What does the N-point function hierarchy of the cosmological matter density field really measures ?

Julien Carron

University of Sussex
with I. Szapudi, Melody Wolk, Mark Neyrinck, Andrew Repp...

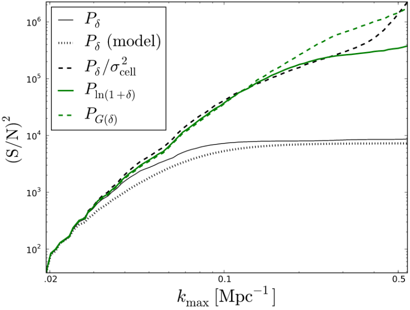
Sept 10, 2015, Warsaw

Why study cosmological information in the density field ?

- Unlike for CMB, standard methods based on the power spectrum extract only a fraction of the available information from LSS or WL surveys. Non-Gaussianity due to gravity.
- Predictions are difficult + Poisson or shape noise, RSD... Using only linear modes throws away most of the data.
- Good news: many ($\propto k^3$, or ℓ^2) high k modes are available
- Bad news: they may not contain much independent information: loss of information due to the tri-spectrum, and super survey modes.
- Mainstream idea is to use higher order statistics (3pt function, etc.)

How can we understand really the best descriptors and observables of LSS ? Tools : study qualitative properties of $\ln p$ and $\partial_\alpha \ln p$, as set by the dynamics.

- Cosmic variance going crazy : number of modes weighted by covariance matrix (here Neyrinck 2011)



- Neyrinck 2011, Mohammed Seljak 2014, JC et al 2015, Repp et al 2015.

$$Cov_{kk'} \sim \frac{2P^2(k)}{N_k} + \sigma^2_{\rm min} P(k)P(k'), \quad \sigma^2_{\rm min} = \sigma^2_{IS}(k_{\rm max}) + \sigma^2_{SS}(k_{\rm min})$$

- Measurement of spectrum amplitude limited by cosmic variance $\sim 1\%$ in Gpc^3/h^3 volume, irrespective of $k_{\rm max}$.

How can I guess which observable is most interesting to me ?

- Just look at $\partial_\beta \ln p$. (always a 'sufficient' statistic, in the frequentist sense).
- Gaussian fields $\ln p$: second order polynomial in the field -> 2 pt statistics are sufficient.
- Time dependence set by the dynamics

$$\frac{dp}{dt} = \partial_t p + \nabla(Fp) = 0,$$

describing how information in the initial conditions flows from two-point statistics to higher orders (and beyond). Difficult.

We restrict ourselves to the one-point density PDF as function of smoothing scale, where we can solve fully the problem (JC Szapudi 2013).

- Perturbation theory with Gaussian initial conditions predicts hierarchical structure of the cumulants

$$\langle \delta^p \rangle_c = \sigma^{2(p-1)} S_p + \text{loop corrections}$$

$$S_3 = \frac{34}{7} - (n+3), \cdots$$

- Use the Edgeworth series expansion of $\ln p$ to obtain the sufficient observables (and information) $\rightarrow$ It is always possible to extract the first m terms of the total information with a $m+2$ polynomial in the field.

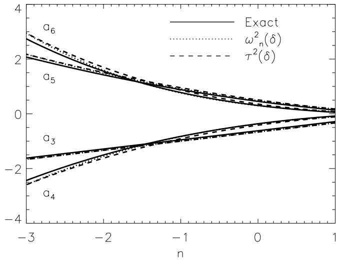
- General result is

$$\partial_{\sigma^2} \ln p(\delta) \simeq \delta^2 + \sum_{j\geq 3} a_j \delta^j \quad \text{with}$$

$$a_j = \frac{2}{j!} \sum_{\sum i k_i = j-2} (-1)^{|\mathbf{k}|} (j - 2 + |\mathbf{k}|)! \prod_{i\geq 3} \frac{S_i^{k_i}}{(i-1!)^{k_i}\, k_i!},$$

$$a_3 = -\frac{S_3}{3}, \quad a_4 = \frac{1}{12}\left(3S_3^2 - S_4\right), \quad \cdots$$

- Square of the optimal 'local' transform. Only leading cumulants, and σ^2 independent !



Two almost perfect matches :

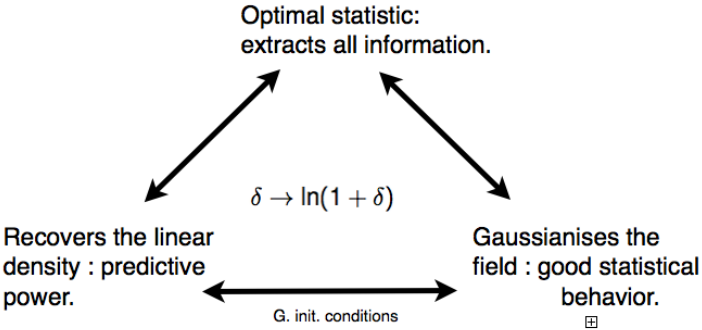
- The power transform

$$\omega_n(\delta) = \left(\frac{(1+\delta)^{(n+1)/3} - 1}{(n+1)/3} \right).$$

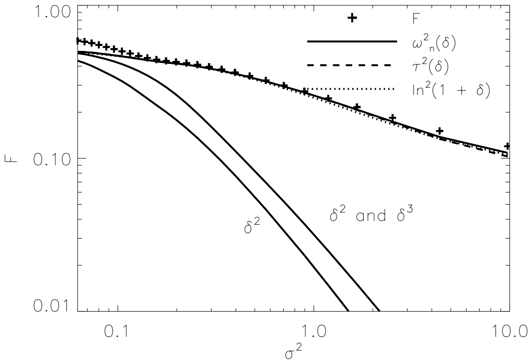
- The linear density contrast

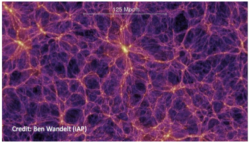
$$\tau(\delta) = \frac{3}{2}(1+\delta)^{(n+3)/6}\left[(1+\delta)^{-2/3} - 1\right]$$

$$\partial_\alpha \ln p(\delta) \simeq \tau^2(\delta) \simeq \left(\frac{(1+\delta)^{(n+1)/3} - 1}{(n+1)/3} \right)^2 \simeq \ln(1+\delta)^2$$



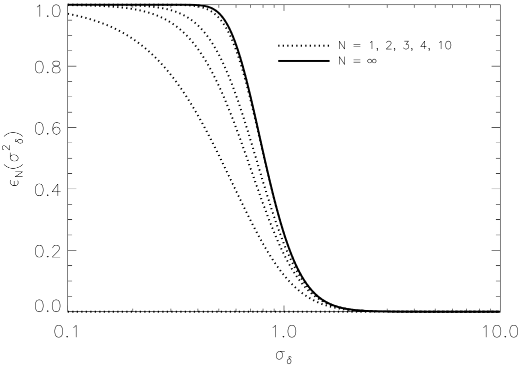
Tests to Millenium simulation density field





Similar to the lognormal PDF (JC 1105.4467)

- Lognormal model : $\ln(1+\delta)$ is Gaussian.
- Qualitatively very good : correct two-point function ξ_δ, $S_3 = 3$ instead of $34/7 - 2$, etc. Same sufficient statistic.



What's happening here ?

- Start with hierarchy of polynomials in the field

$$\delta^0(x), \delta(x), \delta^2(x), \cdots, \delta^N(x) \quad \left\langle \delta^i \delta^j \right\rangle = \Sigma_{ij}.$$

- Transform to an orthogonal basis set

$$P_0(\delta), P_1(\delta), \cdots, P_N(\delta) \quad \langle P_n P_m \rangle = \delta_{nm}$$

- Reproducing kernel :

$$K_N(\delta', \delta) = \sum_{n=0}^{N} P_n(\delta') P_m(\delta) = \sum_{ij=1}^{N} \delta'^i \left[\Sigma^{-1} \right]_{ij} \delta^j$$

$$\tilde{f}_N(\delta) = \int d\delta' p(\delta') f(\delta') K_N(\delta', \delta)$$

What if we include the entire hierarchy ? Two possibilities.

- *If* the polynomials form a *complete* basis, then they are able to resolve fully the field PDF:

$$K_N(\delta',\delta) \to \frac{\delta^D(\delta'-\delta)}{p(\delta)}, \quad \tilde{f}_N(\delta) \to f(\delta)$$

- If not, the kernel keeps a finite width, and information is lost ! This is the case expected for the density field (only asymptotic generating functions), due to sharp nonlinear growth.

$$\left\langle e^{-J\delta} \right\rangle = \int \mathcal{D}\delta_L e^{-\frac{1}{2}\frac{\delta_L^2}{\sigma^2} - J\delta[\delta_L]} \neq \sum_n \frac{(-J)^n \langle \delta^n \rangle}{n!}$$

Two consequences :

- Lack of predictive power ($\tilde{f}_N \neq f$) : N-point moments have access to only part of the theory. There are different fields with identical N-point functions at all orders (JC and Neyrinck 1201.1444).
- Inference on model parameter from $N-$point moments can become inefficient. Fisher info. $F_{\alpha\beta}$ in the moments strictly smaller than Fisher info. in the full field. This is a special case of $\tilde{f}_N \neq f$ with

$$f = \partial_\alpha \ln p$$

What does the hierarchy really measure ?

- We solved that problem fully in the model of the lognormal field (JC and Szapudi 1508.04838), for any two-point function $\langle \delta(x)\delta(y) \rangle$.
- In the decomposition of the log-density at each point

$$\ln(1+\delta(x)) = \theta(x) + k(x)\cdot 2\pi \quad \text{with } \theta \in [-\pi,\pi], \quad k \in \mathbb{Z},$$

measuring the hierarchy is equivalent to measuring $\theta(x)$ only.

Sketch of derivation

- It is possible for this model to calculate the Kernel K_N for $N = \infty$. It results not in the field Gaussian PDF, but a different field PDF, the *wrapped* Gaussian field :

$$p_W(\theta) = \frac{1}{(2\pi)^d} \sum_{\mathbf{n} \in Z^d} \exp\left(-\frac{1}{2}\mathbf{n}^t \xi_A \cdot \mathbf{n} - i\theta \cdot \mathbf{n}\right)$$

- It is the solution of the diffusion equation not in Euclidean space, but on the compact torus, thus with homogeneity as long-term solution, where memory of the initial conditions is completely lost.

Summary - Outlook

- There is nothing special about the two-point function beyond the linear regime. There is nothing special about the N-point functions beyond the mildly nonlinear regime. Information leaks out, unable to handle large deviations.
- Need for better behaved basis sets to describe the nonlinear regime.
- There is definitely something special about the log-density, which is to good approximation the best possible local transform of the field. The transform can be accommodated for discreteness.