# Unbiased contaminant removal for P(k) measurement
## arXiv:1607.02417




Ben Kalus

Will Percival

David Bacon

Lado Samushia

ICG, University of Portsmouth, UK
except Lado, who used to be at the ICG, but is now in Kansas

# Galaxy sample contaminants

We're interested in cosmological fluctuations in the galaxy distribution.

All surveys have window functions describing the spatial probability of finding a galaxy in a small volume element given no information about other galaxies.

Determining the expected spatial distribution of galaxies in a sample (with no cosmological fluctuations) is difficult

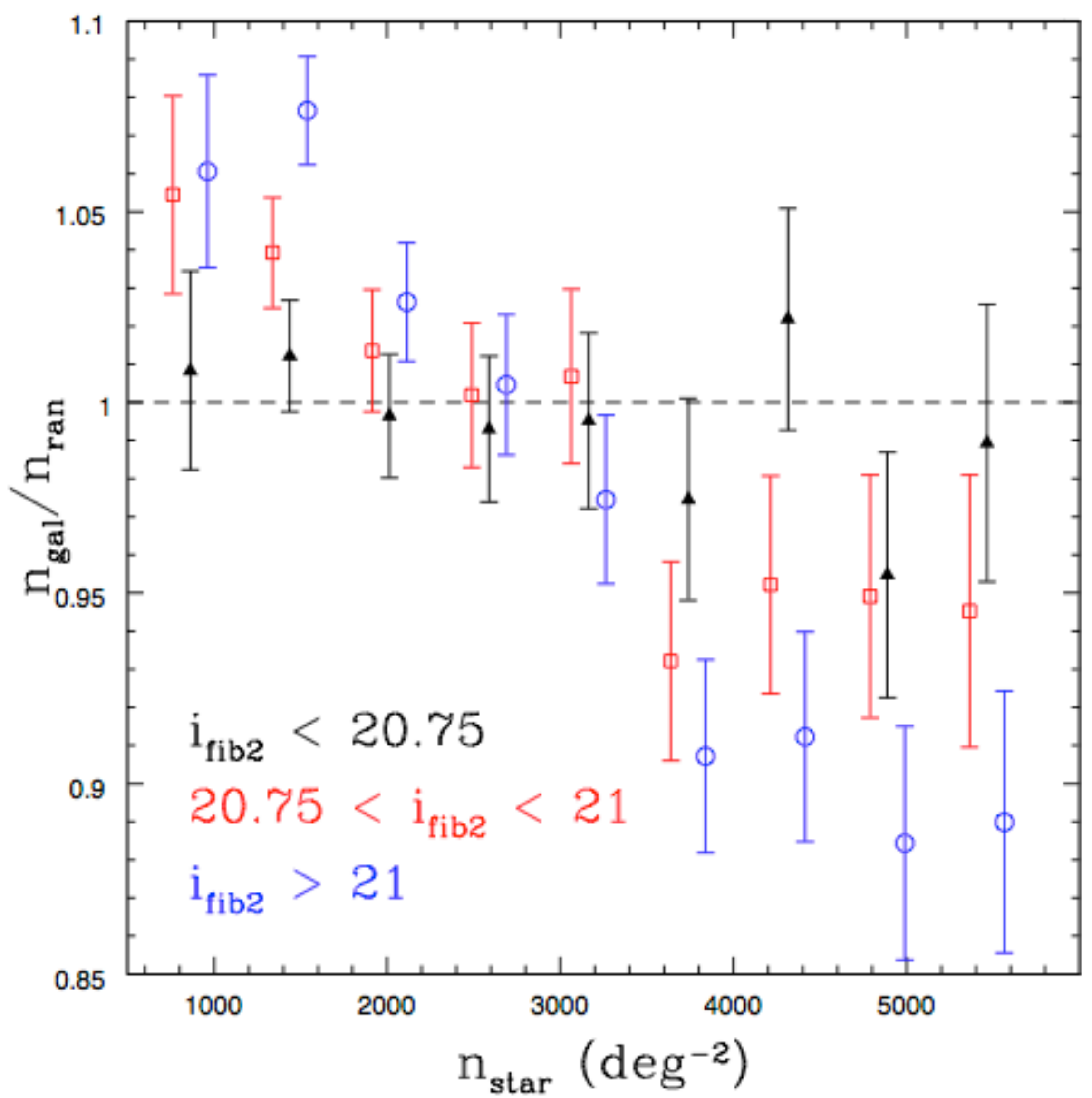
- i.e. cannot perfectly match galaxies and randoms

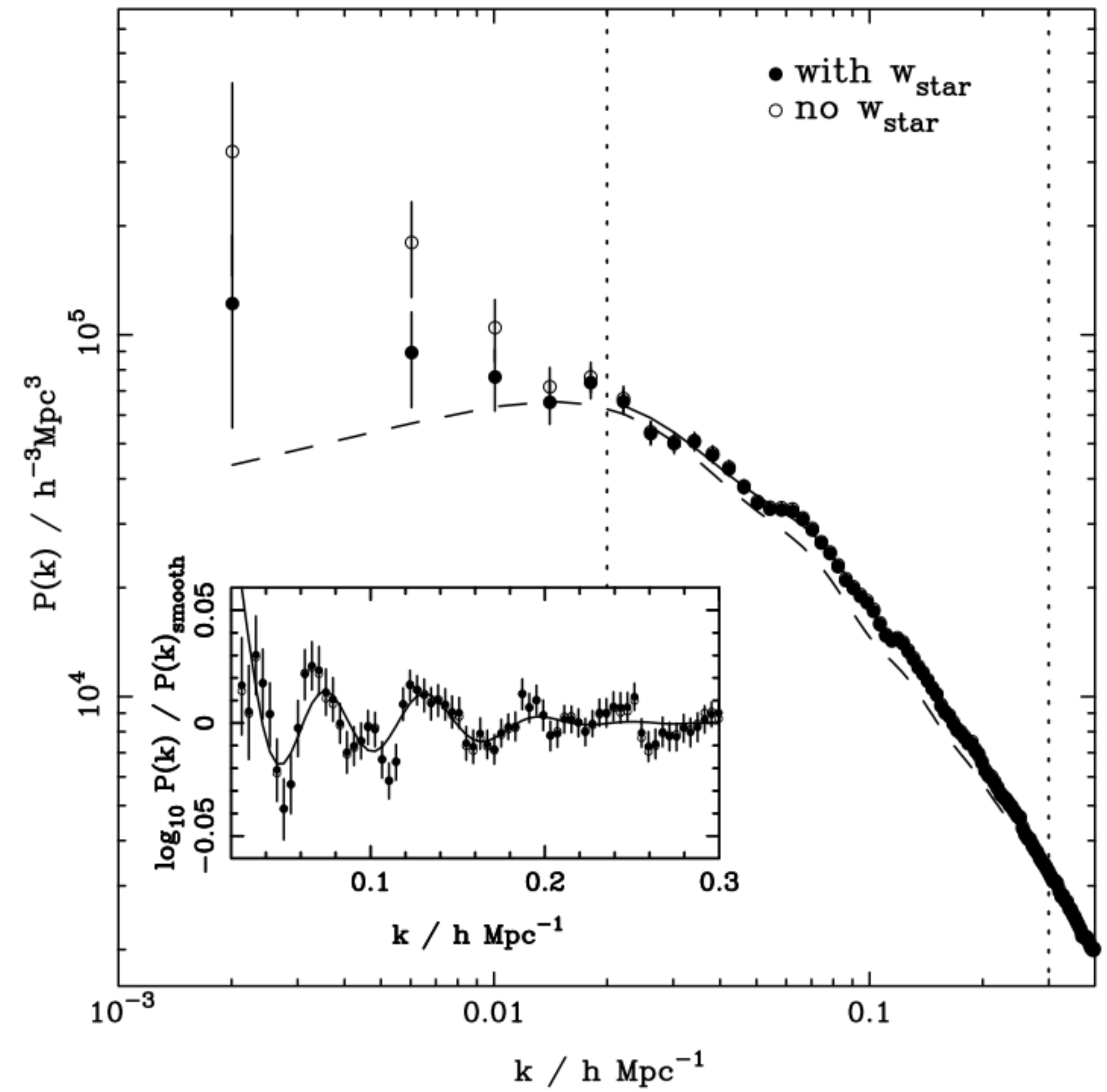
Basic problems:

- stars, extinction, observing conditions, differences across instruments
- for both the target sample, and spectroscopy.

# SDSS-III DR9 spectroscopic sample

DR9 (~1/3 of DR12) contained 264,283 galaxies, z ~ 0.57, and covered 2.2 Gpc²




Ross et al. 2012; arXiv:1203.6499

# Primordial non-Gaussianity: $f_{NL}$ $g_{NL}$

δ is sourced from a potential field Φ, whose form might not be Gaussian

$$\nabla^2\Phi(\mathbf{x}) = 4\pi G\delta(\mathbf{x})$$

$$\Phi(\mathbf{x}) \sim \phi(\mathbf{x}) + f_{NL}\phi^2(\mathbf{x}) + \ldots$$

skewness ~ $f_{NL}$
kurtosis ~ $f_{NL}^2$

Φ is a Gaussian field. the non-linear terms in Φ make Φ non-Gaussian. This map completely specifies Φ statistics.

Salopek and Bond 1990;
Gangui, Lucchin, Matarrese, Mollerach 1994;
Komatsu and Spergel 2001

$$\Phi(\mathbf{x}) \sim \phi(\mathbf{x}) + g_{NL}\phi^3(\mathbf{x}) + \ldots$$

skewness ~ 0
kurtosis ~ $g_{NL}$

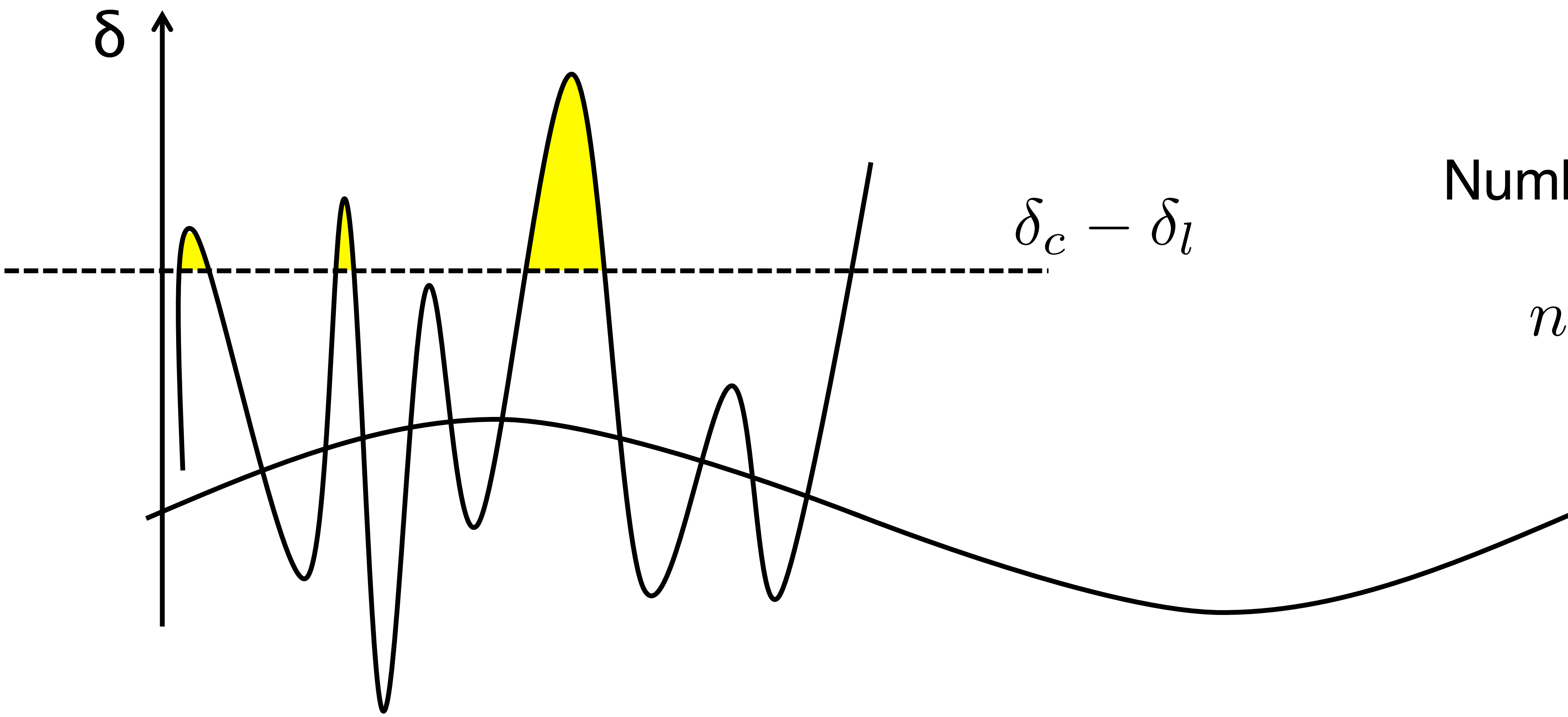
Okamoto and Hu 2002;
Enqvist and Nurmi 2005

$f_{NL}$ is not the only option for local potential fluctuations … you can go even further down this route …

Non-local models introduce non-trivial higher order correlations in **Φ**

# Peak-background split bias model

Halo formation much easier with additional long-wavelength fluctuation




Number density of halos

$$n \rightarrow n - \frac{dn}{d\delta_c}\delta_l$$

Leads to a revised large-scale halo density

$$(1+\delta_{\rm new}) = \left(1+\frac{\Delta n}{n}\right)(1+\delta_l)$$

Directly from the large-scale mode

From the change in Number of haloes

To first order, this leads to a bias

$$\delta_{\rm new} = \left(\delta_l + \frac{\Delta n}{n}\right) \qquad b = \frac{\delta_{\rm new}}{\delta_l} = 1 + \frac{\Delta n}{n\delta_l} = 1 - \frac{d\ln n}{d\delta_c}$$

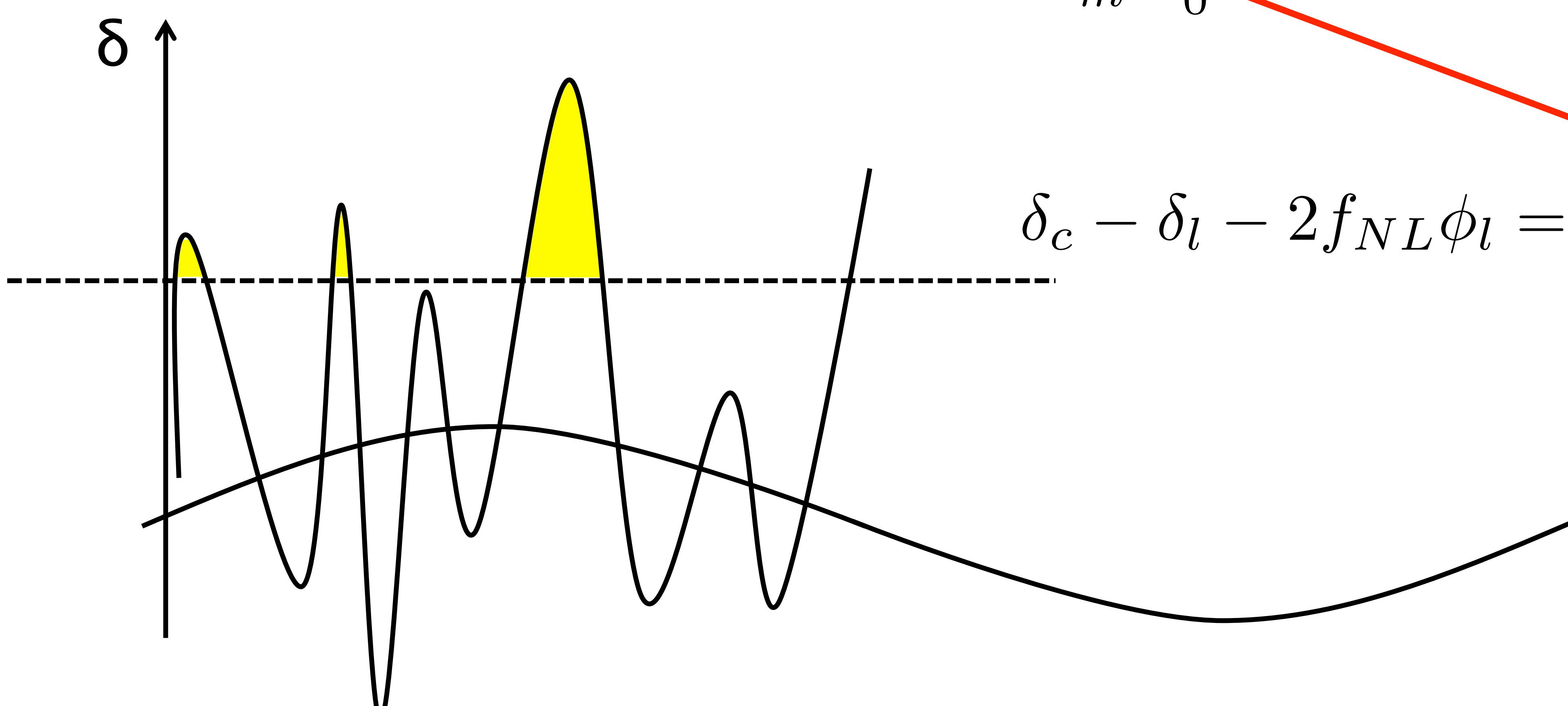
# This is altered by $f_{NL}$ signal

Now split ($\phi+f_{NL}\phi^2$) into long and short wavelength components

$$\Phi(\mathbf{x}) = \phi_l + \underbrace{f_{NL}\phi_l^2}_{\text{small}} + (1 + \underbrace{2f_{NL}\phi_l})\phi_s + f_{NL}\phi_s^2 + \text{cnst}$$

Link between potential and overdensity field shows how changing long wavelength potential component changes “critical density”

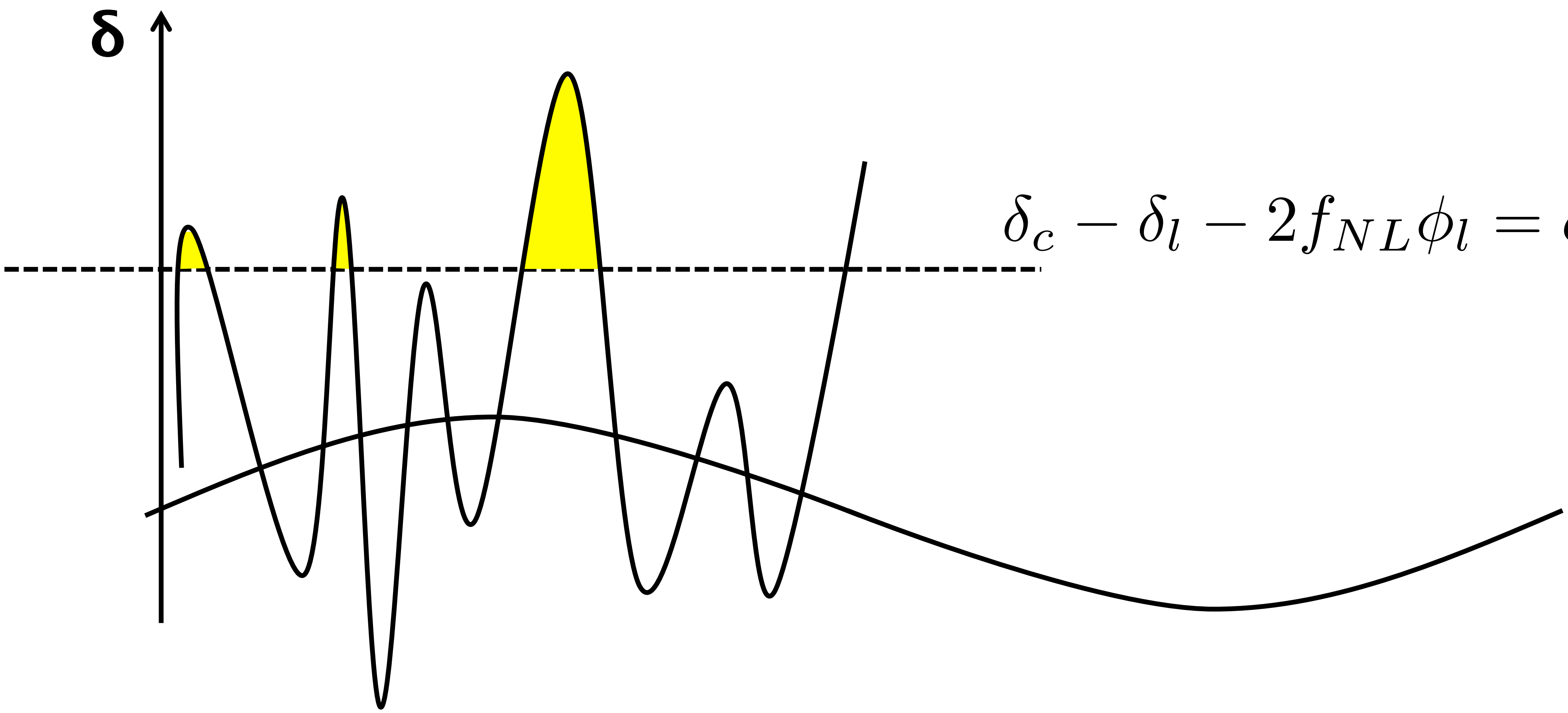
$$\delta_l(k) = \alpha(k)\Phi(k) \quad \alpha(k) = \frac{2c^2k^2T(k)D(z)}{3\Omega_m H_0^2}$$




$$\delta_c - \delta_l - 2f_{NL}\phi_l = \delta_c - \delta_l\left(1 + \frac{2f_{NL}}{\alpha(k)}\right)$$

# Peak-background split for non-Gaussianity

Halo formation much easier with additional long-wavelength fluctuation



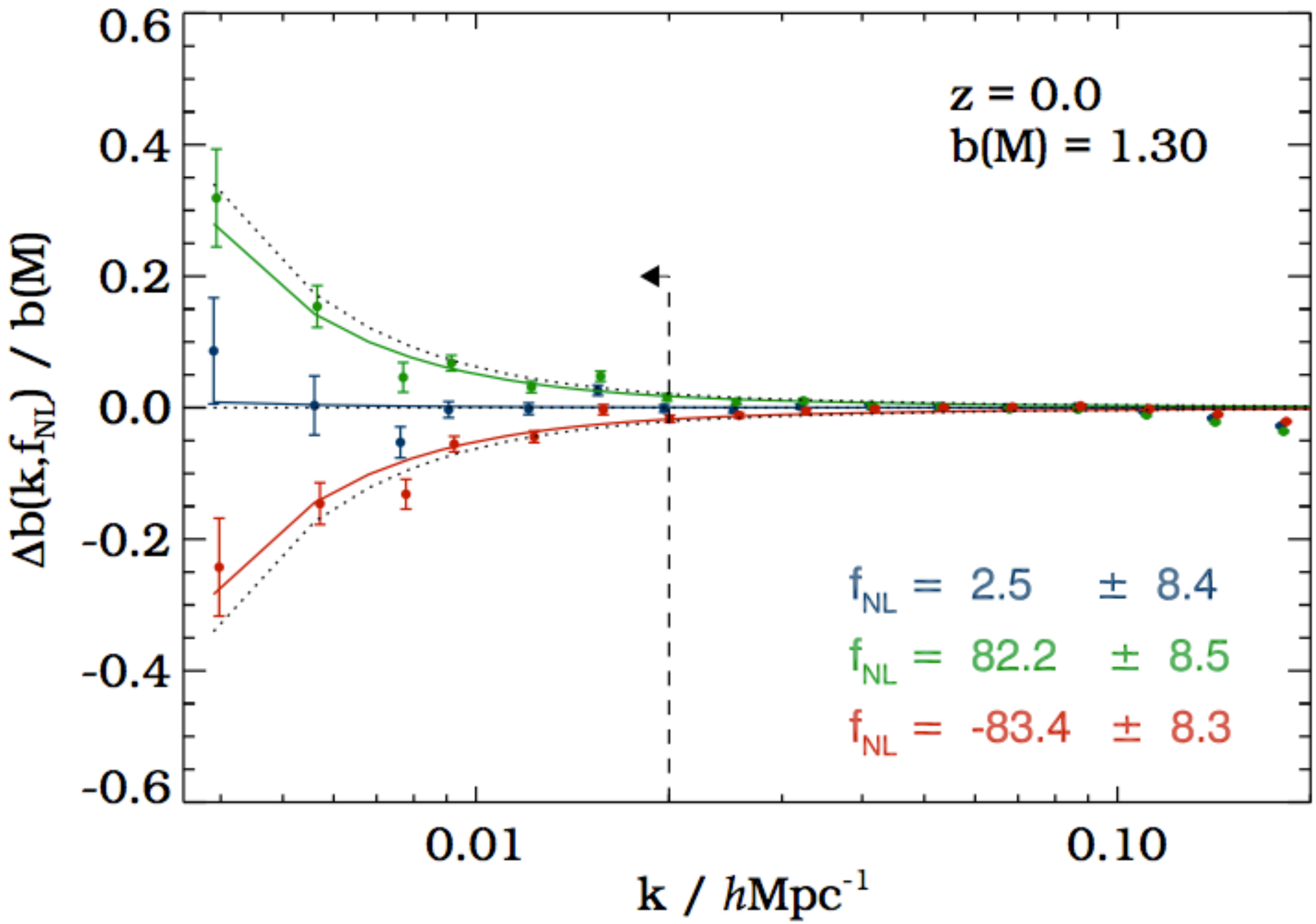

$$\delta_c - \delta_l - 2f_{NL}\phi_l = \delta_c - \delta_l\left(1 + \frac{2f_{NL}}{\alpha(k)}\right)$$

Number density of halos $\quad n \to n - \left(1 + \frac{2f_{NL}}{\alpha(k)}\right)\frac{dn}{d\delta_c}\delta_l$

Leads to a revised bias

$$b = 1 - \left(1 + \frac{2f_{NL}}{\alpha(k)}\right)\frac{d\ln n}{d\delta_c}$$

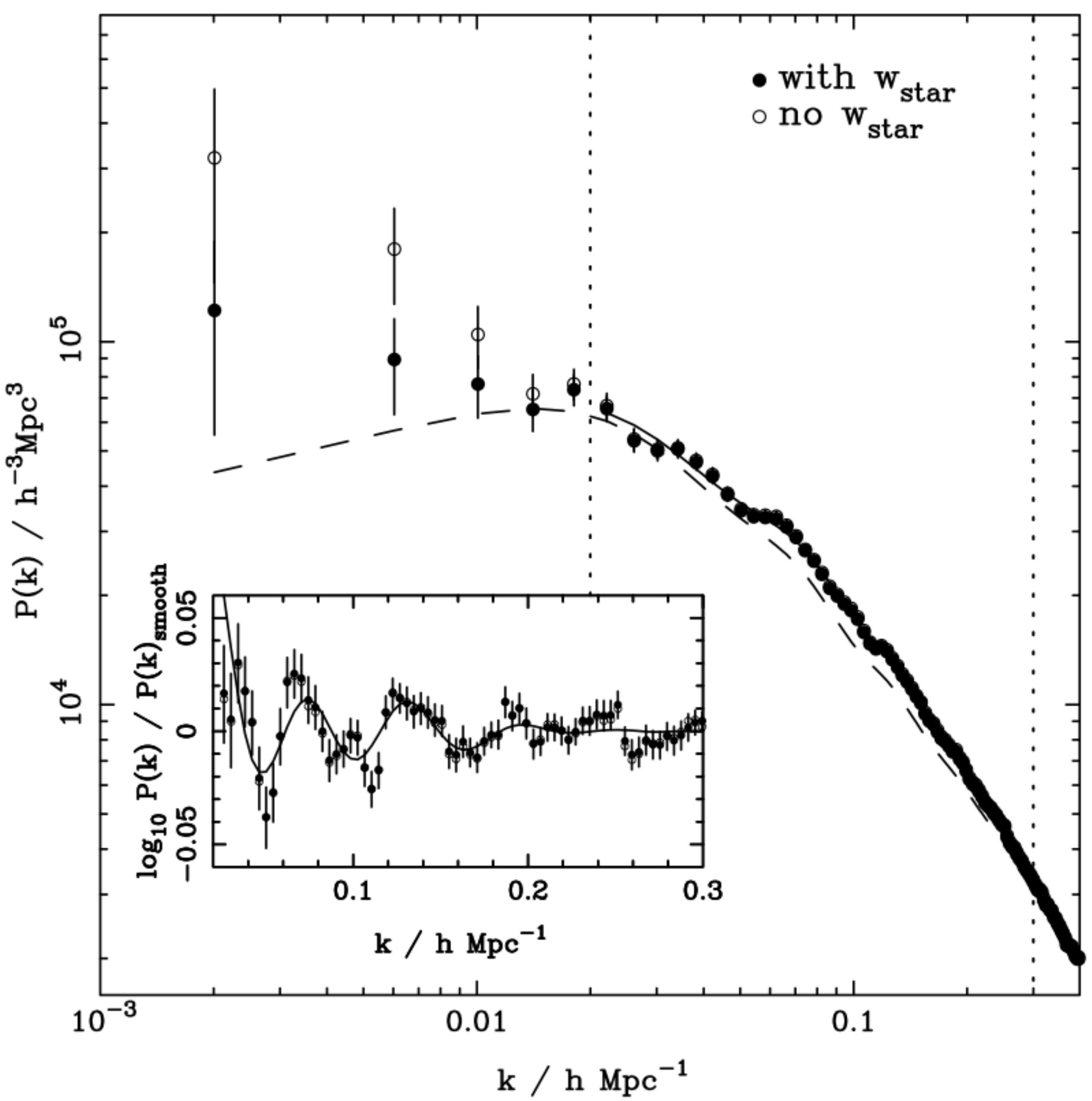
# $K^2$ dependence in simulations

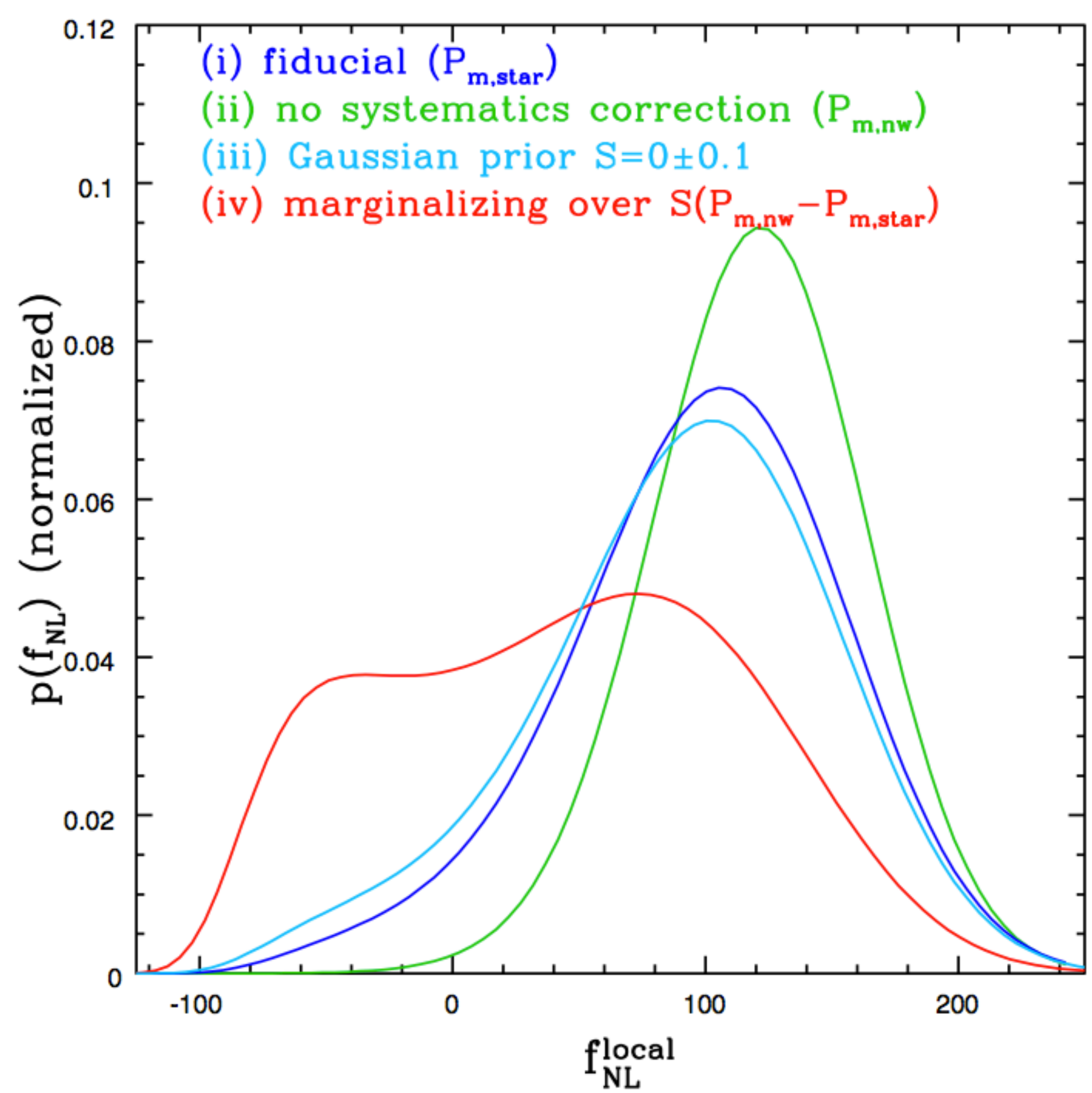



Dalal, Doré, Huterer, Shirokov 2007 ; Smith, LoVerde 2010; Smith, Ferraro, LoVerde 2011; Pillepich, Porciani, Hahn 2008; Desjacques, Seljak, Iliev 2008; Grossi et al 2009; Shandera, Dalal, Huterer 2010; Hamaus et al. 2011

# $f_{NL}$ from BOSS DR9 spectroscopic sample

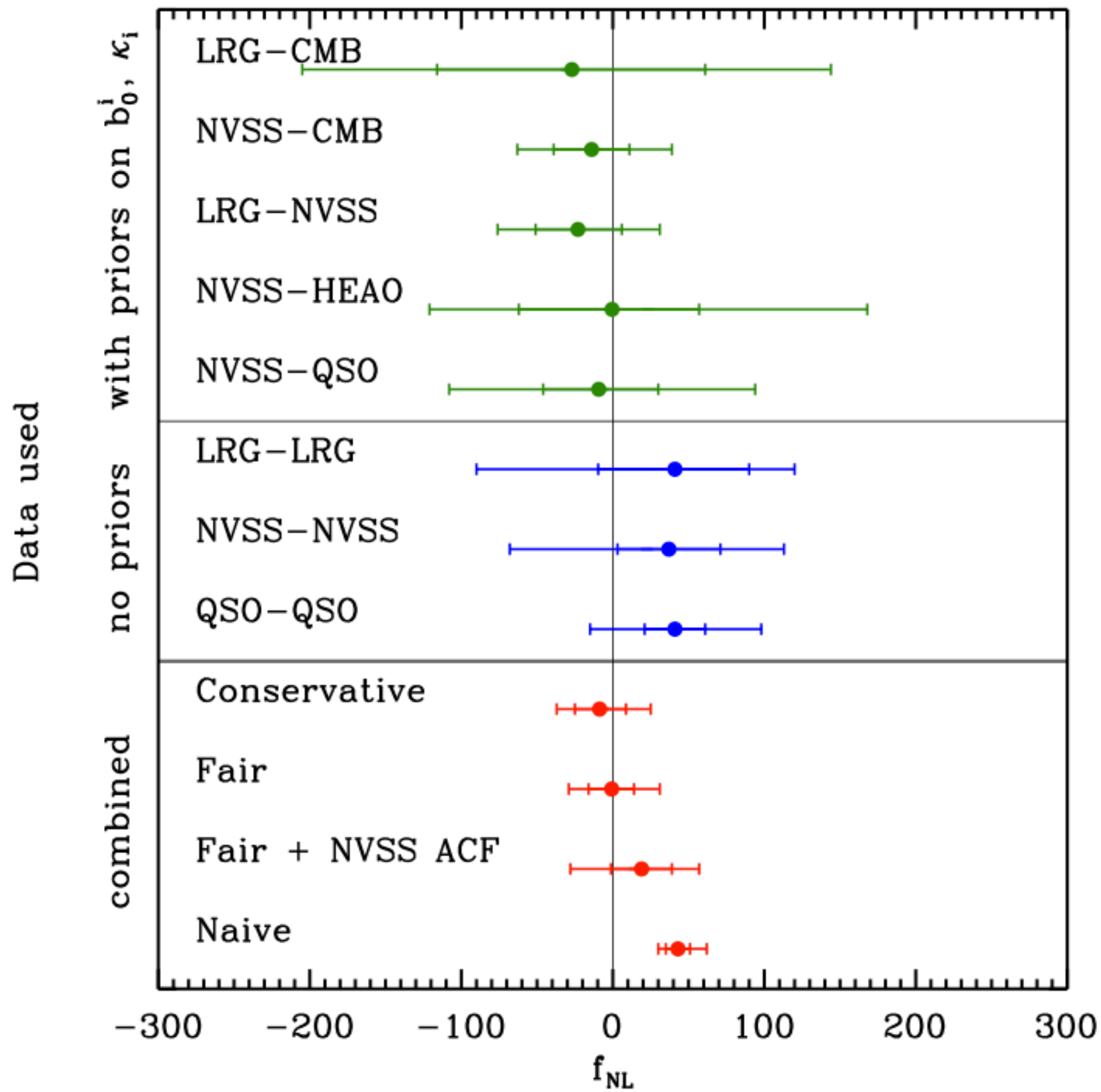
- linked careful analysis of observational systematics with local $f_{nl}$ measurement
- significant effect of stars on measured $f_{nl}$ signal
- $-82 < f_{nl} < 178$ (at 95% confidence)
- no significant evidence for non-zero $f_{nl}$




Ross et al. 2012: arXiv:1208.1491

# $f_{NL}$ from a combination of data




Giannantonio et al. 2013; arXiv:1303.1349

# Removing contaminants from δ : methods

Definition of measured modes of density field (think δ in Fourier space)

$$D(\mathbf{k}) = F(\mathbf{k}) - \epsilon f(\mathbf{k})$$

Cosmological fluctuation

Measurement

Contaminant

**Mode deprojection** removes contaminated modes from the covariance matrix of F

$$C_{\alpha\beta} \rightarrow \tilde{C}_{\alpha\beta} = C_{\alpha\beta} + \lim_{\sigma\rightarrow\infty} \sigma f(\mathbf{k}_\alpha) f^*(\mathbf{k}_\beta)$$

**Mode subtraction** removes the best-fit contaminant

$$D(\mathbf{k}) = F(\mathbf{k}) - \epsilon_{\mathrm{BF}} f(\mathbf{k})$$

# Estimating P(k)

Quadratic Maximum Likelihood (QML) estimator for P(k) is unbiased with any covariance matrix (so can cope with one with missing). Can consider QML as two steps:

mode deprojection (in **E**)

$$\mathbf{p}_j \equiv \sum_{\alpha,\beta} F^*(\boldsymbol{k}_\alpha)\mathbf{E}_{\alpha\beta}(k_j)F(\boldsymbol{k}_\beta)$$

$$\mathbf{E}(k_j) = -\frac{\partial \mathbf{C}^{-1}}{\partial P(k_j)}$$

debiasing / optimsation

$$\widehat{P}(k_i) = \sum_j \mathbf{N}_{ij}^{-1}\mathbf{p}_j$$

$$\mathbf{N}_{ij} = \mathrm{tr}\left\{\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial P(k_i)}\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial P(k_j)}\right\}$$

Practical problem performing QML: we need to manipulate matrices of size $(N_{mode})^2$

QML reduces to FKP estimator only in the limit of diagonal covariance, constant in bin

$$\widehat{P}(k_i) = \frac{1}{N_{\mathbb{k}_i}}\sum_{\boldsymbol{k}_\alpha \in \mathbb{k}_i} |F(\boldsymbol{k}_\alpha)|^2$$

QML: Tegmark (1997) astro-ph/9706198 FKP: Feldman, Kaiser & Peacock (1994) astro-ph/9304022

# Writing mode deprojection as a subtraction

$$\tilde{\mathbf{C}}^{-1}_{\alpha\beta} = \mathbf{C}^{-1}_{\alpha\beta} - \frac{\sum_{\mu\nu} \mathbf{C}^{-1}_{\alpha\mu} f(\boldsymbol{k}_\mu) f^*(\boldsymbol{k}_\nu) \mathbf{C}^{-1}_{\nu\beta}}{\sum_{\mu\nu} f^*(\boldsymbol{k}_\mu) \mathbf{C}^{-1}_{\mu\nu} f(\boldsymbol{k}_\nu)}$$

Sherman-Morrison matrix inversion lemma

Then the $\mathbf{p}_i$ in the QML estimator becomes
(for a diagonal C, although the result is actually more general)

$$\mathbf{p}_i = \sum_{\boldsymbol{k}_\alpha \in \mathbb{k}_i} \left\{ \frac{|F(\boldsymbol{k}_\alpha)|^2}{P^2(k_\alpha)} - \frac{2}{R_P} \mathrm{Re}\left[ S_P \frac{F^*(\boldsymbol{k}_\alpha) f(\boldsymbol{k}_\alpha)}{P^2(k_\alpha)} \right] \right.$$
$$\left. + \frac{|S_P|^2}{R_P^2} \frac{|f(\boldsymbol{k}_\alpha)|^2}{P^2(k_\alpha)} \right\}$$
$$= \sum_{\boldsymbol{k}_\alpha \in \mathbb{k}_i} \frac{\left| F(\boldsymbol{k}_\alpha) - \frac{S_P}{R_P} f(\boldsymbol{k}_\alpha) \right|^2}{P^2(k_\alpha)},$$

Remember that QML has

$$\widehat{P}(k_i) = \sum_j \mathbf{N}^{-1}_{ij} \mathbf{p}_j$$

$$R_P \equiv \sum_\mu \frac{|f(\boldsymbol{k}_\mu)|^2}{P(\boldsymbol{k}_\mu)} \qquad S_P \equiv \sum_{\boldsymbol{k}_\alpha} \frac{F^*(\boldsymbol{k}_\alpha) f(\boldsymbol{k}_\alpha)}{P_\alpha}$$

Kalus et al. 2016; arXiv:1607.02417

# Writing mode deprojection as a subtraction

Suppose instead we'd written out the likelihood, and found the best-fit contaminant

$$-2\ln\mathcal{L} = \ln\left(\prod_{\boldsymbol{k}} P(\boldsymbol{k})\right) + \sum_{\boldsymbol{k}} \frac{|F(\boldsymbol{k}) - \varepsilon f(\boldsymbol{k})|^2}{P(\boldsymbol{k})}$$

$$\frac{\partial \ln\mathcal{L}}{\partial\varepsilon} = \sum_{\boldsymbol{k}} \frac{\mathrm{Re}\left[F_f(\boldsymbol{k})F^*(\boldsymbol{k})\right] - \varepsilon|F_f(\boldsymbol{k})|^2}{P(\boldsymbol{k})}$$

Has solution

$$\varepsilon^{(\mathrm{BF})} = \frac{S_P}{R_P} \qquad R_P \equiv \sum_{\mu} \frac{|f(\boldsymbol{k}_\mu)|^2}{P(\boldsymbol{k}_\mu)} \qquad S_P \equiv \sum_{\boldsymbol{k}_\alpha} \frac{F^*(\boldsymbol{k}_\alpha) f(\boldsymbol{k}_\alpha)}{P_\alpha}$$

Giving

$$\widehat{P}(k_i) = \frac{1}{N_{\mathbb{k}_i}} \sum_{\boldsymbol{k}_\alpha} \left| F(\boldsymbol{k}_\alpha) - \frac{S_P}{R_P} f(\boldsymbol{k}_\alpha) \right|^2$$

This estimate is biased, but shows how linked to mode deprojection

Kalus et al. 2016; arXiv:1607.02417

# Can we debias FKP + mode subtraction?

Finding the expected values of the FKP estimator after mode subtraction

$$\left\langle \left| F_\alpha - \frac{S_P}{R_P} f_\alpha \right|^2 \right\rangle = \mathbf{C}_{\alpha\alpha} + |f_\alpha|^2 - 2|f_\alpha|^2 \left(1 + \frac{1}{R_P}\right) + \left(1 + \frac{1}{R_P}\right) |f_\alpha|^2$$
$$= \mathbf{C}_{\alpha\alpha} - \frac{|f_\alpha|^2}{R_P}.$$

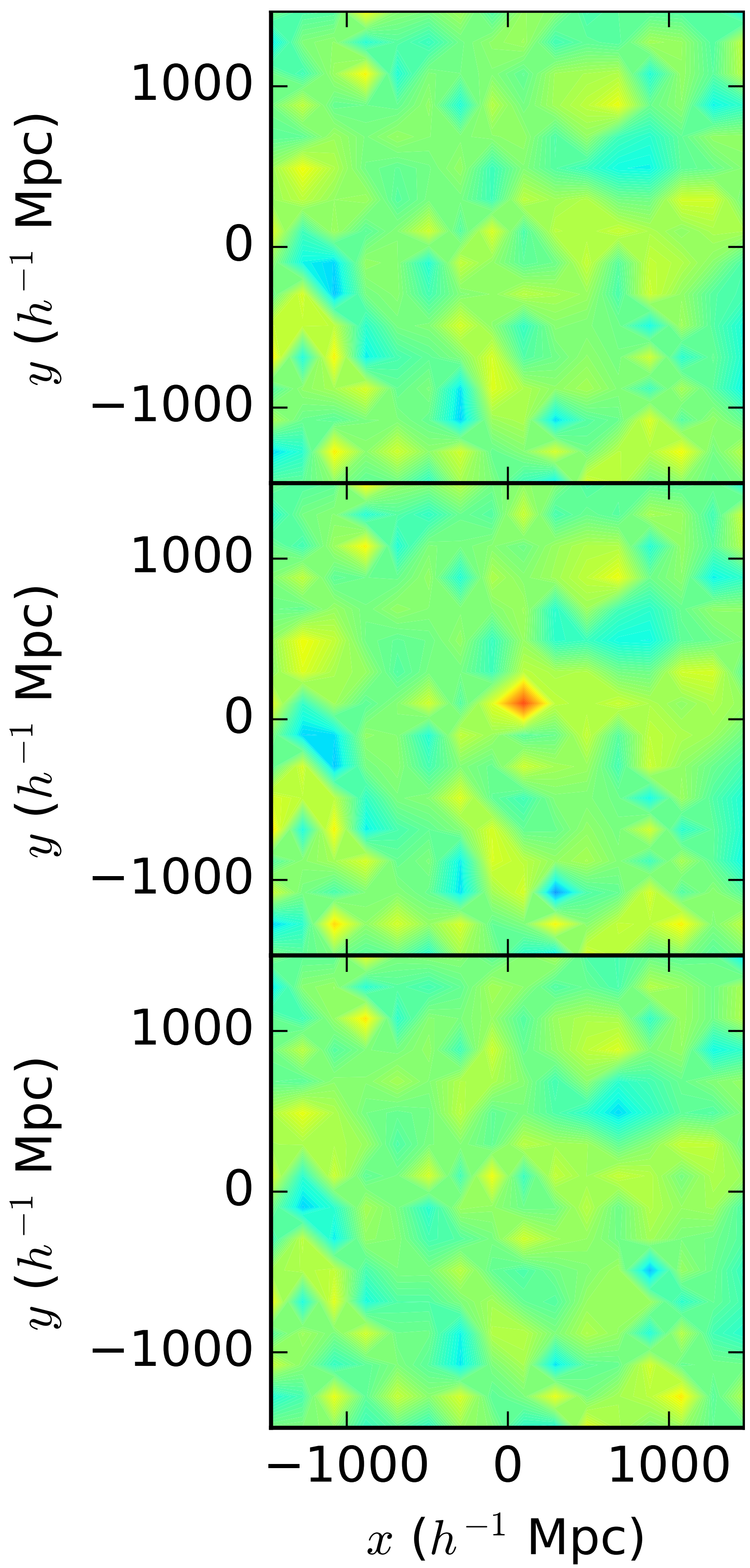
Gives a bias

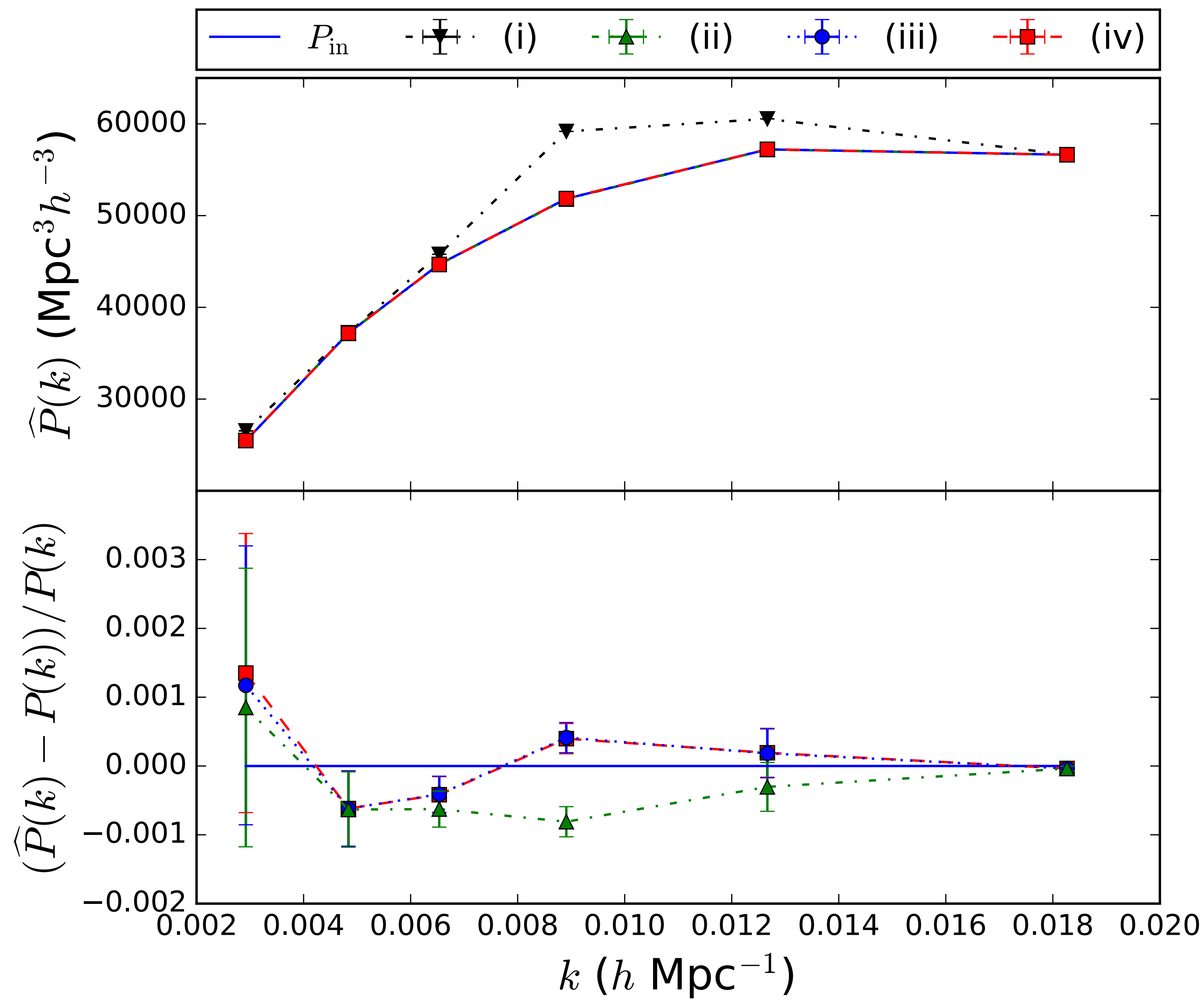
$$\left(1 - \frac{|f(\boldsymbol{k}_\alpha)|^2}{R_P P(k_\alpha)}\right)$$

That we can calculate (and remove) without having to define any $(N_{mode})^2$ matrices

Kalus et al. 2016; arXiv:1607.02417

# Some simple tests - spike removal

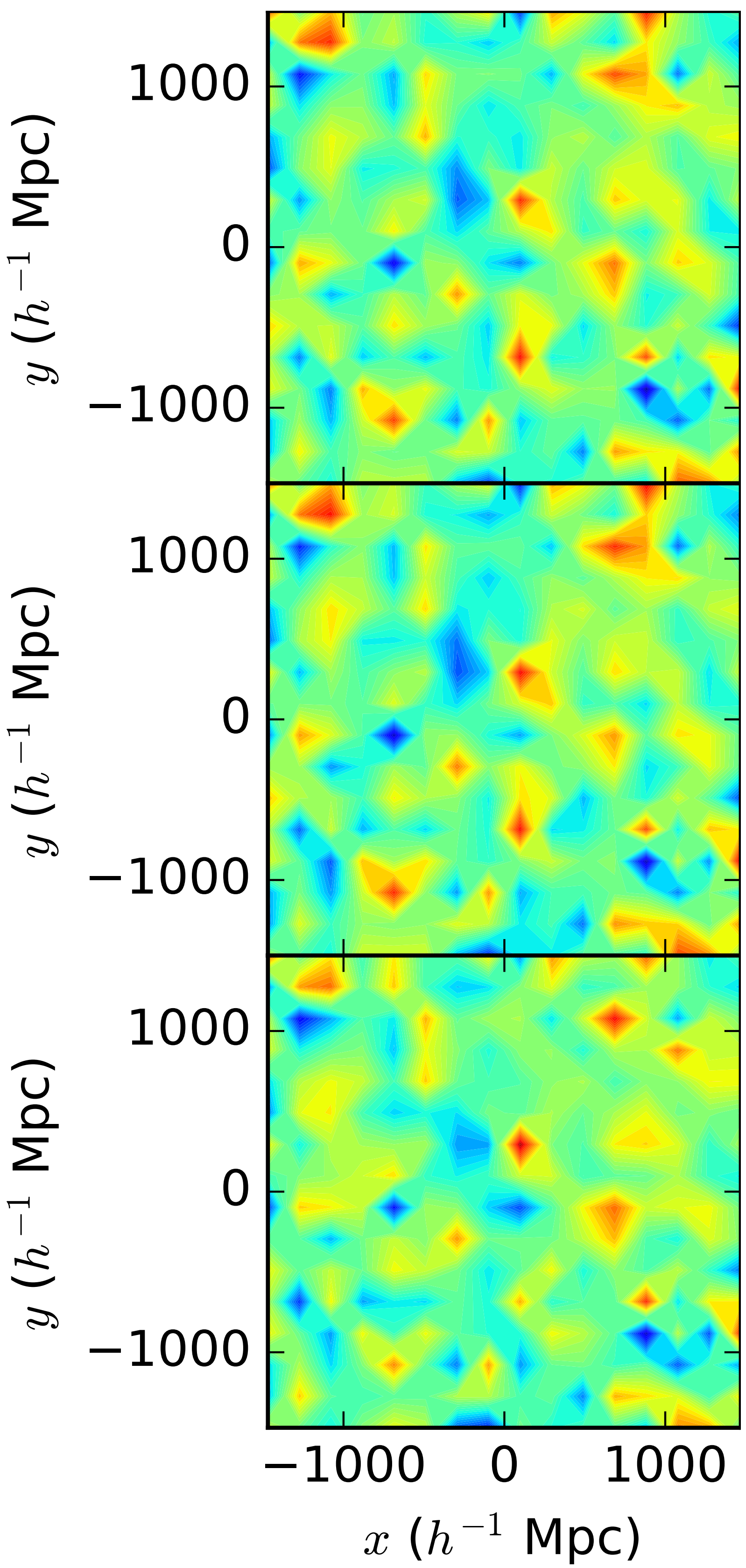


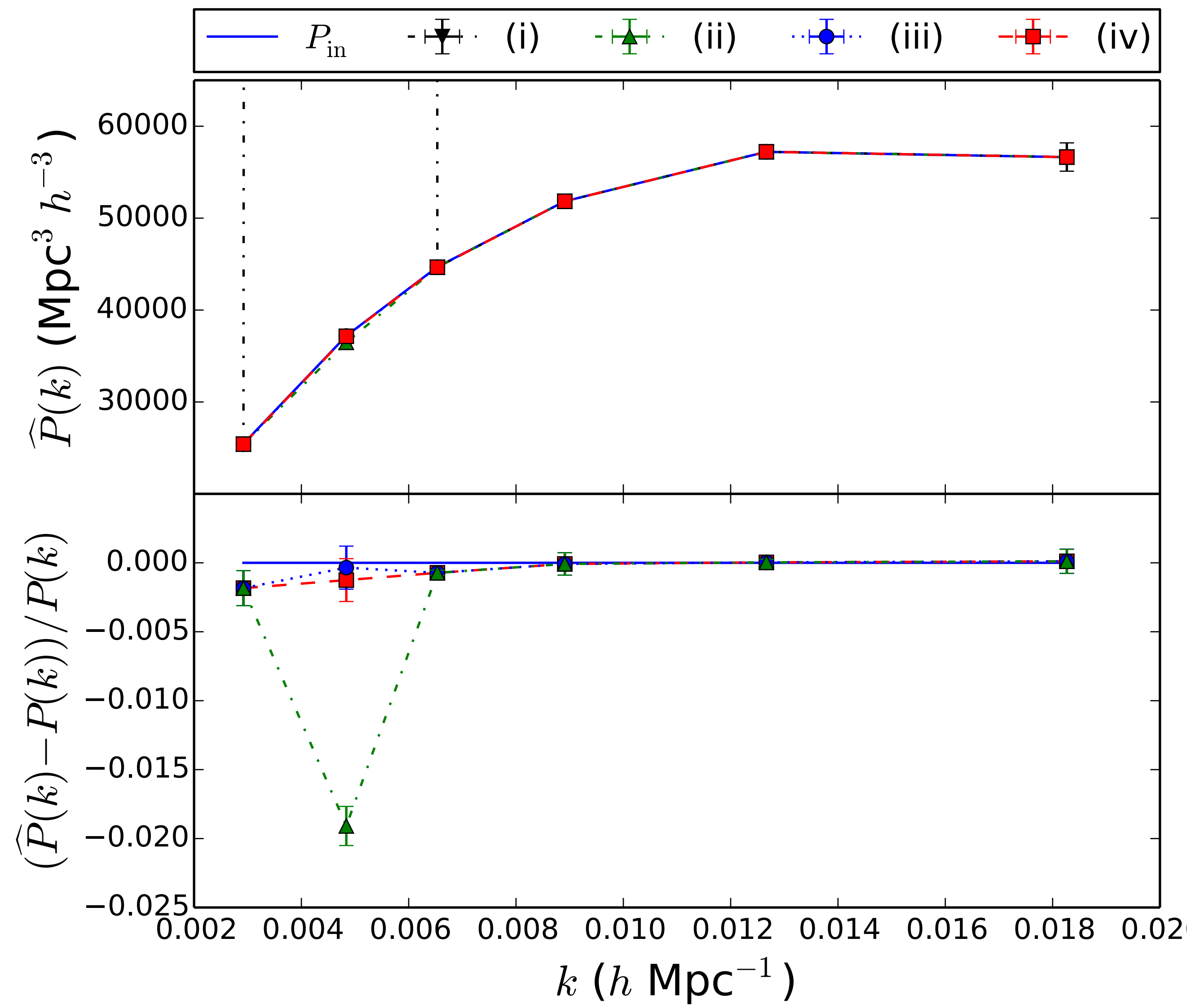




(i) do nothing about contaminant
(ii) best-fit removal of contaminant
(iii) debiased FKP
(iv) QML

Kalus et al. 2016; arXiv:1607.02417

# Some simple tests - contaminated modes




(i) do nothing about contaminant
(ii) best-fit removal of contaminant
(iii) debiased FKP
(iv) QML

Kalus et al. 2016; arXiv:1607.02417

# Conclusions

Finding a best-fit linear multiple, or nulling a mode are equivalent, modulo a bias for each measured mode

QML estimator does the correct debiasing as it's based on a covariance matrix allowing for the contaminant

FKP estimator can also be made unbiased, and has the advantage of not requiring large matrices to be calculated

Extension to multiple contaminants is straightforward (see arXiv: 1607.02417)

Sub-optimality of not using QML for contaminants is generally a lot lower order than ignoring window function

With a window function do need to consider large matrices to debias, but do not need to perform matrix inversion (see arXiv:1607.02417)