \varepsilon_n \right|^2$$

DWT Power Spectrum

$$P_j = \frac{1}{2^j} \sum_{l=0}^{2^j - 1} \left| \varepsilon_{j,l} \right|^2$$

FT vs. DWT Power Spectrum

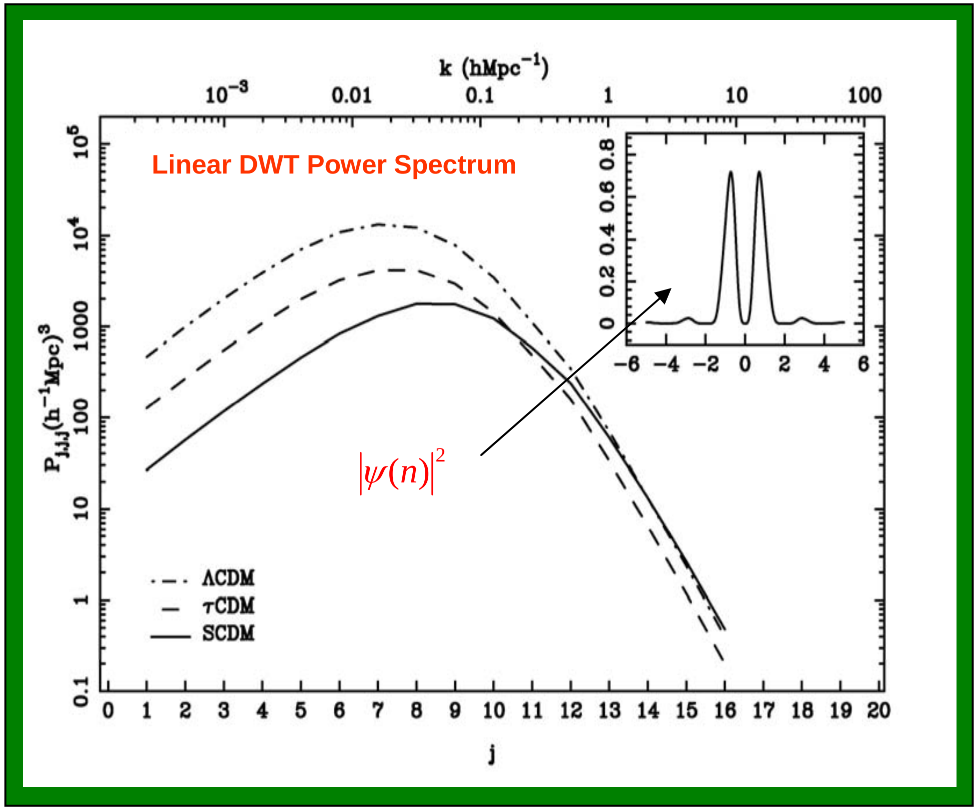
- Parseval's Theorem:

$$\frac{1}{L}\int_0^L \delta^2(x)dx = \sum_n |\delta_n|^2 = \sum_{j=0}^{\infty}\frac{1}{L}\sum_{l=0}^{2^j-1}\varepsilon_{j,l}^2$$

- DWT via Fourier Power Spectrum:

$$P_j = \frac{1}{2^j}\sum\left|\psi(\frac{n}{2^j})\right|^2 P(n) = \sum_n W_j(n)P(n)$$

DWT Power Spectrum: Example



- **Band Average FT Power Spectrum**

$$P_j = \sum_n W_j(n) P(n)$$

$$W_j(n) = \frac{1}{2^j} \left| \psi\left(\frac{n}{2^j}\right) \right|^2$$

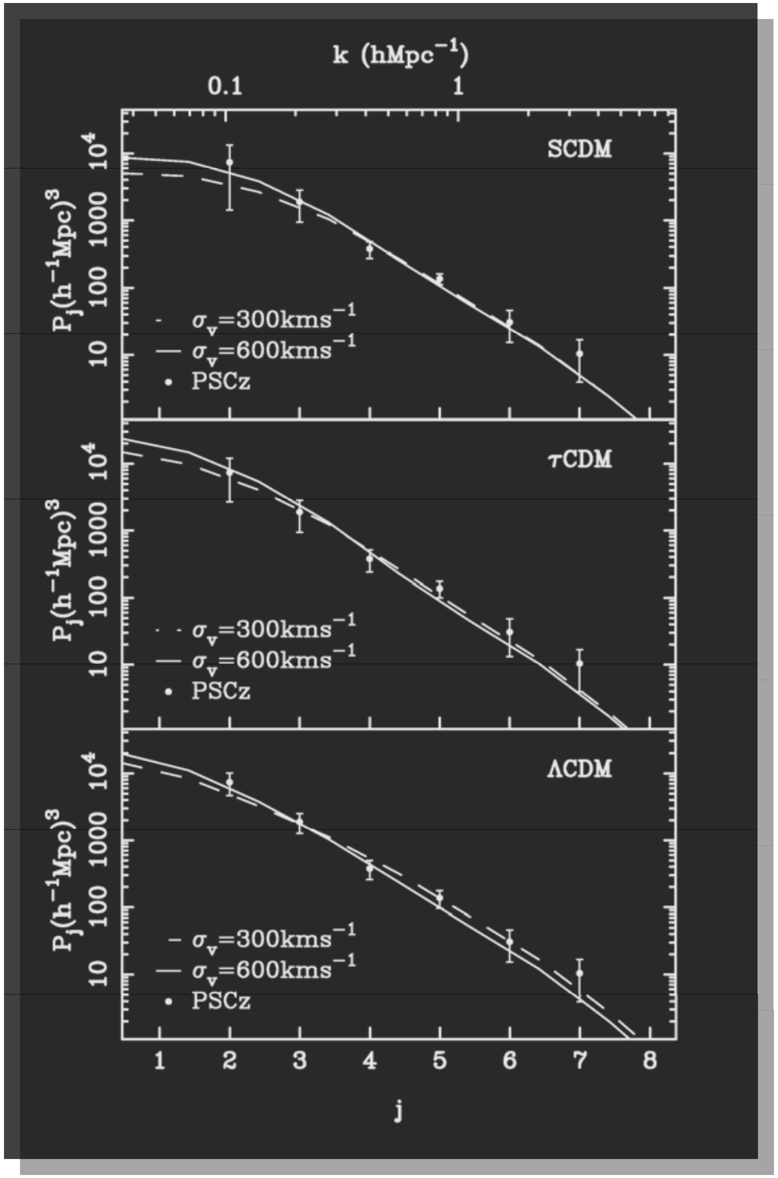
- **In the Logarithmic Spacing of n**

$$\log n = (\log 2) j + \log n_p$$

$$\Delta \log n = \Delta n_p \,/\, n_p$$

Fang & Feng, 2000

Measuring the DWT Power Spectrum in the PSCz



The IRAS Point Source Catalog Redshift Survey (PSCz): covers 84% of the sky with a total of about 15000 redshifts. The median redshift is 8500kms^{-1}. The selection function could take a parametric form by

$$\Phi(r) = \Phi^* \left(\frac{r}{r_*}\right)^{-\alpha} \left[1 + \left(\frac{r}{\beta r_*}\right)^{\gamma}\right]^{-\frac{\beta}{\gamma}}$$

$$\Phi^* = 0.0077, r_* = 86.4$$

$$\alpha = 1.82, \beta = 4.43, \gamma = 1.56$$

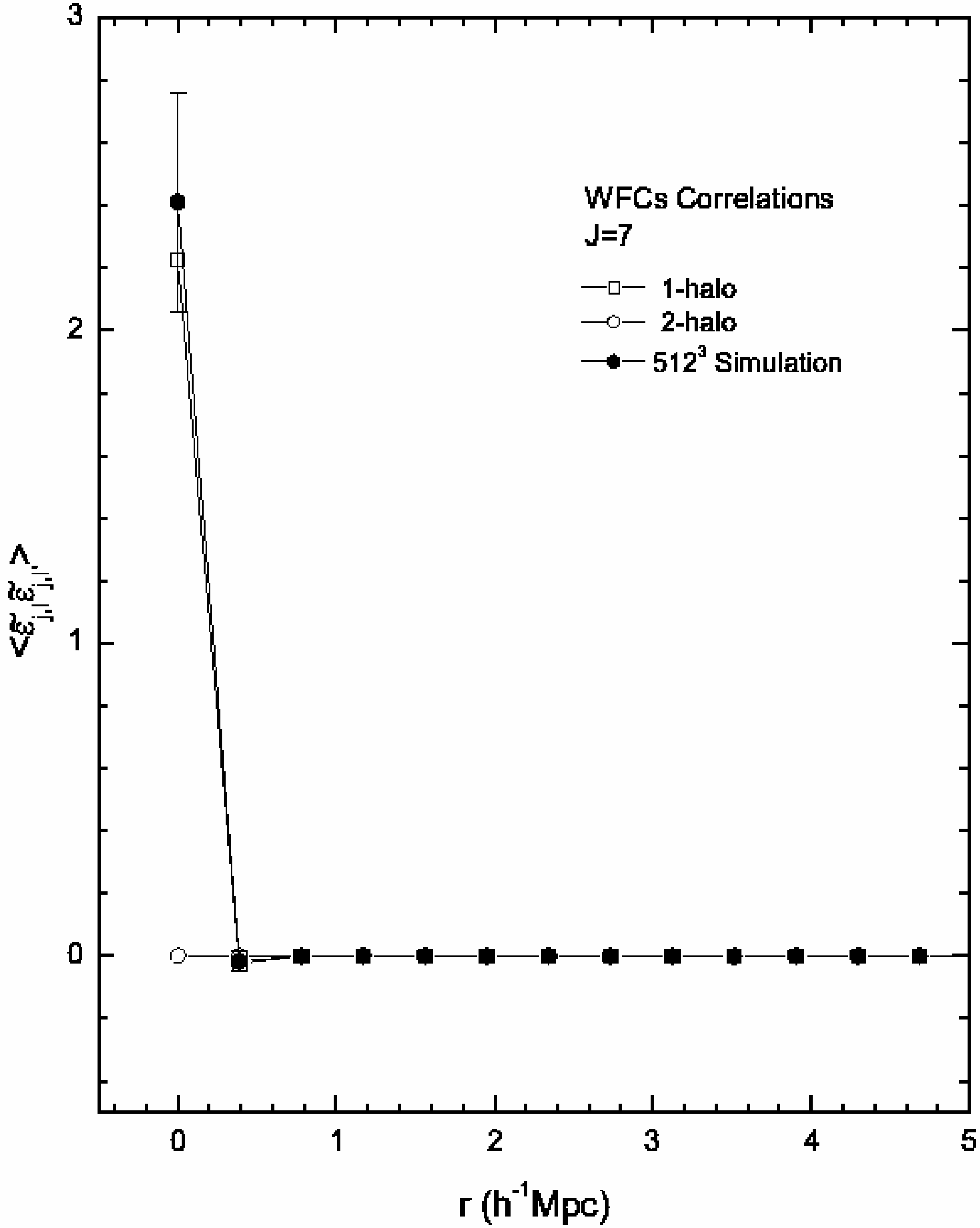
The Model Fitting Parameters

Model	b	β	σ_v (km/s)
SCDM	1.89 ± 0.15	0.53 ±0.04	600
τCDM	2.12 ± 0.16	0.47 ±0.03	600
ΛCDM	1.17 ± 0.10	0.42 ±0.04	600

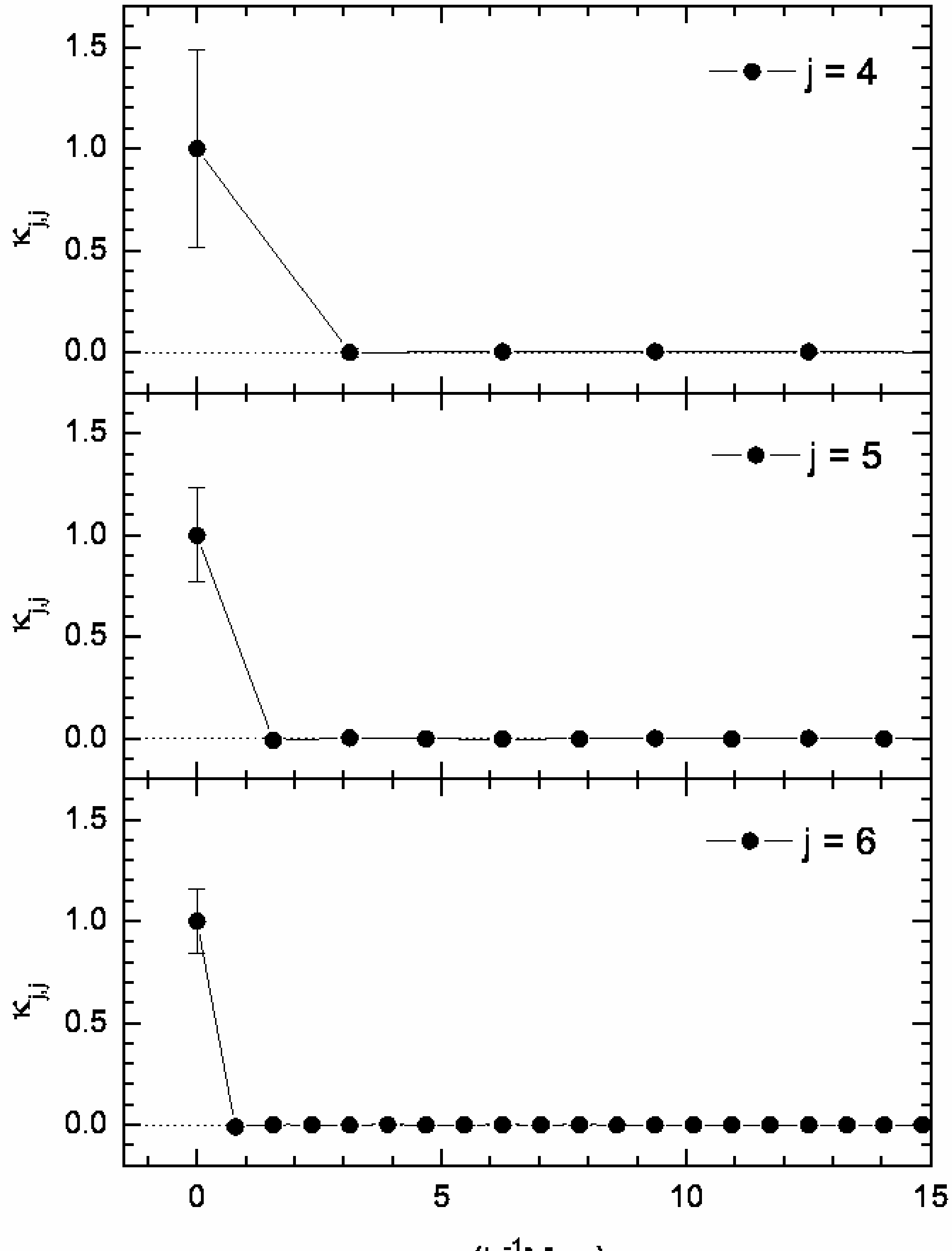
Quasi-Local Evolution of the Cosmic Density Fluctuations

- In the linear regime, the different *k-modes* evolve independently.
- In the non-linear regime, the *{j.l}-modes* at different positions are uncorrelated?

covariance matrix of {j,l}-modes?



WFCs Correlations
J=7
1-halo
2-halo
512^3 Simulation
r (h^-1Mpc)



j = 4
j = 5
j = 6
κ_{j,j}
1.5
1.0
0.5
0.0
0
5
10
15
r (h^-1Mpc)

Constructing the Power Spectrum

DWT → FT Power

The covariance matrix of WFCs is nearly diagonal

$$P(n) = \sum_{j=0}^{\infty} P_j |\psi(n/2^j)|^2$$

The FT power spectrum is determined by a finite set of DWT power spectrum {P_j,j=0,1,2,…..}

The Unequally Spaced FFT

We intend to evaluate a trigonometric sums

$$g_n = \sum_{i=1}^{N} w_i \exp(i2\pi \mathbf{n} \cdot \mathbf{x}_i)$$

Alternatively, written by an integral

$$g_{\mathbf{n}} = \int n_g(\mathbf{x}) \exp(i2\pi \mathbf{n} \cdot \mathbf{x}) d\mathbf{x}$$

with

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

Beylkin G., 1995; 1998; Fang & Feng, 2000

Algorithm

- *step 1*: compute the decomposition coefficients using mth B-spline function,

$$\epsilon_{\mathbf{l}} = \sum_{i=0}^{N} w_i \beta^m (2^j \mathbf{x}_i - \mathbf{l})$$

- *step 2*: evaluate the sum using FFT

$$\tilde{g}_{\mathbf{n}} = \frac{1}{\sqrt{a(\mathbf{n}/2^j)}} \sum_{k=0}^{2^j-1} \epsilon_{\mathbf{k}} e^{2\pi i \mathbf{n}\cdot\mathbf{k}/2^j}$$

where

$$a^m(\xi) = \sum_{l=-m}^{l=m} \beta^{(2m+1)}(l) e^{2\pi i l \xi}$$

The Underlying Math.

Lemma: let w be a test function, then

$$w(x_0) - \int p_0(x)w(x)dx = O(h^{(m+1)}), \qquad h = 2^{-j}$$

where

$$p_0(x) = \sum_{k \in \mathbf{Z}} \phi_{jk}^m(x_0)\phi_{jk}^m(x)$$

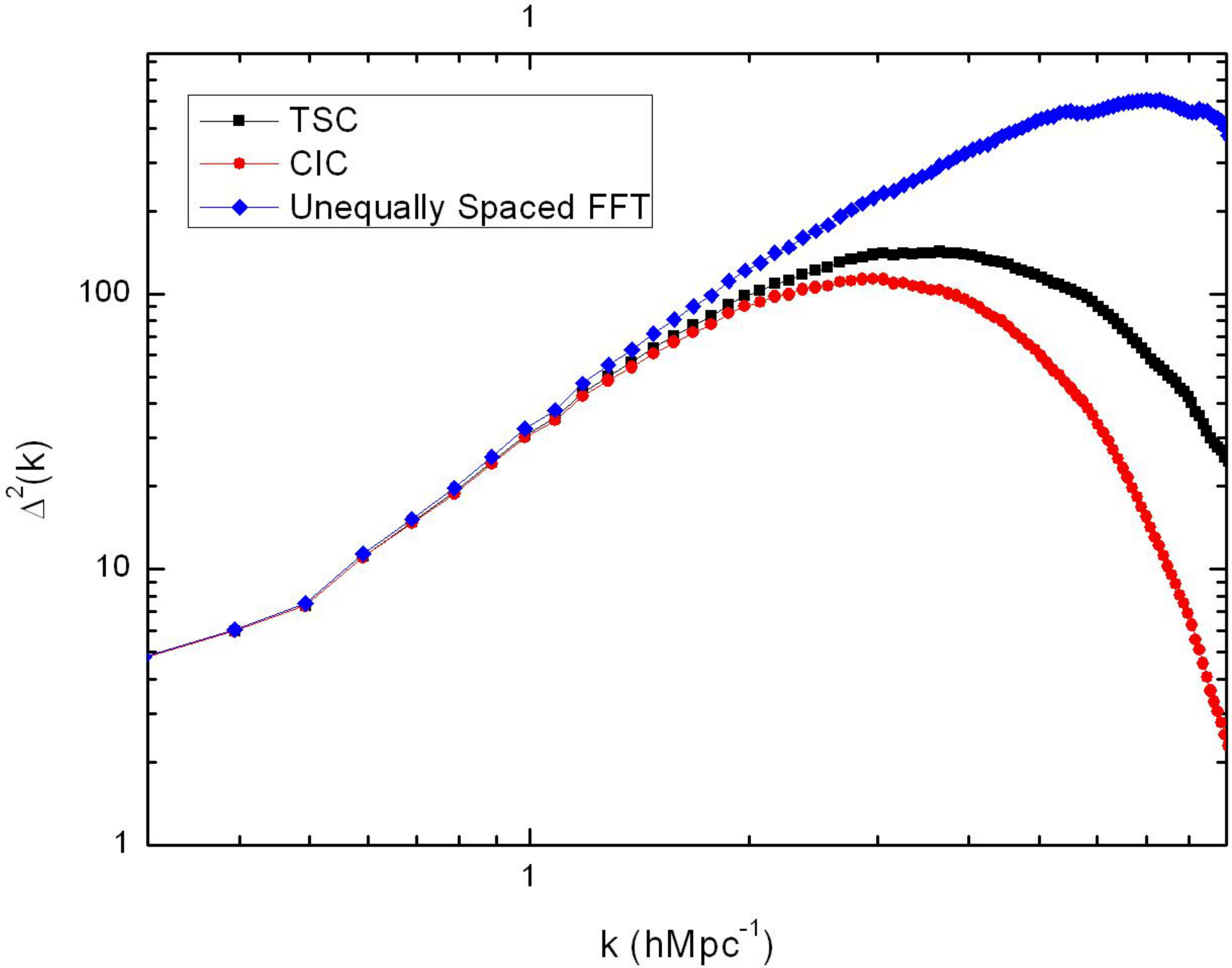
So, the particle distribution could be approximated by

$$n_g(\mathbf{x}) = \sum_{i=1}^{N} w_i \delta^3(\mathbf{x} - \mathbf{x_i}) \approx \sum_{k \in \mathbf{Z}} \epsilon_k \phi_{jk}^m(\mathbf{x})$$

with

$$\epsilon_{\mathbf{k}} = \sum_{i=0}^{N} w_i \phi^m(2^j \mathbf{x}_i - \mathbf{k})$$

The FT Power Spectrum Measured in the simulation sample (Virgo)



Applications

Convolving n(x) with a kernel G(x) ⟶ a fast algorithm of summation:

$$\sum_{i=1}^{N} G(x - x_i)$$

- G(x): spherical top hat counting particle–pairs,
- G(x): cubic top hat count-in-cell
- G(x)=1/|**x**| P-P gravity


Thank you !
天文臺