Optimal weak lensing data analysis

Uroš Seljak
UC Berkeley/LBNL
Berkeley, Jan 15 2019

Outline: goals of optimal data analysis

- We would like to optimally extract information from the data: if we have the full likelihood and the prior we can obtain the posterior. We have minimized Bayes risk and obtained optimal results. Often we use summary statistics as intermediate stage.
- The complication is that we only have an implicit likelihood as a function of many parameters, most of which we do not care about: we need to marginalize.
- Model evaluation can be very expensive (a full simulation)
- MCMC is often too expensive
- MAP/VI approximate and often wrong (inconsistent)
- We would like an analysis that is as good as MCMC, at a fraction of computational cost (as few likelihood evaluations as possible)
- Collaborators: G. Aslanyan, Y. Feng, B. Horowitz, C. Modi, B. Yu…

Linear case example: from implicit likelihood to power spectrum analysis

- We can write the probability distribution as a function of data **d** and modes **s,** where **d=Rs+n**: implicit likelihood

$$p(\boldsymbol{s},\boldsymbol{d}|\boldsymbol{S}) = (2\pi)^{-(N+M)/2} det(\boldsymbol{S}\boldsymbol{N})^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{s}^\dagger \boldsymbol{S}^{-1}\boldsymbol{s} + (\boldsymbol{d}-\boldsymbol{R}\boldsymbol{s})^\dagger \boldsymbol{N}^{-1}(\boldsymbol{d}-\boldsymbol{R}\boldsymbol{s})\right)$$

- By integrating over **s** (marginalizing) we can write the probability distribution of the data **d**: explicit likelihood

$$L(\boldsymbol{d}|\boldsymbol{\Theta}) = (2\pi)^{-N/2} \det(\boldsymbol{C})^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{d}^\dagger \boldsymbol{C}^{-1}\boldsymbol{d}\right)$$

- $C=RSR^T+N$
- We can rewrite this into an optimal quadratic estimator, which requires C^{-1}**d**

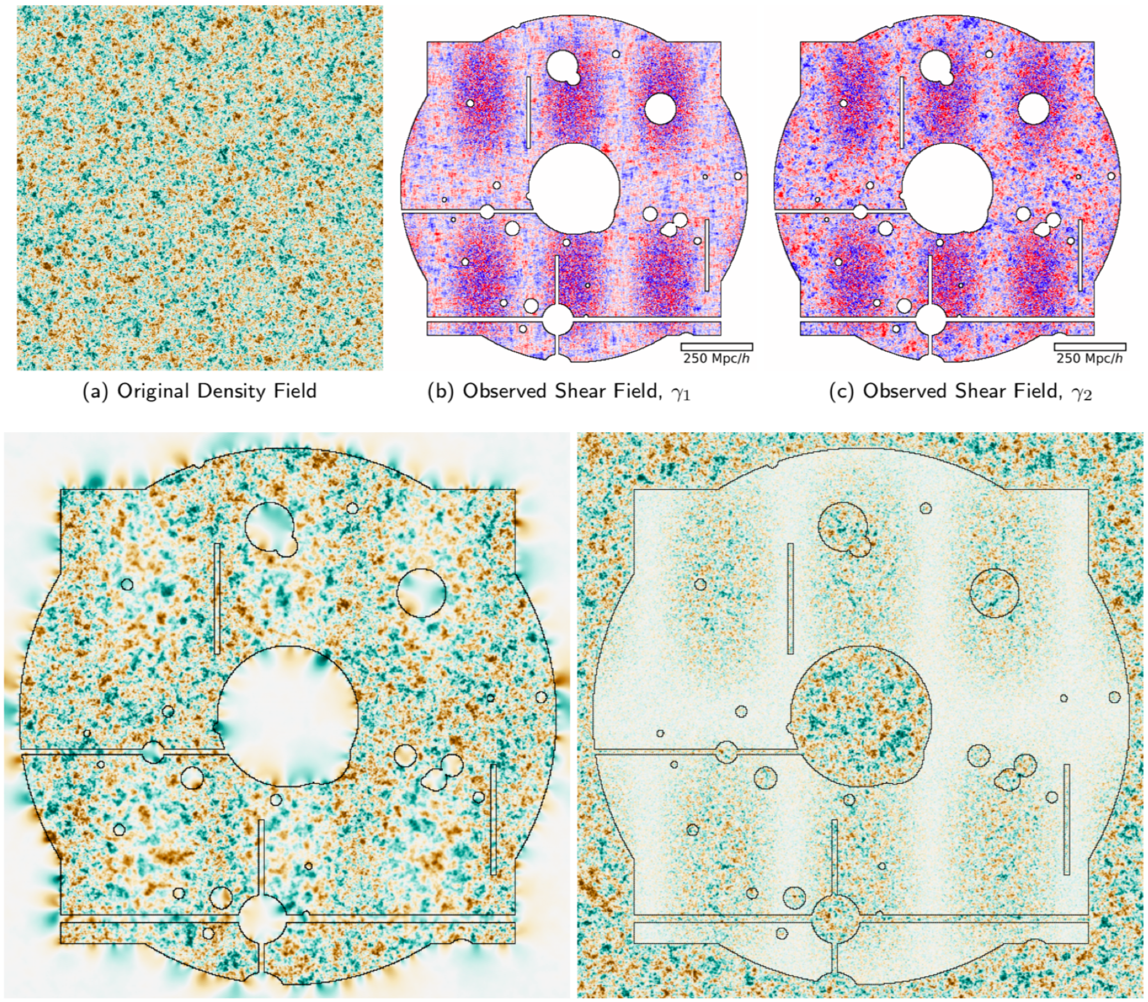
$$(\boldsymbol{F}\Theta)_l = \frac{\boldsymbol{F}}{2}\sum_{l'} F_{ll'}^{-1}(\boldsymbol{d}^\dagger \boldsymbol{C}^{-1}\boldsymbol{Q}_{l'}\boldsymbol{C}^{-1}\boldsymbol{d} - b_{l'})$$

- We can simplify by simpler weighting (pseudo-Cl=FKP)

Which is easier?

- To get just the power spectrum pseudo-Cl analysis is easiest, since there is no C^{-1} operation needed
- It is suboptimal on large scales due to the mask, but nearly optimal on small scales (hence used in CMB etc)
- This comes at a price: no obvious path to get the covariance matrix
- In practice it is modeled with simulations (mocks) or theory
- In contrast, if one sticks to the likelihood analysis one gets the covariance from the shape of the likelihood at the peak
- In explicit form this requires repeated $C^{-1}\mathbf{d}$: expensive
- In implicit form this requires finding the peak posterior of **s**: **Wiener filter**
- Sampling of the modes very expensive (Gibbs sampling), but has been attempted in CMB

Example: WL analysis Wiener filter

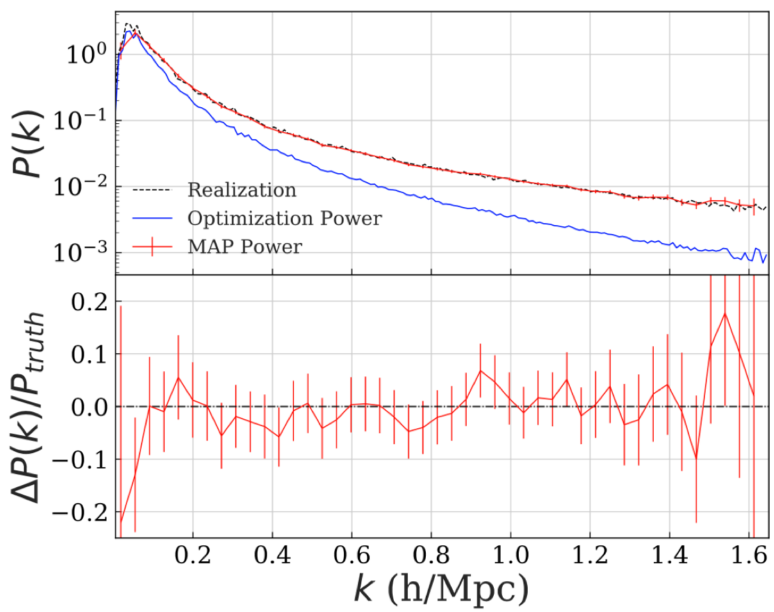


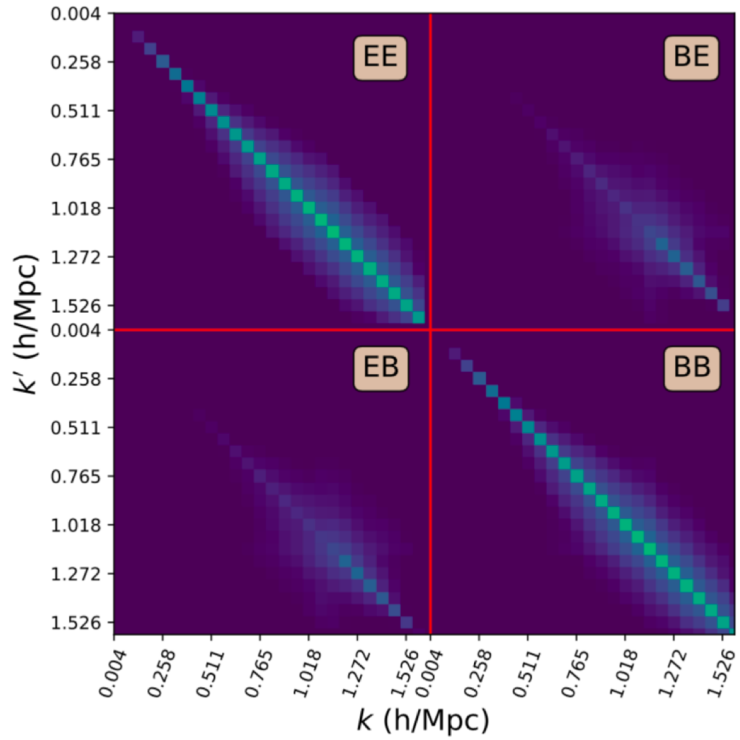
(a) Original Density Field (b) Observed Shear Field, γ_1 (c) Observed Shear Field, γ_2

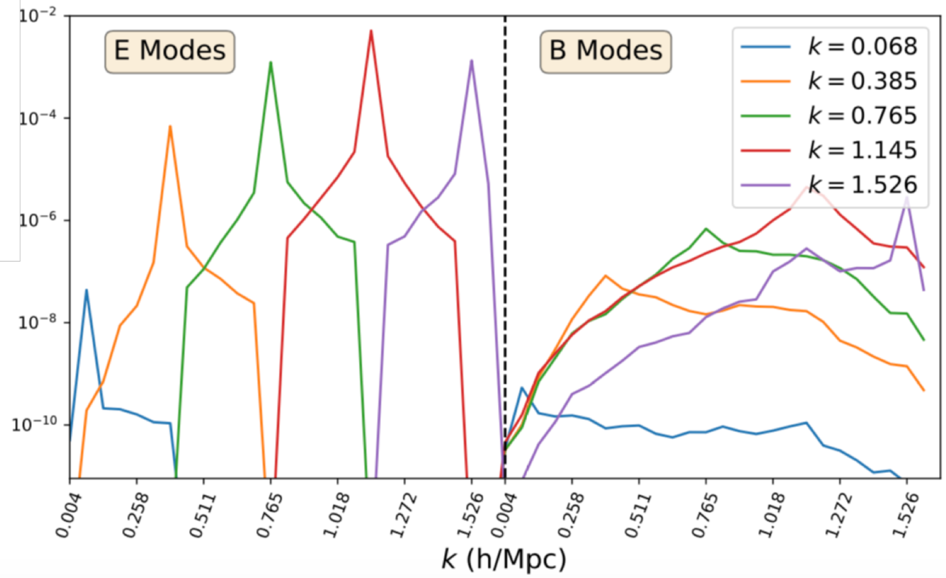
(d) Reconstructed Field (e) True Field - Reconstructed

Horowitz, US, Aslanyan 2018

Example: WL power spectrum analysis







What about the nonlinear case?

- We could follow "moment matching" path: evaluate all the N-point functions
- We would also need to get their covariance matrix. This is already very hard for 2-pt function, becomes impossible analytically for higher orders
- If one has N simulations then covariance matrix becomes singular with M>N summary statistics
- We can however try some specially powerful summary statistics (e.g. next talks)
- **Alternative: likelihood analysis**
- Writing down implicit likelihood is easy: d=f(s)+n

$$P(\boldsymbol{s}|\boldsymbol{d}) = (2\pi)^{-(M+N)/2} \det(\boldsymbol{S}\boldsymbol{N})^{-1/2} \exp\left(-\frac{1}{2}\left\{\boldsymbol{s}^{\dagger}\boldsymbol{S}^{-1}\boldsymbol{s} + [\boldsymbol{d} - \boldsymbol{f}(\boldsymbol{s})]^{\dagger}\boldsymbol{N}^{-1}[\boldsymbol{d} - \boldsymbol{f}(\boldsymbol{s})]\right\}\right).$$

- f(s) is a simulation of the data
- Need to first find peak posterior of s (MAP)

Finding MAP of s in 10^{10} dim parameter space

- Maximize posterior=minimize the loss function (d=x)

$$\chi^2(\boldsymbol{s}) = \boldsymbol{s}^\dagger \boldsymbol{S}^{-1}\boldsymbol{s} + [\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})]^\dagger \boldsymbol{N}^{-1}[\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})],$$

$$\chi^2(\boldsymbol{s}) = \chi_0^2 + 2\boldsymbol{g}\boldsymbol{\Delta s} + \boldsymbol{\Delta s}\boldsymbol{D}\boldsymbol{\Delta s},$$

$$\boldsymbol{g} = \frac{1}{2}\frac{\partial \chi^2}{\partial \boldsymbol{s}} = \frac{\boldsymbol{s}_m}{S} - \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}[\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s}_m)],$$

$$R_{ij} = \frac{\partial F(\boldsymbol{s}_m)_i}{\partial s_j}$$

gradient

$$\boldsymbol{D} = \frac{1}{2}\frac{\partial \chi^2}{\partial \boldsymbol{s}\partial \boldsymbol{s}} = \boldsymbol{S}^{-1} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1}\boldsymbol{R} + \boldsymbol{F}''[\boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s}_m)]$$

Hessian

$$\frac{\partial \chi^2(\boldsymbol{s})}{\partial \boldsymbol{\Delta s}} = 0,$$

$$\boldsymbol{\Delta s} = -\boldsymbol{D}^{-1}\boldsymbol{g}.$$

Newton's method

Need a gradient R_{ij}: derivative of a full simulated data wrt all initial modes s dotted with a vector: no large matrices needed
Also need nonlinear model F(s): a full simulation
Need to compute fast F(s) and its gradient
We can drop F"(d-F) in Gauss-Newton approximation (good when close to the minimum)
We are doing L-BFGS or Steihaug-CG
(Gauss Newton with trust region and conjugate gradient)

Nonlinear case: from implicit to explicit likelihood

- Integrate out the modes around the minimum variance map (approximate multivariate gaussian integrals)

$$L(\boldsymbol{d}|\boldsymbol{\Theta}) = \int P(\boldsymbol{s}, \boldsymbol{d} - \boldsymbol{F}(\boldsymbol{s})) d^M \boldsymbol{s}$$

$$= (2\pi)^{-(M+N)/2} \det(\boldsymbol{S})^{-1/2} \det(\boldsymbol{N})^{-1/2} \exp\left(\frac{1}{2}[\hat{\boldsymbol{s}}^\dagger \boldsymbol{D} \hat{\boldsymbol{s}} - \tilde{\boldsymbol{d}}^\dagger \boldsymbol{N}^{-1} \tilde{\boldsymbol{d}}]\right) \times$$

$$\int \exp\{-\frac{1}{2}[\boldsymbol{s} - \hat{\boldsymbol{s}}]^\dagger \boldsymbol{D} [\boldsymbol{s} - \hat{\boldsymbol{s}}]\} d^M \boldsymbol{s}$$

$$= (2\pi)^{-N/2} \det(\boldsymbol{S}\boldsymbol{N}\boldsymbol{D})^{-1/2} \exp\left(\frac{1}{2}[\hat{\boldsymbol{s}}^\dagger \boldsymbol{D} \hat{\boldsymbol{s}} - \tilde{\boldsymbol{d}}^\dagger \boldsymbol{N}^{-1} \tilde{\boldsymbol{d}}]\right).$$

- Hessian D in **s** basis: not sparse

$$\boldsymbol{D} = \frac{1}{2} \frac{\partial \chi^2}{\partial \boldsymbol{s} \partial \boldsymbol{s}} = \boldsymbol{S}^{-1} + \boldsymbol{R}^\dagger \boldsymbol{N}^{-1} \boldsymbol{R}$$

- This is **explicit likelihood**: no longer depends on **s**
- It maps data likelihood into a gaussian
- D determinant needed to preserve probability (i.e. Jacobian)

What just happened?

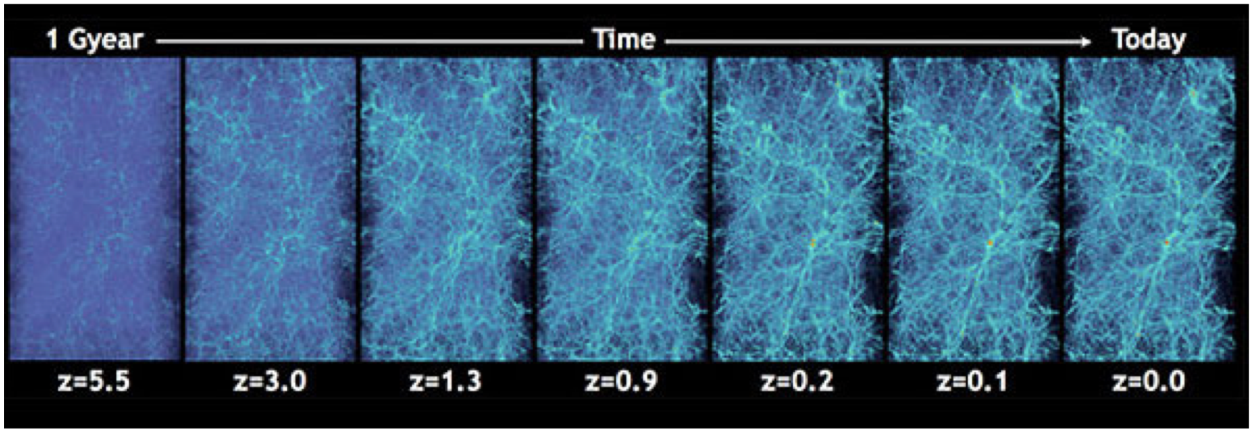
- Iterative solution to MAP has found a nonlinear mapping of the data to a gaussian distribution
- Likelihood analysis ensures optimal weighting of all the higher order statistics: this is the power of likelihood analysis
- If gravity creates non-linearity one can view this operation as reversing gravity
- All the higher order moments have been mapped back to the 2^{nd} moment (power spectrum)
- Summarizing information in the data is now easy, since it is a gaussian: everything is in power spectrum (and forward model parameters such as matter density)
- The only problem is that determinant: in high dimensions it is impossible to evaluate it

$$\nabla_{z}\mathcal{L}(\hat{s}) - \nabla_{z}\ln\det D$$

- We can determine 2^{nd} term using forward model simulations: we run MAP on the simulation and evaluate the above gradient
- Gradient has to vanish if we evaluate the gradient at the value of z used to generate the simulation

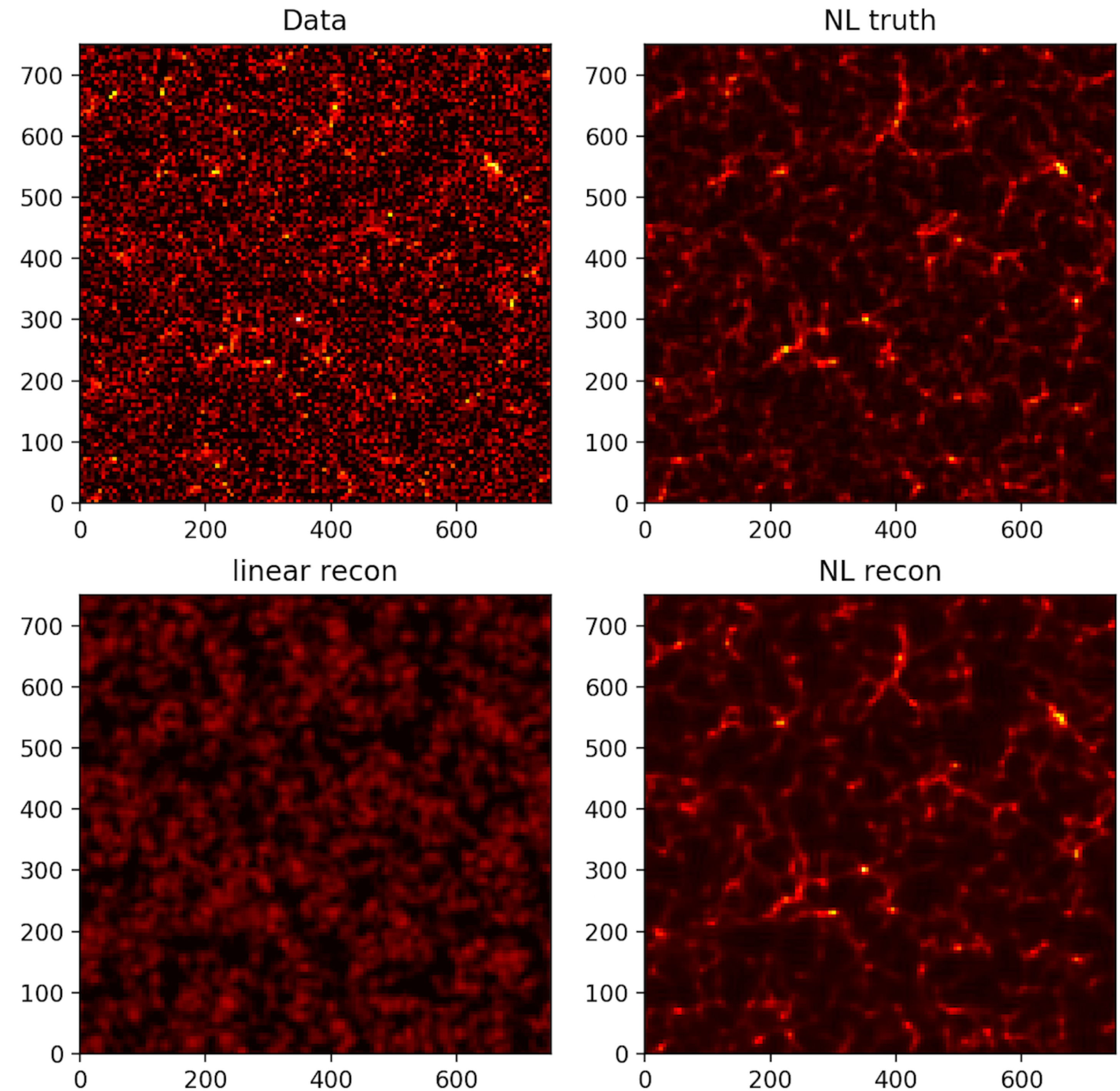
Applications to cosmology problems

- Forward model: FastPM (Yu etal) N-body simulation: we can do 10^{10} particles

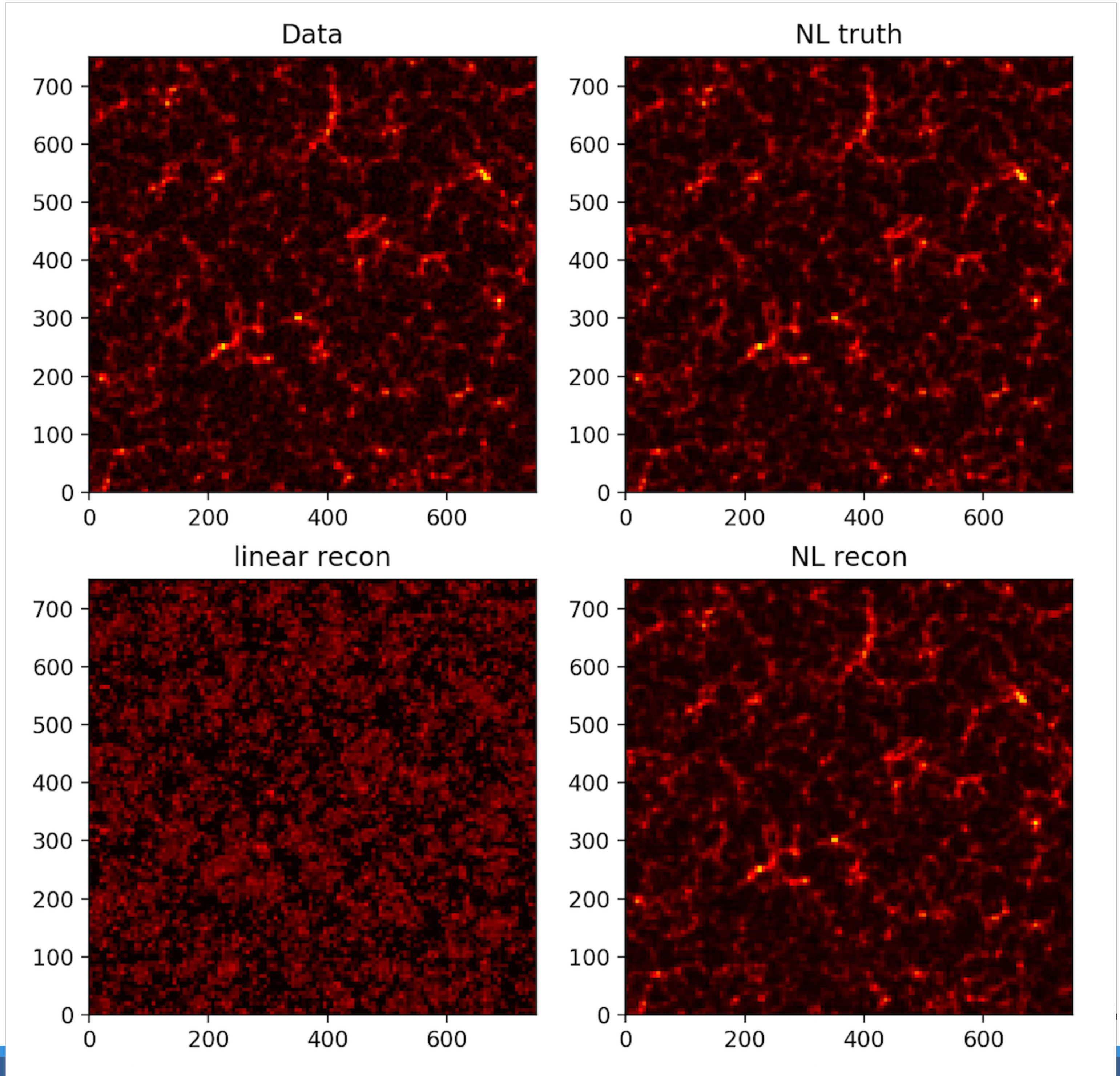


- We marginalize over these latent variables and determine the mean and covariance of summary statistics, which are their power as a function of scale (we use 30-40 bandpowers)

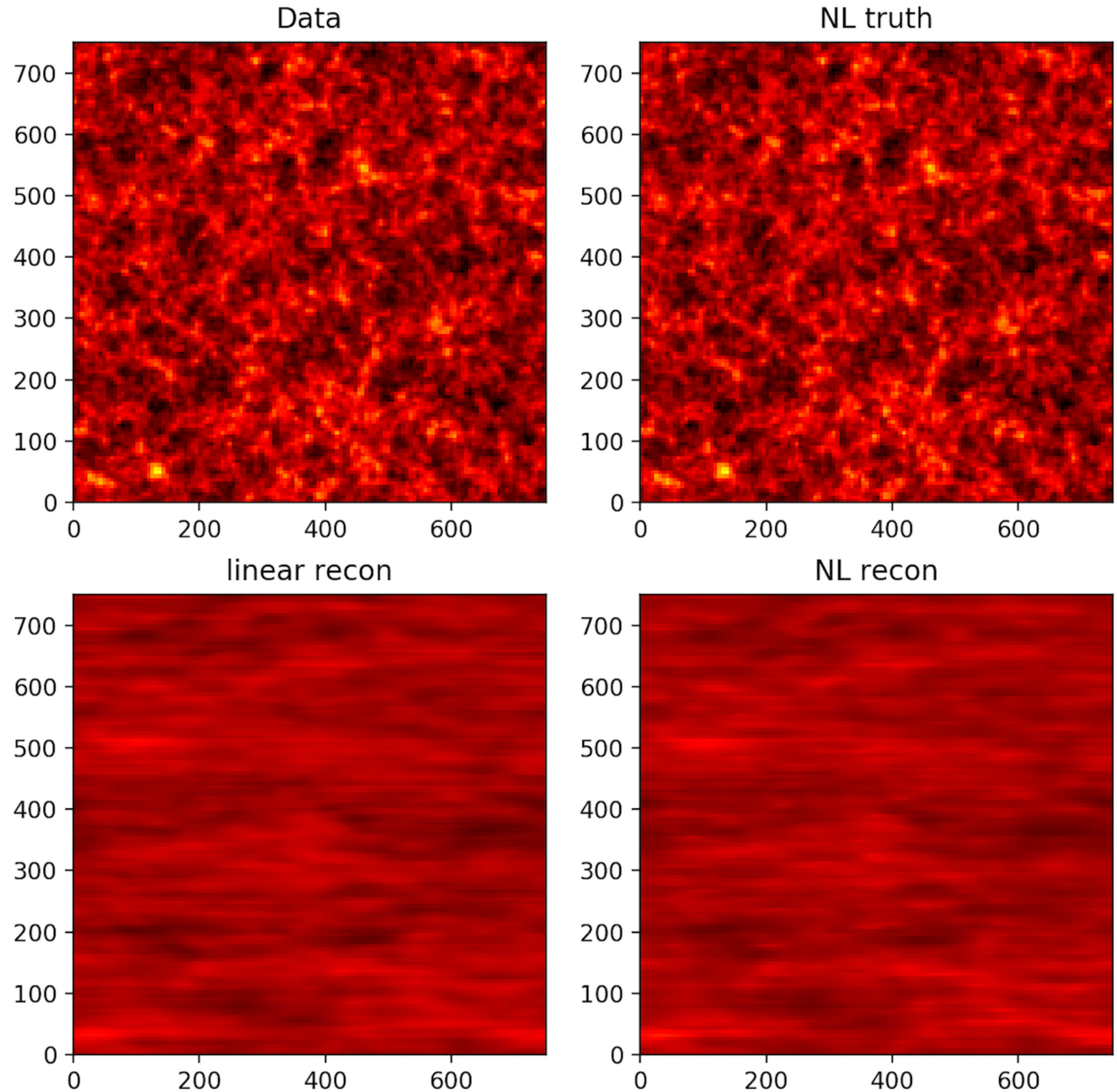
- high noise (P=1000Mpc/h^3), low smoothing
- 750Mpc/h box, 128^3
- High k suppressed
- Slices 6Mpc/h



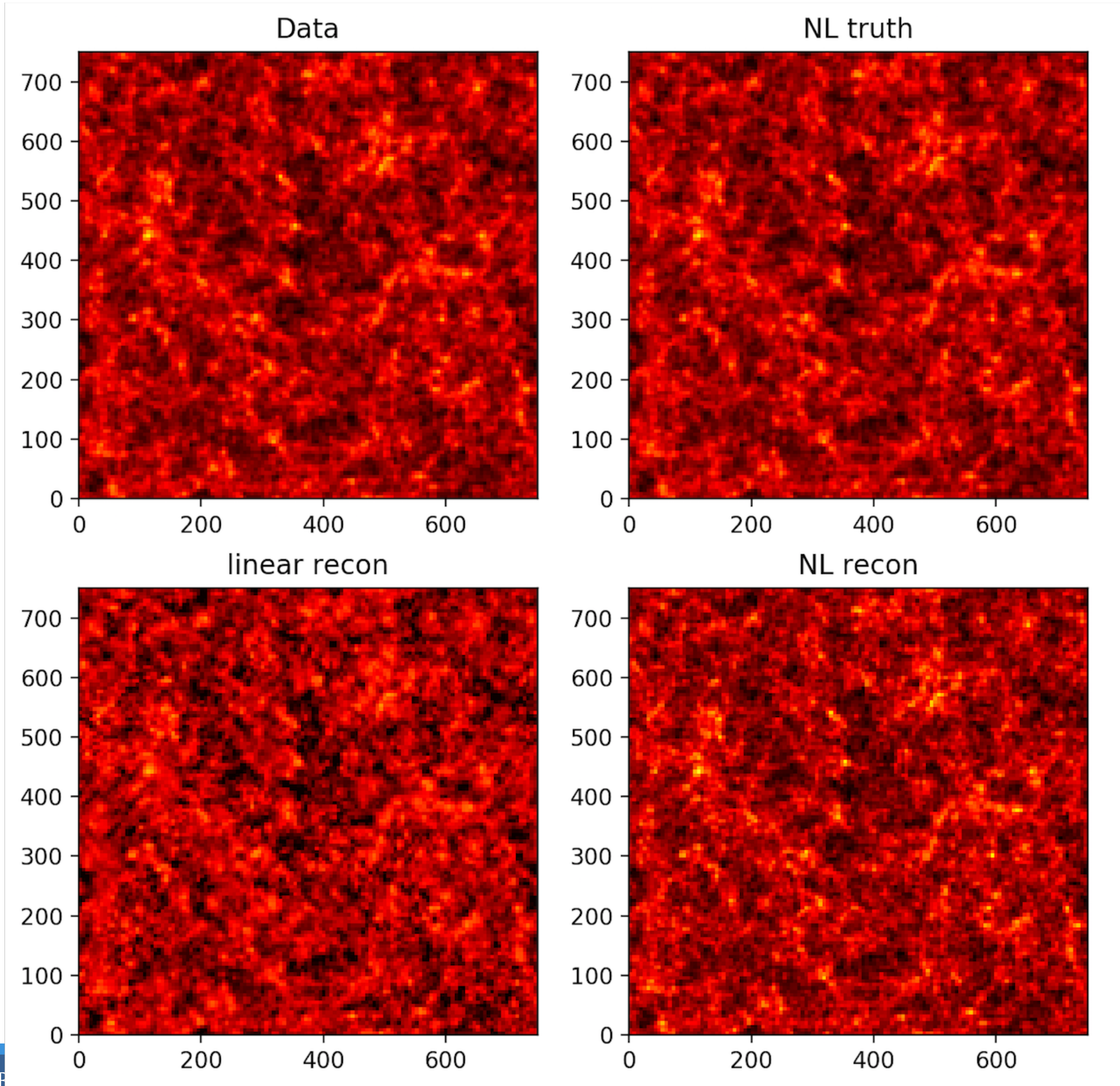
- Low noise, low smoothing
- 750Mpc/h box, 128^3
- Seems to reconstruct well all scales



- 2d projections (weak lensing)
- No reconstruction along line of sight, as expected

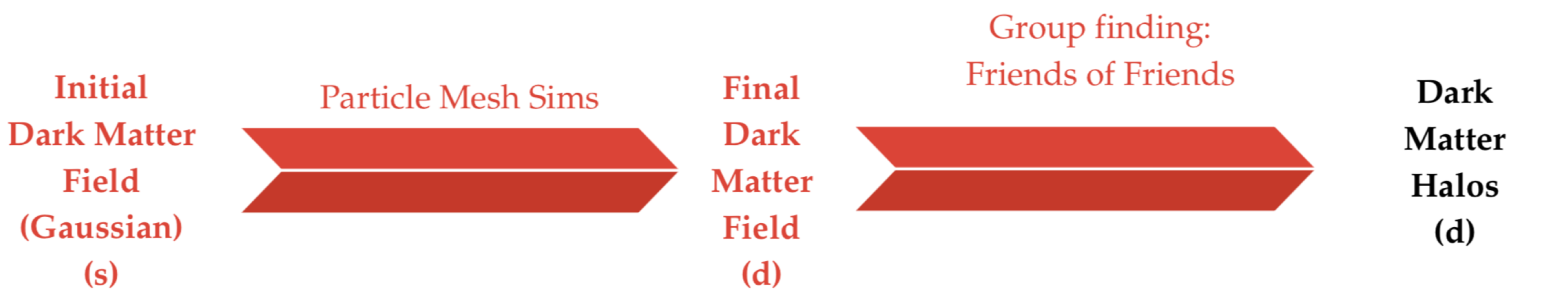


- 2d projections (weak lensing)
- Good reconstruction transverse to line of sight
- More gaussian because of wider projection



Forward model to galaxies: from initial to final dark matter to galaxies

Modi etal 2018



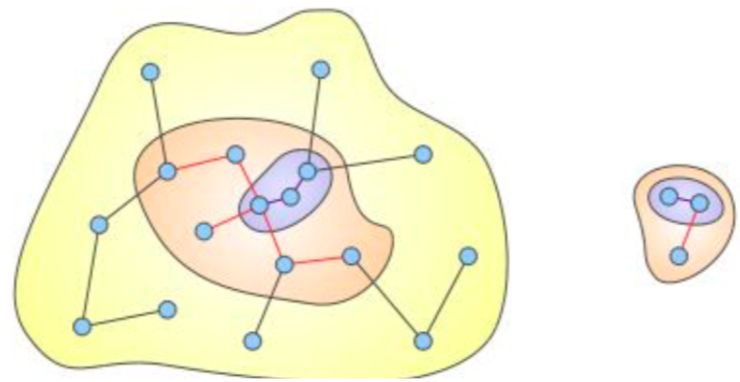
Leapfrog evolution

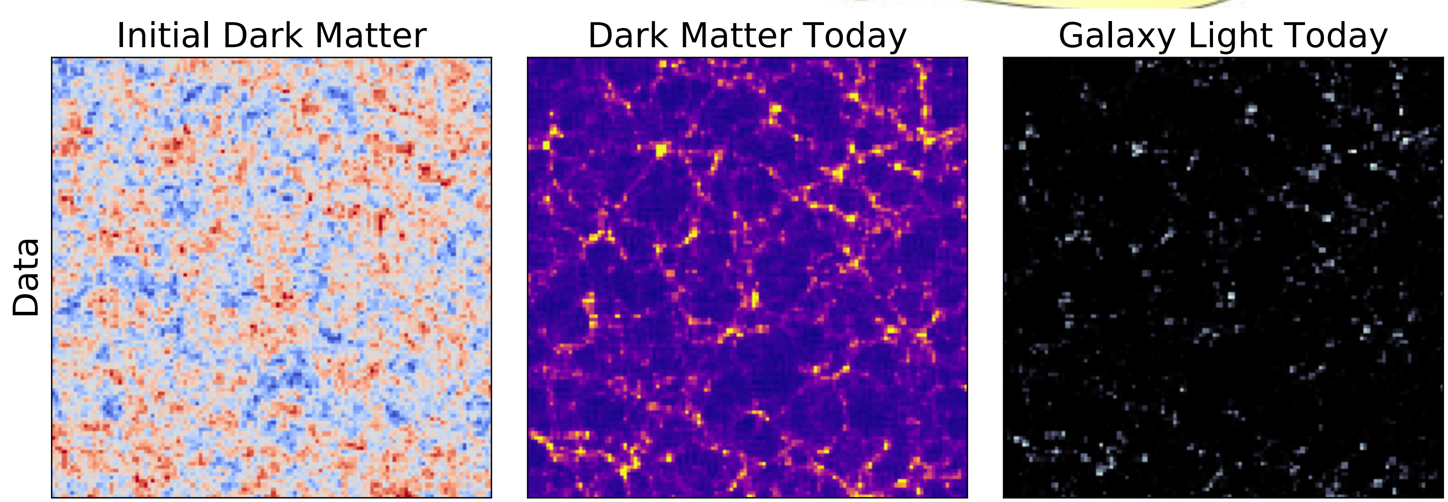
$$\mathbf{p}_{n+1/2} = \mathbf{p}_{n-1/2} - F(a_n)\nabla\phi_n\Delta a,$$
$$\mathbf{x}_{n+1} = \mathbf{x}_n + \frac{F(a_{n+1/2})}{a^2_{n+1/2}}\mathbf{p}_{n+1/2}\Delta a$$

Poisson Equation

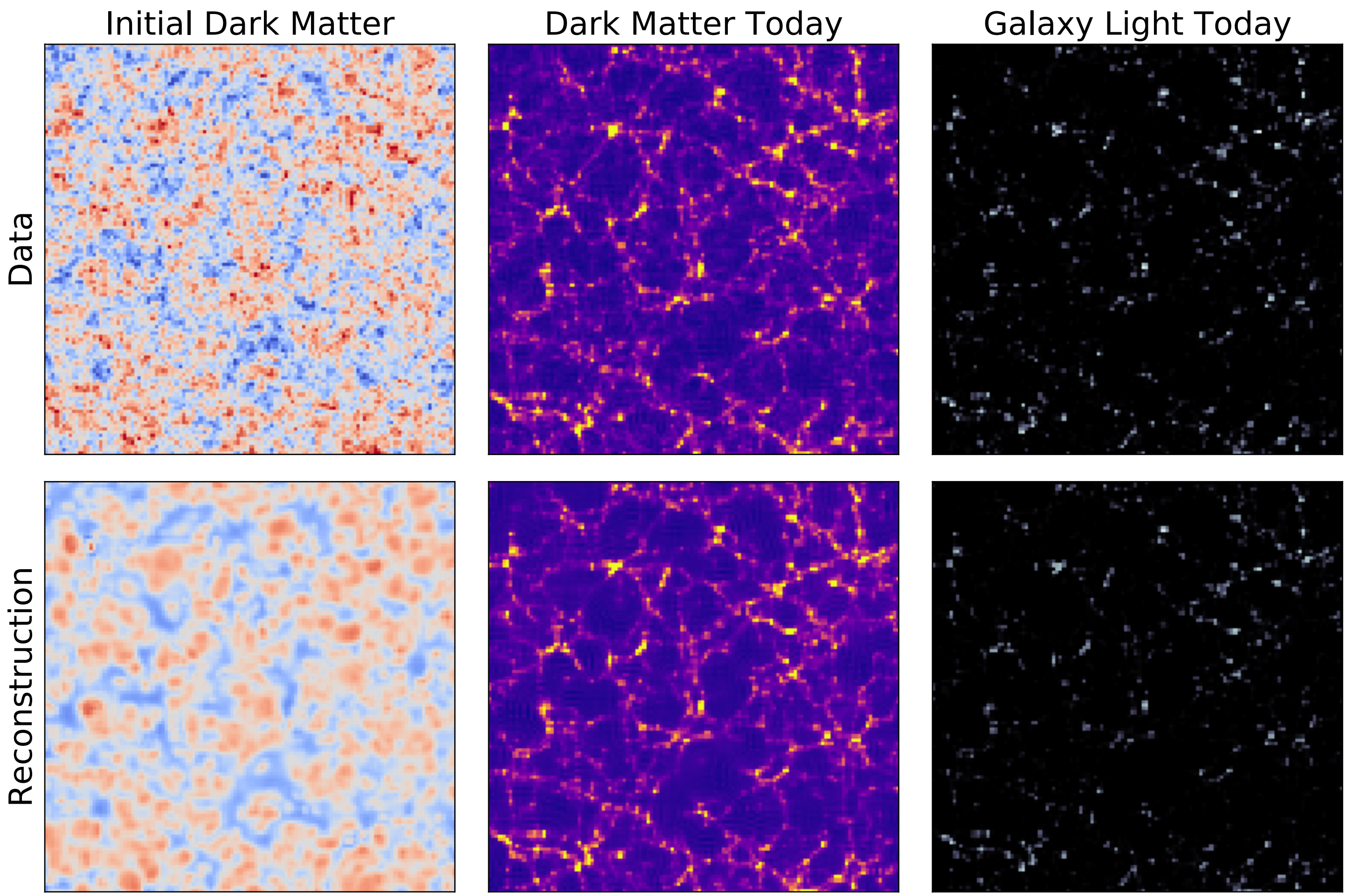
$$\tilde{\nabla}^2\tilde{\phi} = \frac{3}{2}\frac{\Omega_0}{a}(\tilde{\rho} - 1),$$

For N time steps





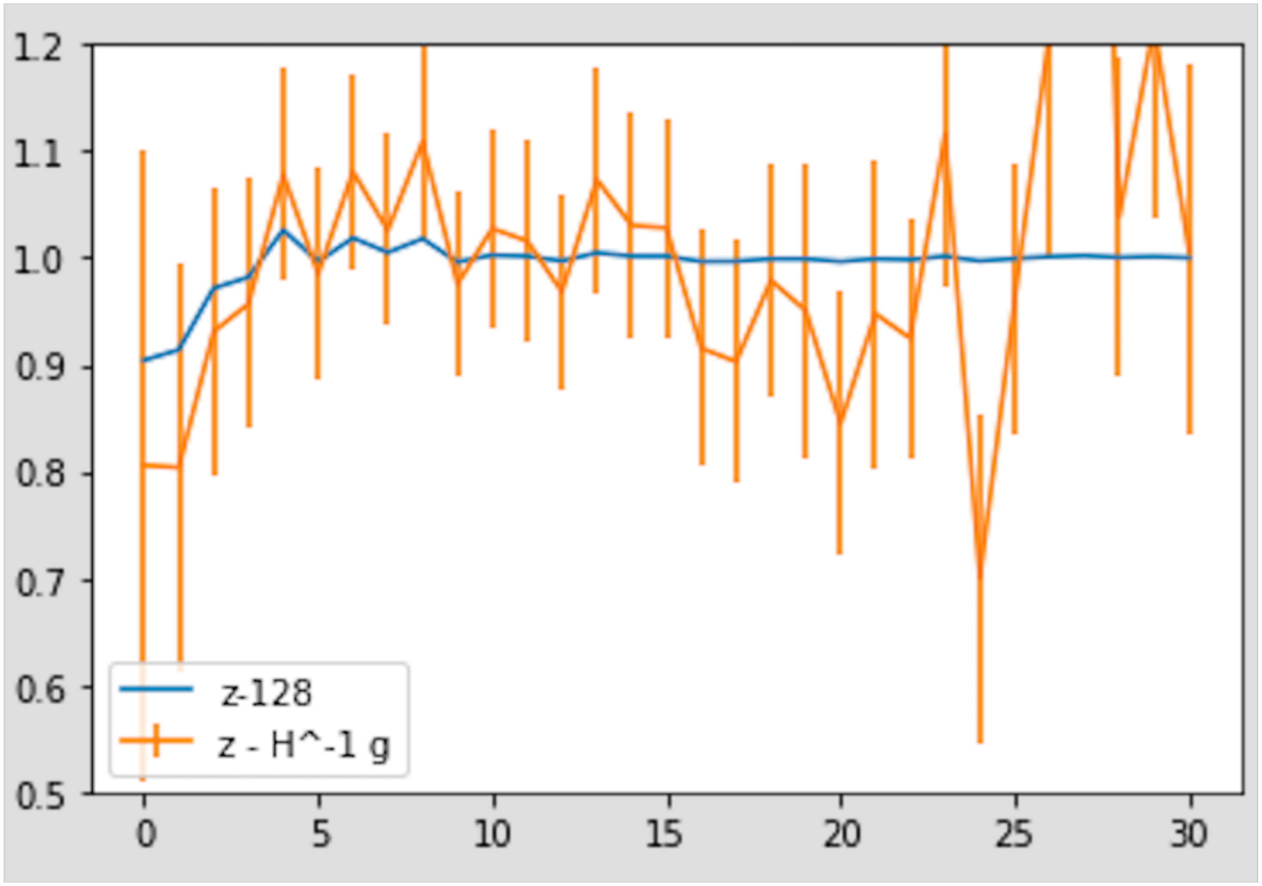
Example of MAP galaxy reconstruction



We use optimization that finds the best solution in terms of final data. This 3-d example optimizes in 2 million dimensions

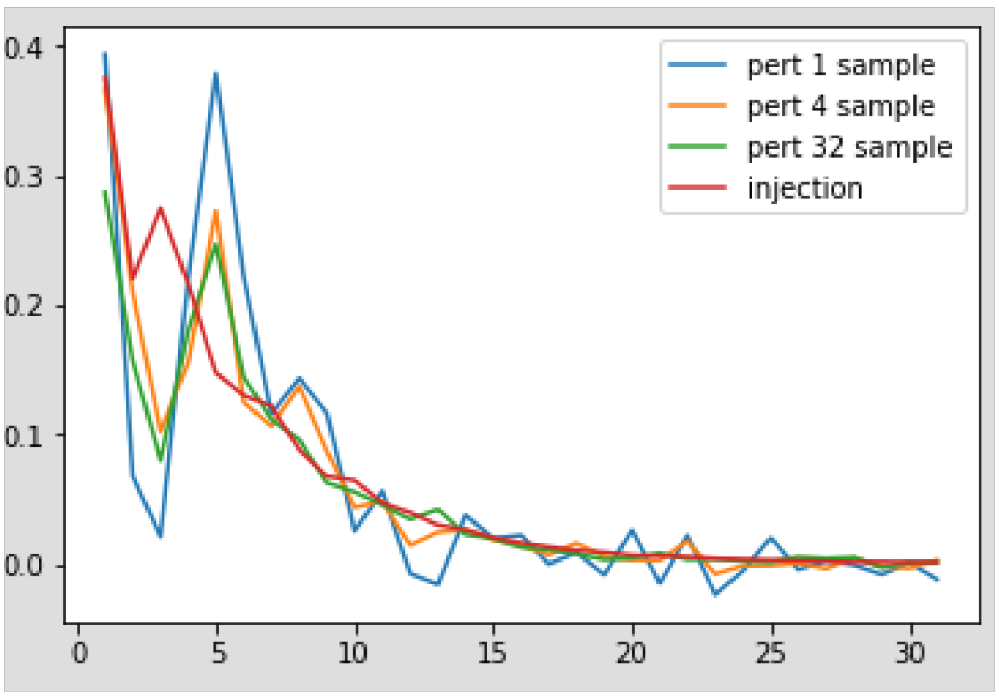
Reconstructing the power spectrum

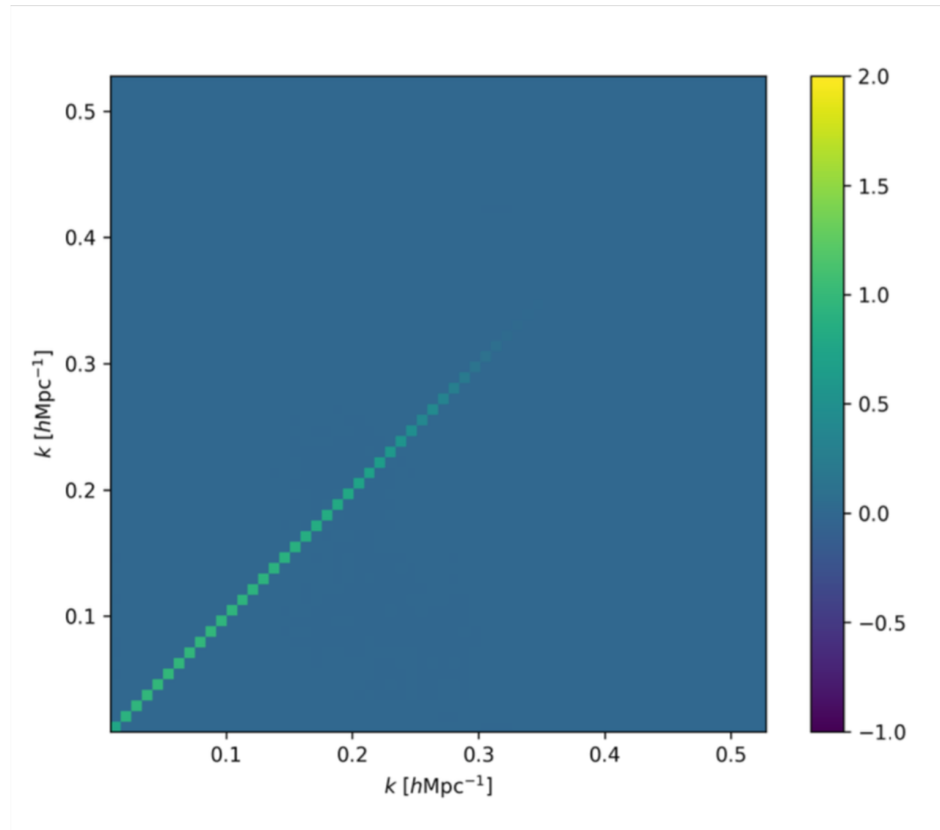
- Plot shows ratio of reconstructed power to true power
- We get unbiased results, with expected scatter



Hessian: inverse covariance matrix

- More samples reduce noise
- Finite difference with 2 simulations best (power injection method)
- Diagonal elements: we reconstruct more power at low k due to noise
- Off diagonals: almost zero
- Asymptotic limit: mean and covariance suffice for posterior





Beyond gaussian distribution

- Mapping a gaussian to an inverse Wishart distribution: if a bandpower has few modes then its distribution is not gaussian

$$-\ln L(\mathbf{\Theta}|\hat{\mathbf{\Theta}}) = \sum_l X_l \left(x_l - \ln x_l - 1\right)$$

$$X_l = \frac{((\boldsymbol{MF\Theta})_l + b_l)^2}{(\boldsymbol{MFM})_{ll}}$$

$$x_l = \frac{\hat{\Theta}_l + b_l}{\Theta_l + b_l},$$

- Nuisance parameters: baryonic effects (Biwei Dai talk), shear systematics etc.

Marginals and posteriors

- We have some summary statistics of the data with its covariance matrix: both can be model dependent
- The model depends on a number of parameters, which are all correlated with each other
- We only care about certain parameters: we marginalize over the others
- We are left with the posterior of the parameters we care: we would like to quantify this posterior in terms of its 1-d PDF and various summary statistics such as mean, mode, median, 68% and 95% credible intervals…
- Sometimes we also show 2-d PDF, but these are more for qualitative use (e.g. how correlated are two variables) than for any quantitative applications.
- We (almost) never look above 2-d (too difficult to visualize)

Astronomy: MCMC dominated

- We are confusing Bayesian marginal analysis with MCMC analysis, at a great CPU cost
- If the posterior is gaussian there is nothing wrong with a MLE or MAP analysis
- Can we develop a method that starts at MAP and expands around it to go to non-gaussian posteriors only if needed
- Can we unchain posterior inference?

Stochastic VI: KL divergence is noisy when sampling

- We have data **x** and parameter z. Assume gaussian q(z) and also assume p(z|**x**) is gaussian in z, but we do not know it.

$$\mathcal{L}_q = -\ln q(z),\ q(z) = N(z;\mu,\Sigma) \qquad \mathcal{L}_p = -\ln p(\boldsymbol{z}|\boldsymbol{x})$$

$$\mathrm{KL}(q||p) = \langle \mathcal{L}_q - \mathcal{L}_p \rangle_q,\ \mathrm{KL}(p||q) = \langle \mathcal{L}_p - \mathcal{L}_q \rangle_p$$

- We cannot analytically evaluate KL, so we have to sample from p or q. KL(p||q) leads to MC sampling:

$$\mathrm{KL}(p||q) = \sum_k \mathcal{L}_p(z_k) - \frac{(z_k-\mu)^2}{2\Sigma} - \frac{\ln(2\pi\Sigma)}{2}$$

Let us minimize $\mathrm{KL}(p||q)$ with respect to μ and Σ.

$$\mu = N_k^{-1}\sum_k z_k,\ \Sigma = N_k^{-1}\sum_k (z_k-\mu)^2$$

- Converges as $N_k^{-1/2}$, the usual MC scaling.

Our proposal: EL_2O (on arxiv today)

With Byeonghee Yu

$$\mathcal{L}_q = -\ln q(z),\ q(z) = N(z;\mu,\Sigma)$$

$$\mathcal{L}_p = -\ln p(\boldsymbol{z}|\boldsymbol{x})$$

- We propose to minimize L_2 norm between L_p and L_q. It needs to be sampled from some fiducial probability distr, which can be q
- if q covers p it is noiseless, if not it finds the closest solution to it
- **EL_2O: expectation with L_2 optimization** $\mathrm{EL_2O} = \langle (\mathcal{L}_q - \mathcal{L}_p - c)^2 \rangle_{\tilde{p}}$
- For the problem above $\mathrm{EL_2O} = N_k^{-1} \sum_k \left[\frac{(z_k-\mu)^2}{2\Sigma} - \mathcal{L}_p(z_k) - c' \right]^2$

$$c' = c - (\ln 2\pi\Sigma)/2$$

- This is quadratic in z, linear least square (and thus convex) in

$$-c' + \mu^2/2\Sigma,\ -\mu/\Sigma \text{ and } 1/\Sigma$$

- N_k=3 samples suffice to give complete solution. **No sampling noise** 24

EL_2O with gradient, Hessian…

- Modern trend in ML/stats: automatic derivatives (backpropagation or adjoints): huge gains in information

$$\mathcal{L}_p(\boldsymbol{z}_k + \Delta \boldsymbol{z}_k) = \sum_{n=0}^{\infty} \frac{1}{n!} \nabla^n_{\boldsymbol{z}} \mathcal{L}_p(\boldsymbol{z}_k)(\Delta \boldsymbol{z}_k)^n$$

$$\mathcal{L}_q(\boldsymbol{z}_k + \Delta \boldsymbol{z}) = -\ln q(\boldsymbol{z}_k + \Delta \boldsymbol{z}) = \sum_{n=0}^{\infty} \frac{1}{n!} \nabla^n_{\boldsymbol{z}} \mathcal{L}_q(\boldsymbol{z}_k)(\Delta \boldsymbol{z})^n$$

$$\mathrm{EL_2O} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\langle N_{\mathrm{der}}^{-1} \sum_{n=0}^{n_{\max}} \sum_{i_1,..i_n} \left[\nabla^n_{\boldsymbol{z}} \mathcal{L}_p(\boldsymbol{z}) - \nabla^n_{\boldsymbol{z}} \mathcal{L}_q(\boldsymbol{z}, \boldsymbol{\theta})\right]^2 \right\rangle_{\tilde{p}}$$

$$q(\boldsymbol{z}) = N(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-N/2} \det \boldsymbol{\Sigma}^{-1/2} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})},$$

$$\mathcal{L}_q = \frac{1}{2}\left[\ln\det\boldsymbol{\Sigma} + (\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) + N\ln(2\pi)\right].$$

For $n = 2$ we optimize

$$\mathrm{EL_2O} = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}}{\operatorname{argmin}} N_{\mathrm{der}}^{-1} \left\langle \sum_{i,j\le i}^{M} \left\{\nabla_{z_i}\nabla_{z_j}\mathcal{L}_p - \nabla_{z_i}\nabla_{z_j}\mathcal{L}_q\right\}^2 + \sum_{i=1}^{M} \left\{\nabla_{z_i}\mathcal{L}_p - \nabla_{z_i}\mathcal{L}_q\right\}^2 \right\rangle_{\tilde{p}}$$

Beyond full rank gaussian: bijective transformations

Full rank Gaussian is the only correlated distribution that we know how to analytically marginalize: compute Hessian, invert to get covariance, remove the unwanted variables, invert again to get Hessian of remaining parameters. We can enhance it using 1d transforms which allow easy marginals

bijective 1d transform

$$q(\boldsymbol{z}) = N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Pi_i |J_i|, \;\; J_i = \frac{dy_i}{dz_i},$$

- the resulting posterior can accommodate more of the variation of $\mathcal{L}_p$ – corresponding to skewness and curtosis in 1d $u_i = (z_i - \mu_i)/\Sigma_{ii}^{1/2}$

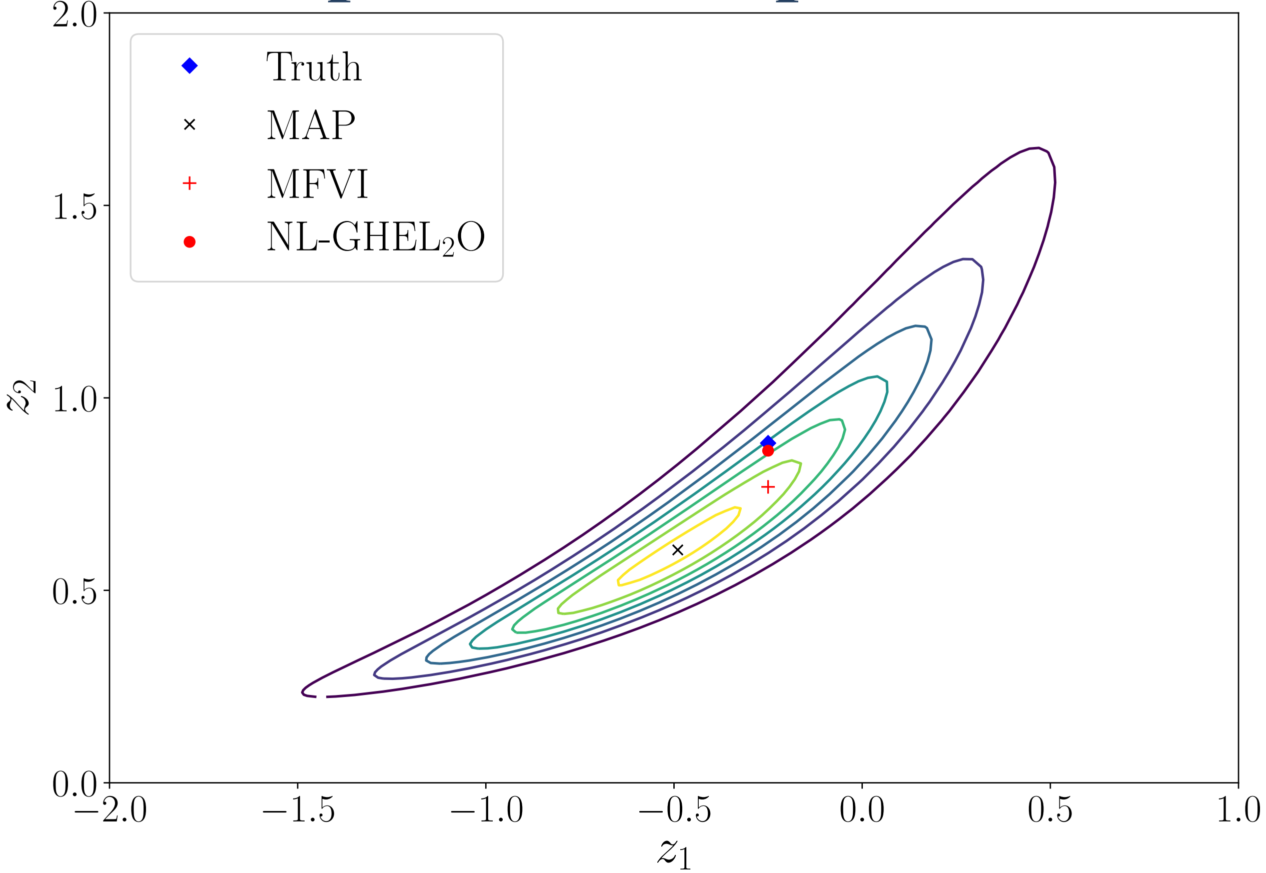
$$y_i(z_i) = \sinh_\eta \left[\frac{\exp(\epsilon_i u_i) - 1}{\epsilon_i} \right]$$

$$y_i(z_i) = \sinh_\eta u_i \quad \text{for } \epsilon_i = 0$$

$$\sinh_\eta(x) = \begin{cases} \eta^{-1}\sinh(\eta x) & (\eta > 0) \\ x & (\eta = 0) \\ \eta^{-1}\text{arsinh}(\eta x) & (\eta < 0) \end{cases}$$

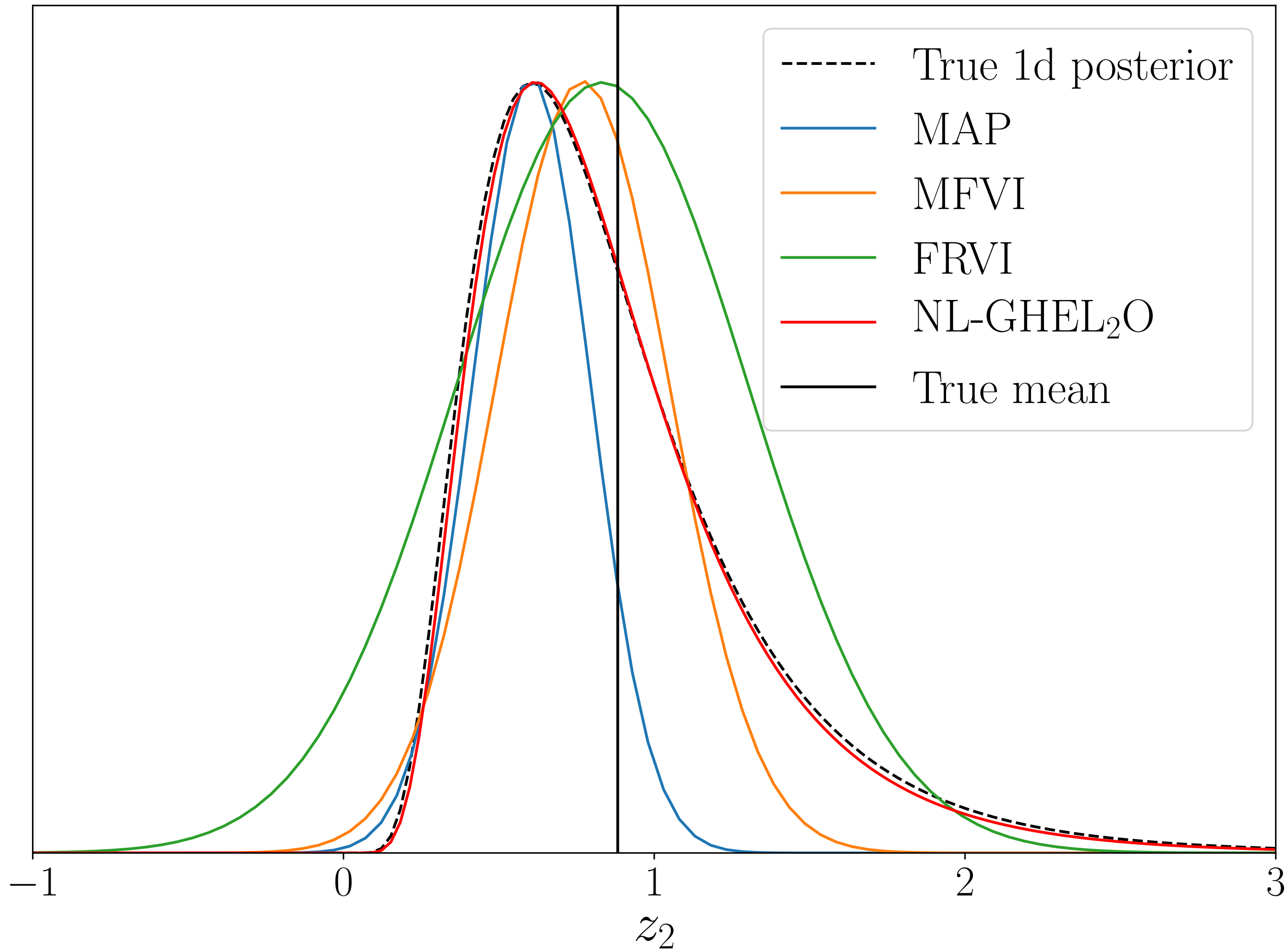
- Model the posterior using $\boldsymbol{\varepsilon}$, $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$
- Can apply the transform multiple times

Example: banana posterior



- **MAP** does not get the mean correctly.
- **MFVI** is better
- **EL$_2$O** fully accommodates variation of the Hessian. Convergence is achieved very quickly.

Example: banana posterior



We get almost perfect PDF (true PDF not in the q family)

Beyond full rank gaussian: gaussian mixtures

- Gaussian mixtures can handle multimodal posteriors and non-bijective mappings

$$q(\boldsymbol{z}) = \sum_j w_j N(\boldsymbol{y}^j; \boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j)\Pi_i \left|\frac{dy_i^j}{dz_i}\right| \equiv \sum_j w_j q^j(\boldsymbol{z})$$

$$\sum_j w_j = 1$$

$$w_j(\boldsymbol{z}) = \frac{w_j q^j(\boldsymbol{z})}{q(\boldsymbol{z})}$$

$$\nabla_{\boldsymbol{z}}\mathcal{L}_q = \sum_j w_j(\boldsymbol{z})\nabla_{\boldsymbol{z}}\mathcal{L}_q = \sum_j w_j(\boldsymbol{z})(\boldsymbol{\Sigma}^j)^{-1}(\boldsymbol{z}-\boldsymbol{\mu}^j)$$

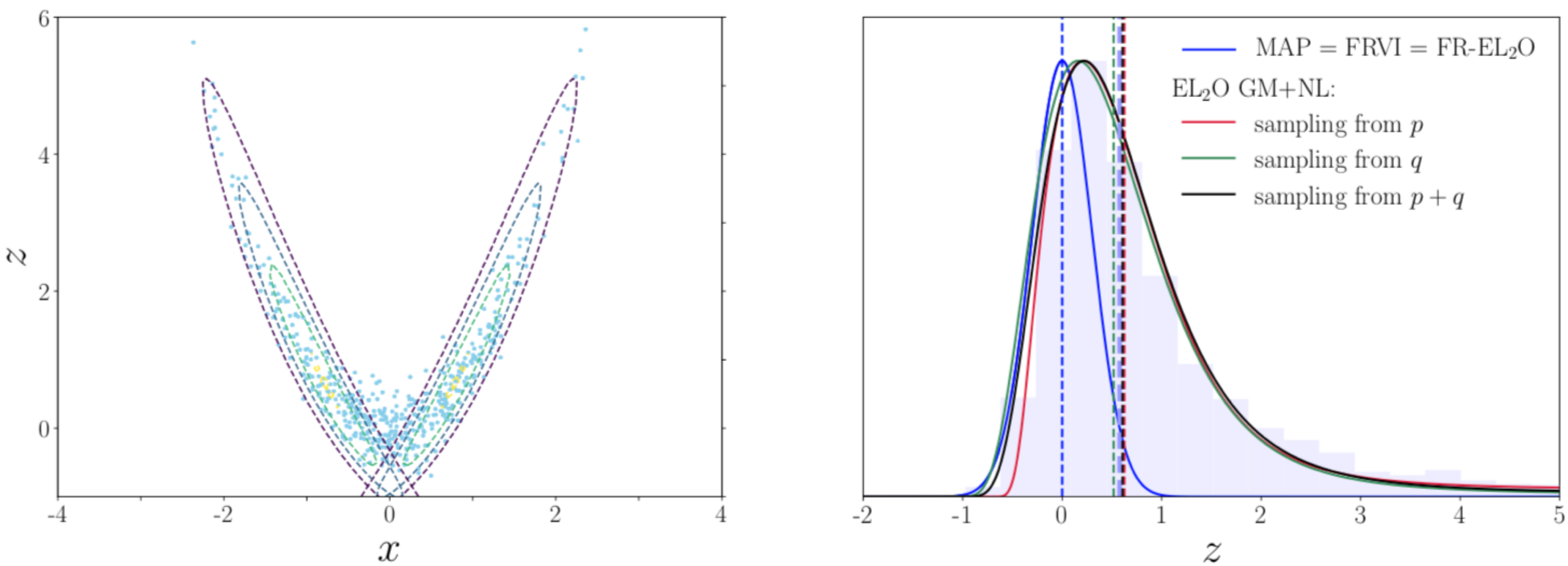
$$\nabla_{\boldsymbol{z}}\nabla_{\boldsymbol{z}}\mathcal{L}_q = \sum_j [\nabla_{\boldsymbol{z}} w_j(\boldsymbol{z})\nabla_{\boldsymbol{z}}\mathcal{L}_q + w_j(\boldsymbol{z})\nabla_{\boldsymbol{z}}\nabla_{\boldsymbol{z}}\mathcal{L}_q]$$

$$= \sum_j w_j(\boldsymbol{z})(\boldsymbol{\Sigma}^j)^{-1} - \sum_i \sum_{j\neq i} \frac{w_i(\boldsymbol{z})w_j(\boldsymbol{z})}{w_i}\left[(\boldsymbol{\Sigma}^j)^{-1}(\boldsymbol{z}-\boldsymbol{\mu}^j)(\boldsymbol{\Sigma}^i)^{-1}(\boldsymbol{z}-\boldsymbol{\mu}^i)\right]$$

Example: forward model posterior

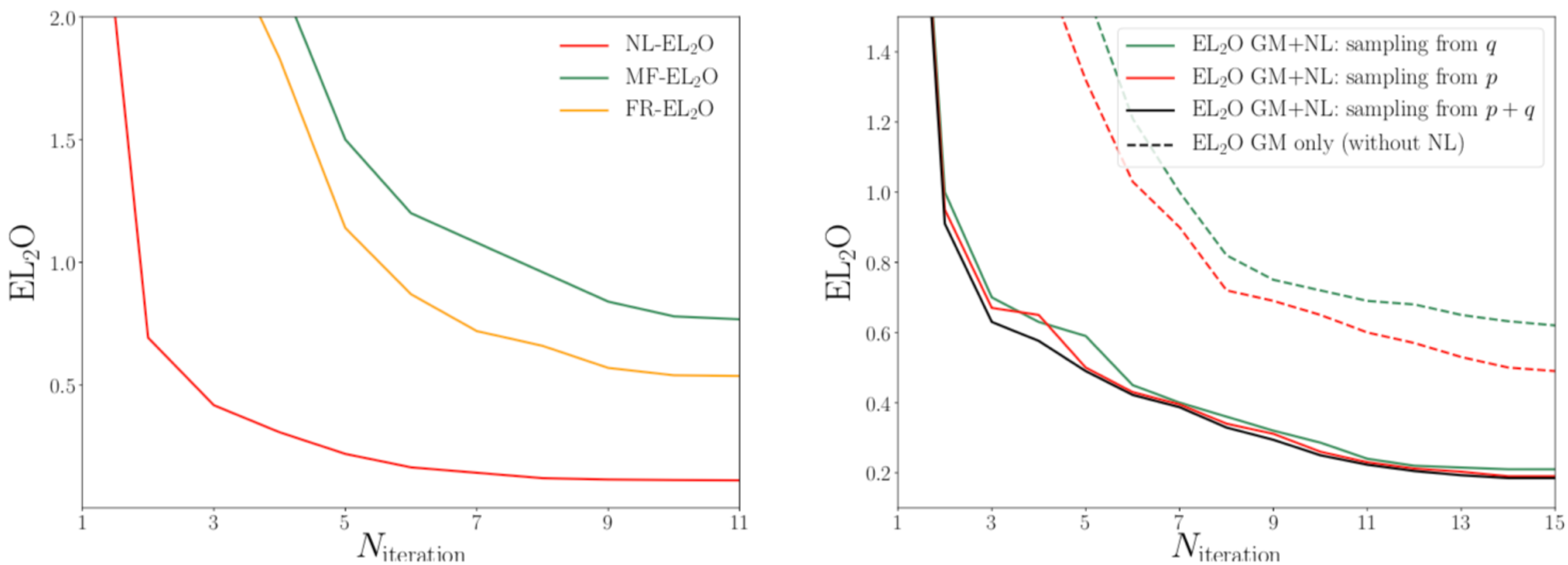
$$\tilde{\mathcal{L}}_p = \frac{1}{2}\left[x\Sigma^{-1}x + (z - x^2)Q^{-1}(z - x^2)\right]$$

- MAP or gaussian VI completely fail
- Solve with 2 symmetric gaussians



EL_2O gives error estimate

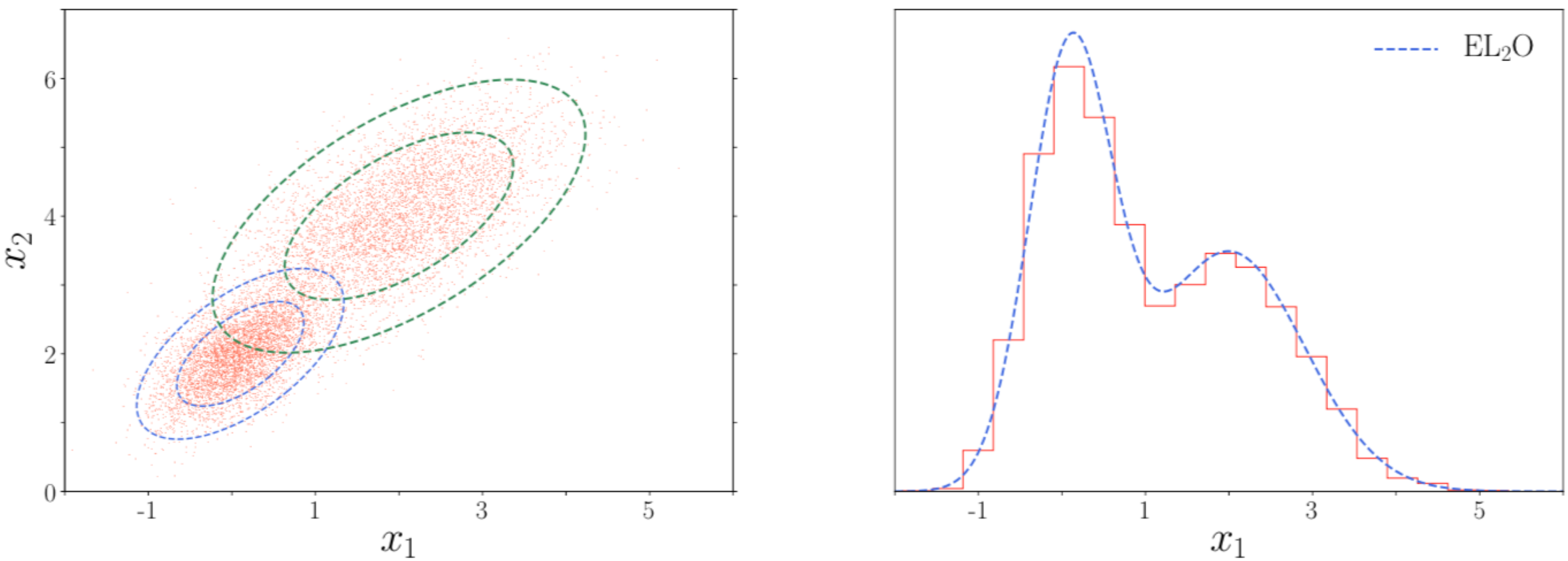
- Another advantage is that L_2 distance can tell us how well the approximation works
- In VI ELBO is meaningless on its own
- EL_2O less than 0.2 is very good



Multi-modal posterior

-Single starting point finds both maxima once we use 2 GM

-Several starting points lead to two different maxima and gaussian mixtures properly normalizes the two



Example: BOSS RSD analysis

- Take summary statistics of galaxy clustering $P_l(k)$, where $l = 0, 2, 4$ are the multipoles of the power spectrum and k is the wavevector.

- **Data**: Measured $P_l(k)$ of the BOSS DR12 galaxies (LOWZ+CMASS)
- **Covariance**: nearly diagonal, but model dependent (sampling variance component)
- **Model**: Predicted $P_l(k)$ which depends on 13 parameters, presented in Hand etal (arXiv:1706.02362)

$$P_{gg}^S(\mathbf{k}) = (1 - f_s)^2 P_{cc}^S(\mathbf{k}) + 2 f_s (1 - f_s) P_{cs}^S(\mathbf{k}) + f_s^2 P_{ss}^S(\mathbf{k})$$

Sample	Description
type A centrals	isolated centrals (no satellites in the same halo)
type B centrals	non-isolated centrals (at least one satellite in same halo)
type A satellites	isolated satellites (no other satellites in same halo)
type B satellites	non-isolated satellites (at least one other satellite in the same halo)

Power Spectrum Model

Free Parameters	
Name [Unit]	Prior
$\alpha_\perp$	$\mathcal{U}(0.8, 1.2)$
$\alpha_\parallel$	$\mathcal{U}(0.8, 1.2)$
f	$\mathcal{U}(0.6, 1.0)$
$\sigma_8(z_\mathrm{eff})$	$\mathcal{U}(0.3, 0.9)$
b_{1,c_A}	$\mathcal{U}(1.2, 2.5)$
f_s	$\mathcal{U}(0, 0.25)$
f_{s_B}	$\mathcal{U}(0, 1)$
$\langle N_{>1,s} \rangle$	$\mathcal{N}(2.4, 0.1)$
σ_c [h^{-1}Mpc]	$\mathcal{U}(0, 3)$
σ_{s_A} [h^{-1}Mpc]	$\mathcal{U}(2, 6)$
γ_{s_A}	$\mathcal{N}(1.45, 0.3)$
γ_{s_B}	$\mathcal{N}(2.05, 0.3)$
$f^{1h}_{s_B s_B}$	$\mathcal{N}(4, 1)$

13 physically motivated parameters:

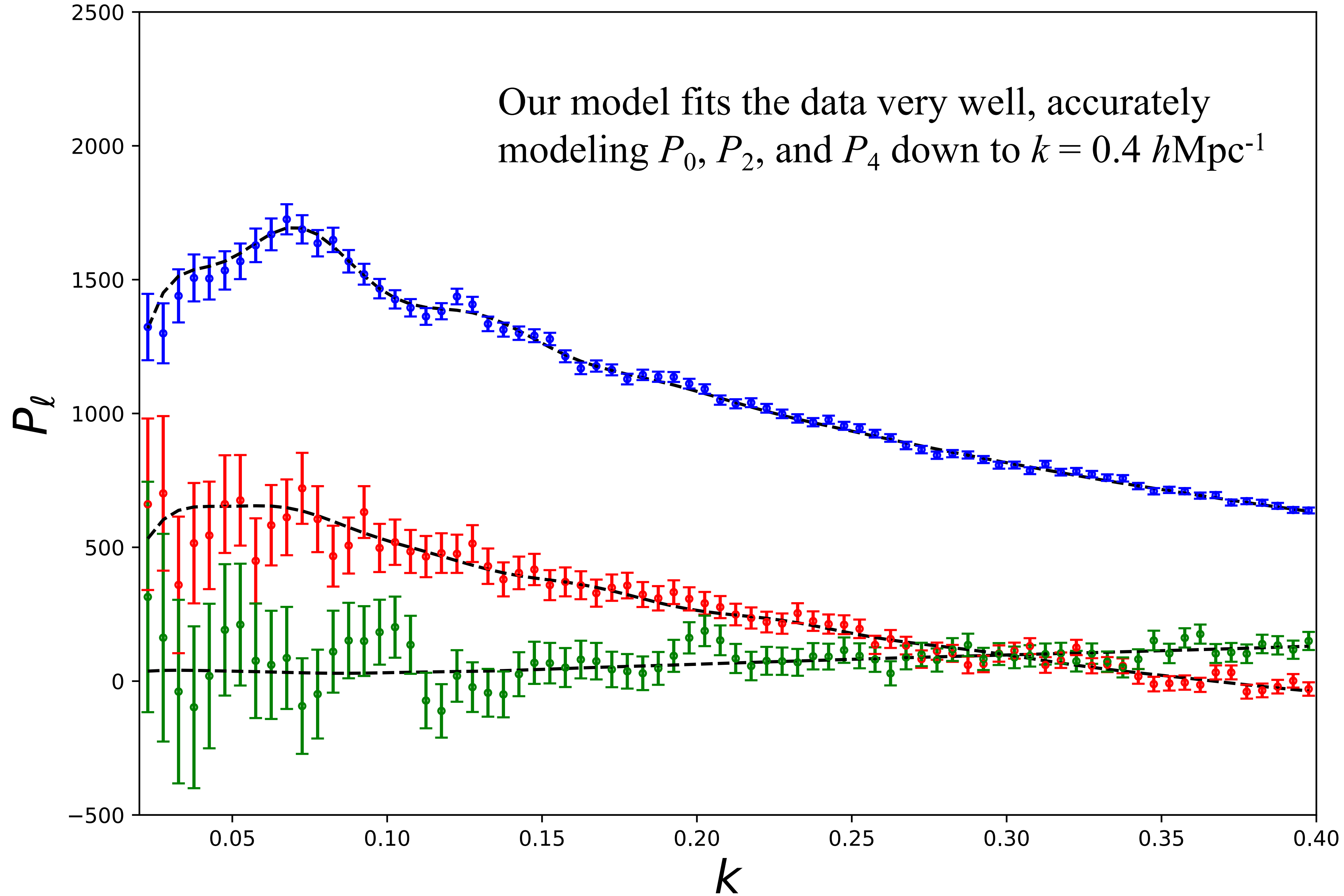
- Cosmology parameters
- Linear bias parameters
- Sample fractions
- Velocity dispersions
- 1-halo amplitudes

- Model evaluation cost: seconds to minutes, because it is PT based

SDSS RSD analysis (B. Yu etal)

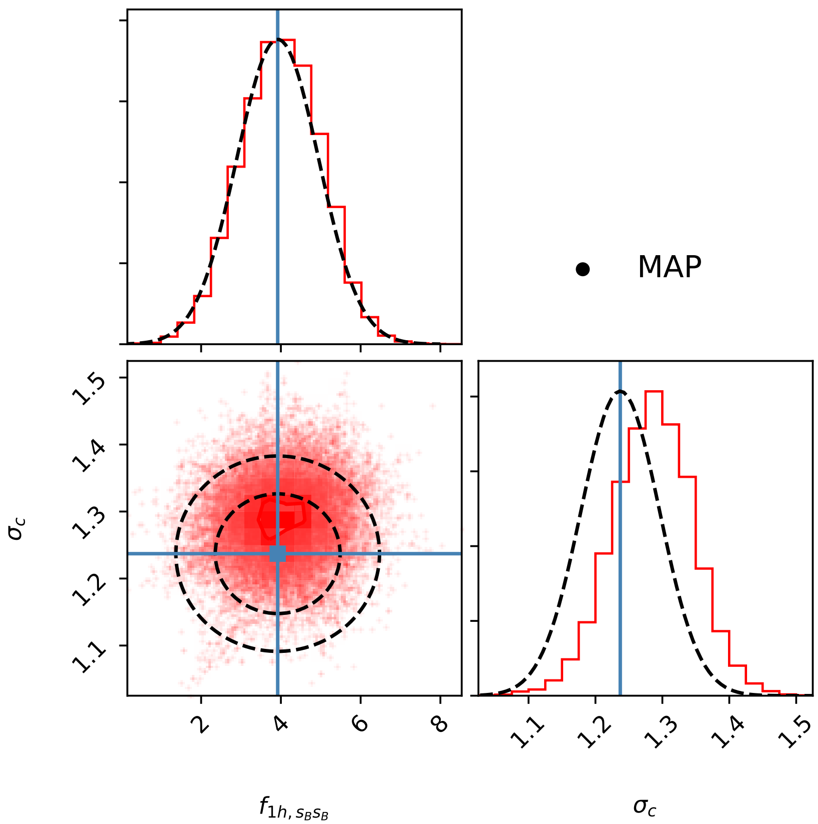
- Analytic derivatives are available for 9 parameters, leaving the remaining 4 to numerical finite difference method.
- Use **Gauss-Newton** approximation to get the Hessian.
- We get a good fit to the data with about 20 iterations
- Different starting points help find global minimum
- Adding a bit of stochasticity helps get out of shallow local minima
- Additionally a few samples to get a good posterior. Total number of iterations 25 (5 calls per iteration because of finite difference)
- Emcee starting at MAP converges with 10^5++ calls
- Emcee starting far from MAP never converges

Example: BOSS RSD analysis

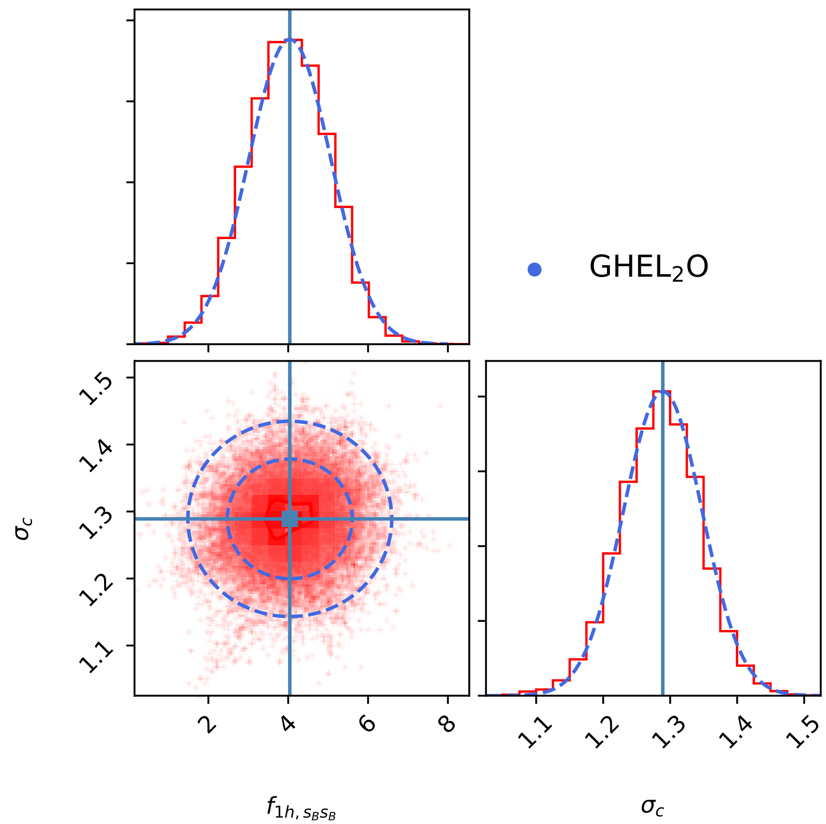


Example: BOSS RSD analysis

MAP



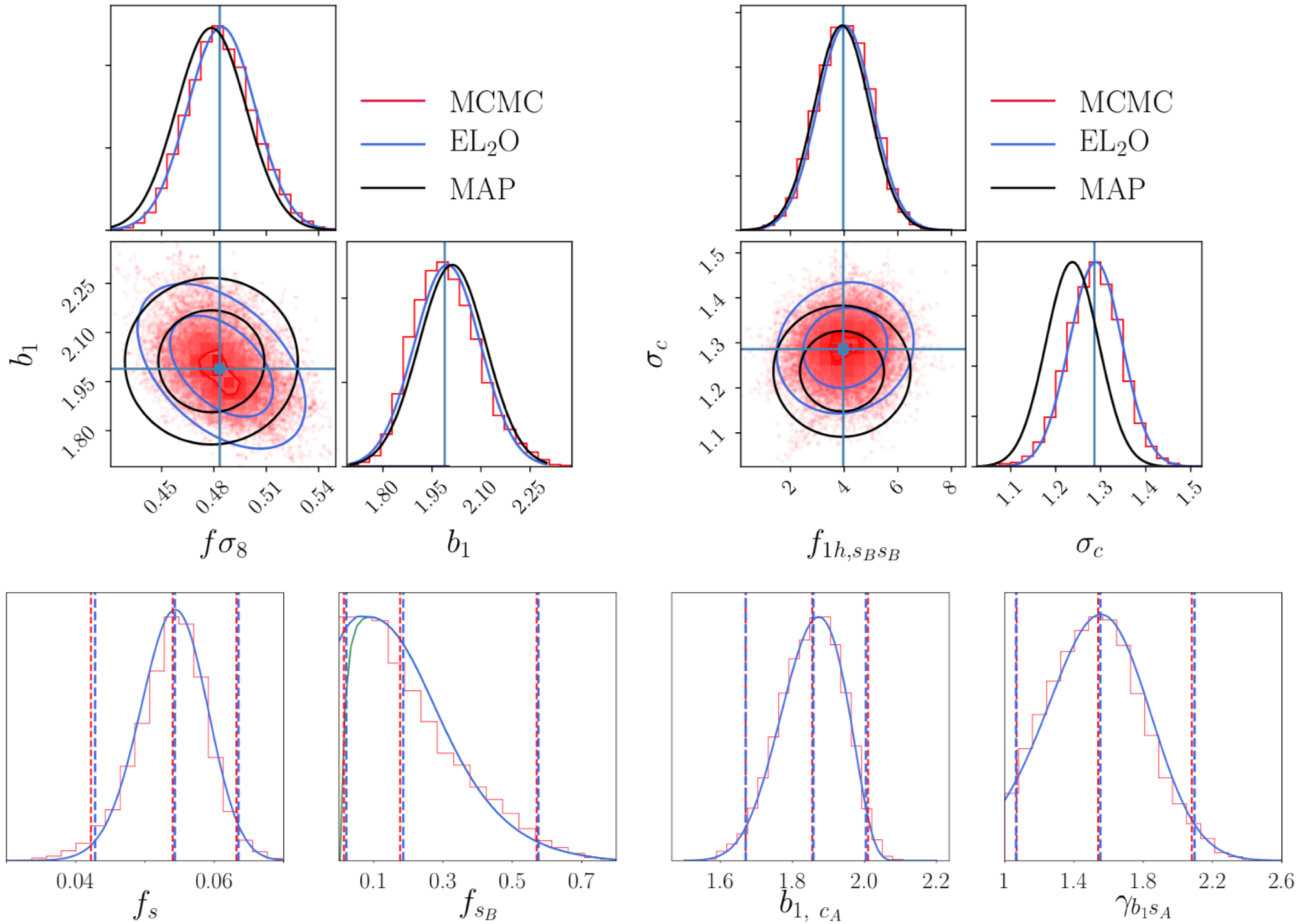
EL$_2$O



EL$_2$O is better than **MAP** and equal to **MCMC**:

Averaging Hessian over samples smooths out small scale noise

LOWZ+CMASS, bin 2 ($0.4 < z < 0.6$)



MCMC
EL$_2$O
MAP
MCMC
EL$_2$O
MAP
b_1
$f\sigma_8$
b_1
σ_c
f_{1h,s_Bs_B}
σ_c
f_s
f_{s_B}
$b_{1,\ c_A}$
$\gamma_{b_1 s_A}$

Summary

- Full analysis of WL data requires sophisticated statistical methods
- These can be broken into several components:
- implicit to explicit likelihood: MAP (initial field reconstruction)
- compression of explicit likelihood into optimal summary statistics and their probability distribution: bandpower analysis, covariance matrix, inverse Wishart
- Bayesian posterior analysis: from summary statistics to cosmological parameters.
- For the last step EL_2O looks very promising as a tool to do inference and may even some day replace MCMC