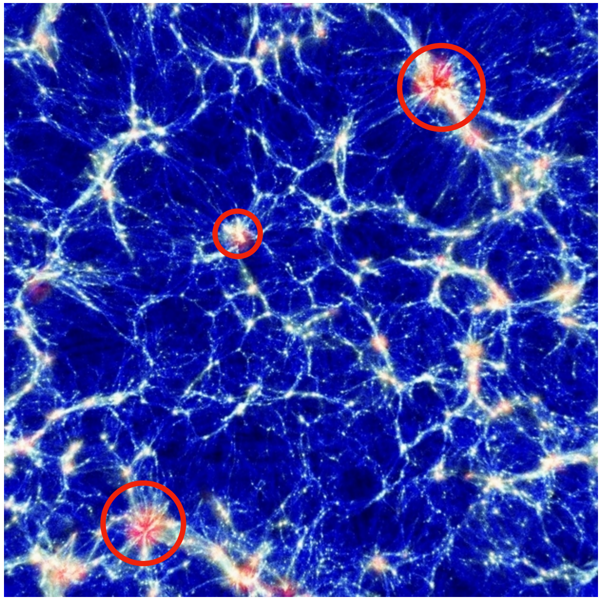
Likelihoods for cluster count cosmology

Constantin Payerne, Calum Murray, Céline Combet.
Laboratoire de Physique Subatomique et de Cosmologie, Grenoble

Rubin LSST France meeting, 05/2022

Are the largest gravitationally bound objects in the Universe

- Form within the largest dark matter halos
- $M > 10^{14}\ M_\odot$
- size of ≈ 1 Mpc
- Recently formed objects, redshift $z \leq 2$: Final step of hierarchical large scale structure formation



Numerical simulations Credits: Klaus Dolag

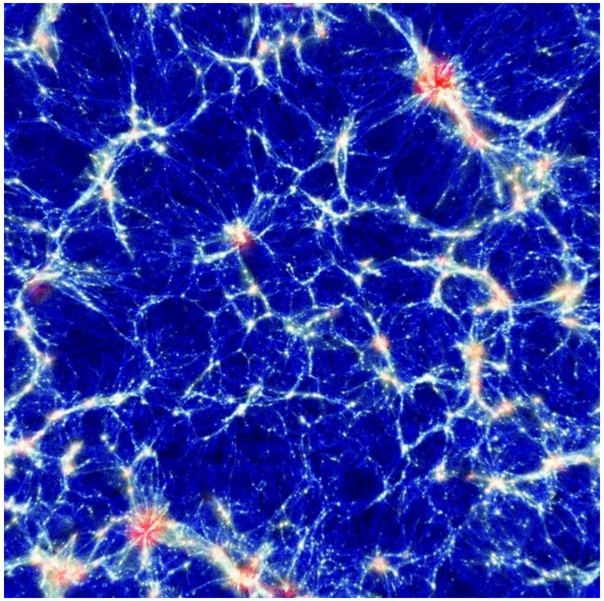
The evolution of mass and redshift distribution of halos is sensitive to cosmology

Basic recipe for cluster abundance cosmology (ideal case)

- From a galaxy cluster survey with known redshifts, masses
- Count the number of galaxy clusters in bins of redshift and mass

$$N(\theta) = \Omega_s \int_{z_1}^{z_2} dz \underbrace{\frac{d^2V(z)}{dz d\Omega}}_{\text{Differential comoving volume (cosmology)}} \int_{M_1}^{M_2} dM \underbrace{\frac{dn(M,z)}{dM}}_{\text{Halo mass function } (\Omega_m, \sigma_8)}$$

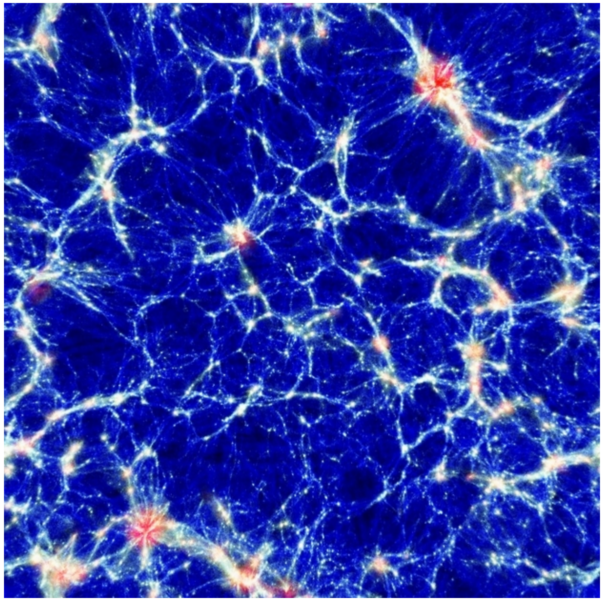
Comparing the observed abundance $\widehat{N}$ to the prediction N = know the statistical properties of cluster count



Cluster abundance as a Poisson variable ?

Counting experiment

- discrete
- un-correlated
- $\rightarrow \widehat{N} \sim \mathscr{P}(\mu = N)$
- Poisson shot noise $\sigma^2(\widehat{N}) = N$



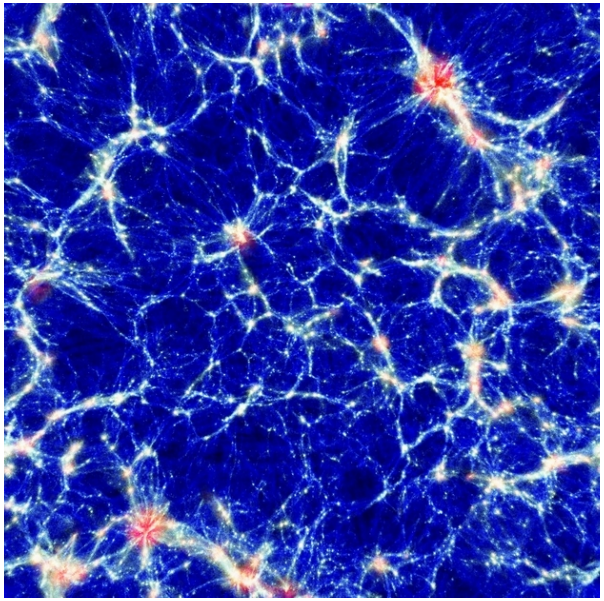
Cluster abundance as a Poisson variable ?

Counting experiment

- discrete
- un-correlated
- $\rightarrow \widehat{N} \sim \mathscr{P}(\mu = N)$
- Poisson shot noise $\sigma^2(\widehat{N}) = N$

The local halo density has spatial fluctuations

- $\delta n_h(\vec{x}) = b\delta_{\rm m}(\vec{x})$
- Cluster count follows matter density field



Cluster abundance as a Poisson variable ?

Counting experiment

- discrete
- un-correlated
- $\rightarrow \widehat{N} \sim \mathscr{P}(\mu = N)$
- Poisson shot noise $\sigma^2(\widehat{N}) = N$

The local halo density has spatial fluctuations

- $\delta n_h(\vec{x}) = b\delta_{\rm m}(\vec{x})$
- Cluster count follows matter density field

Additional variance to cluster abundance shot noise

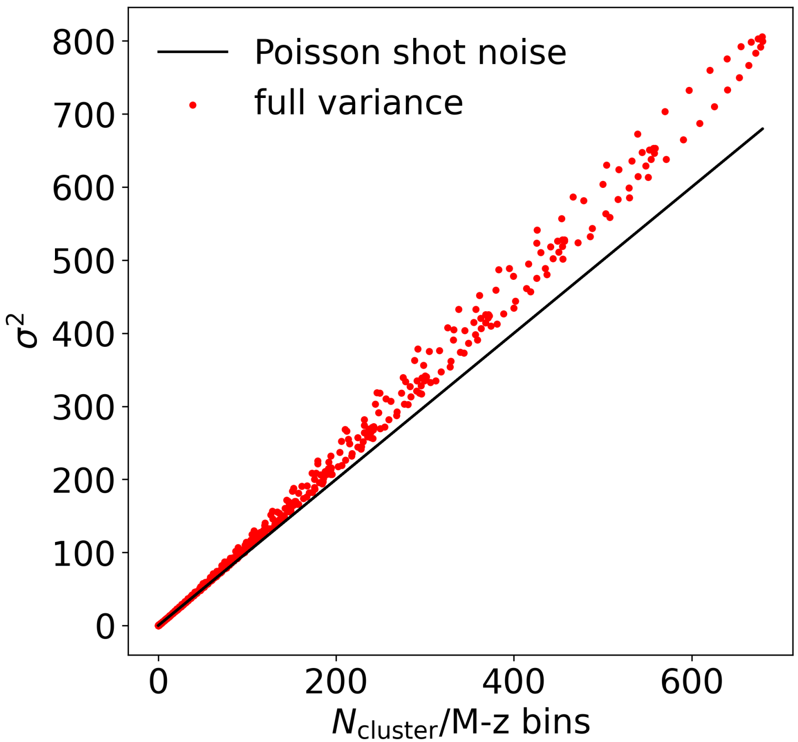
$$\sigma^2(\widehat{N}) = N + \sigma^2_{\rm sample}$$

- $P_{\rm mm}(k)$: matter power spectrum
- Survey geometry (redshift binning, sky area)
- Mass binning

$\rightarrow \sigma^2_{\rm sample}$ increases with the number of halos N per mass-z bins

Covariance matrix for cluster count

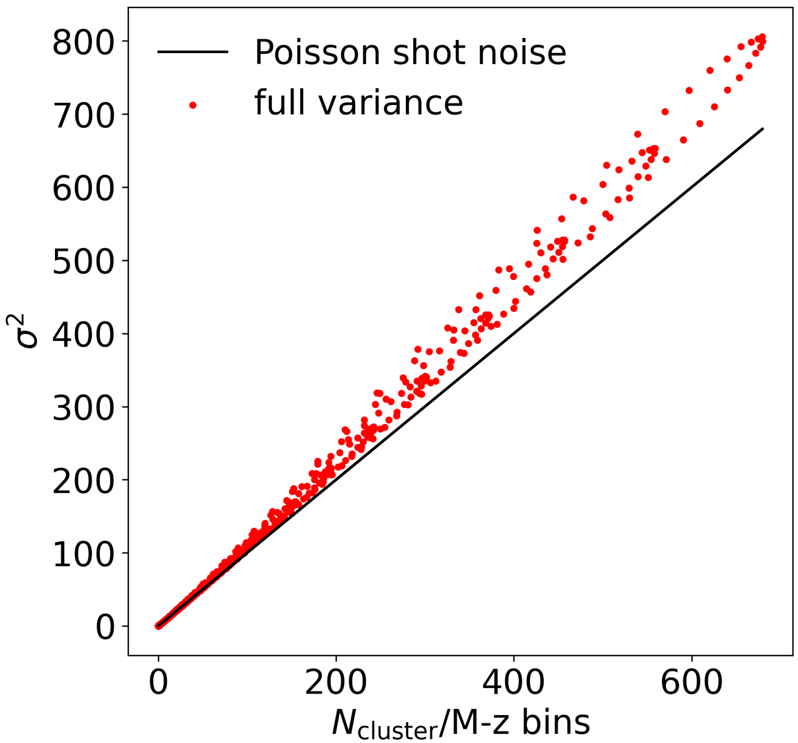
$$\sigma^2(\widehat{N}) = N + \sigma^2_{\rm sample}(N)$$



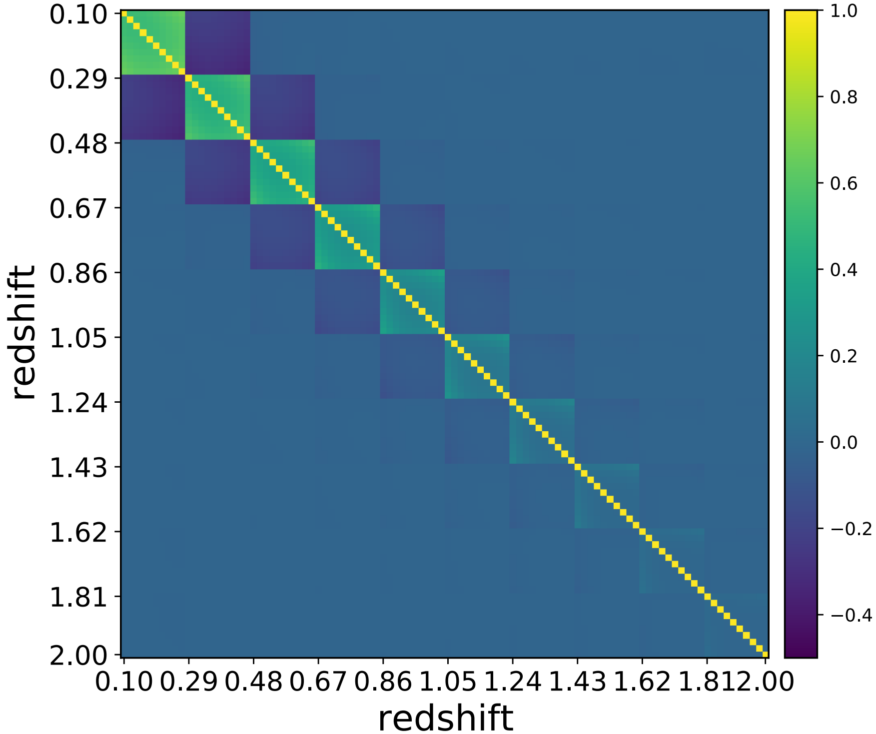
Variance computed with

- PySSC (Lacasa et al. 2021)
- CCL (Chisari et al. 2018)

$$\sigma^2(\widehat{N}) = N + \sigma^2_{\rm sample}(N)$$



+ correlation

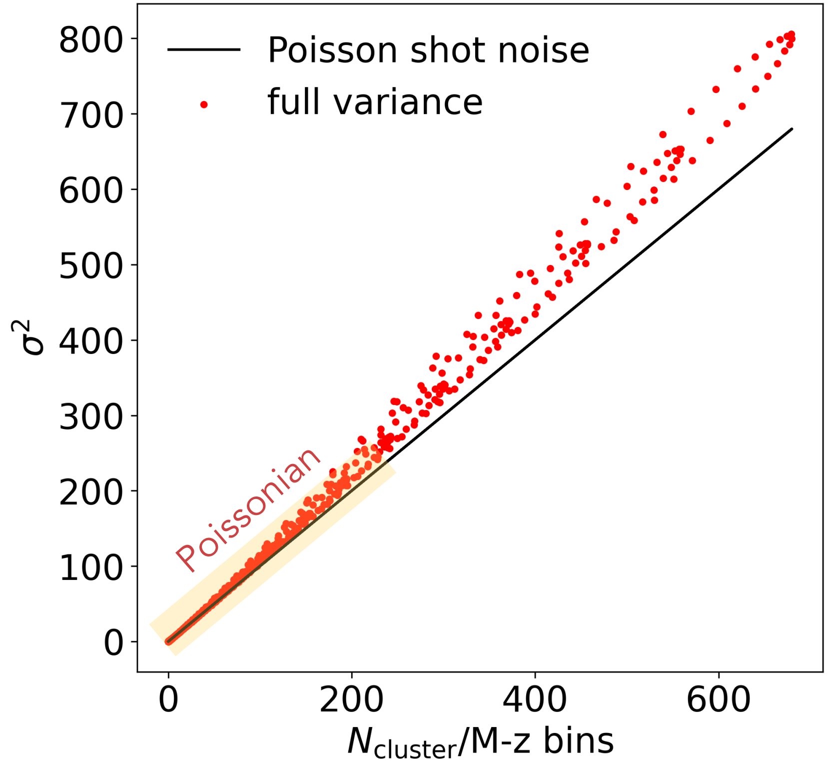


Variance computed with

- PySSC (Lacasa et al. 2021)
- CCL (Chisari et al. 2018)

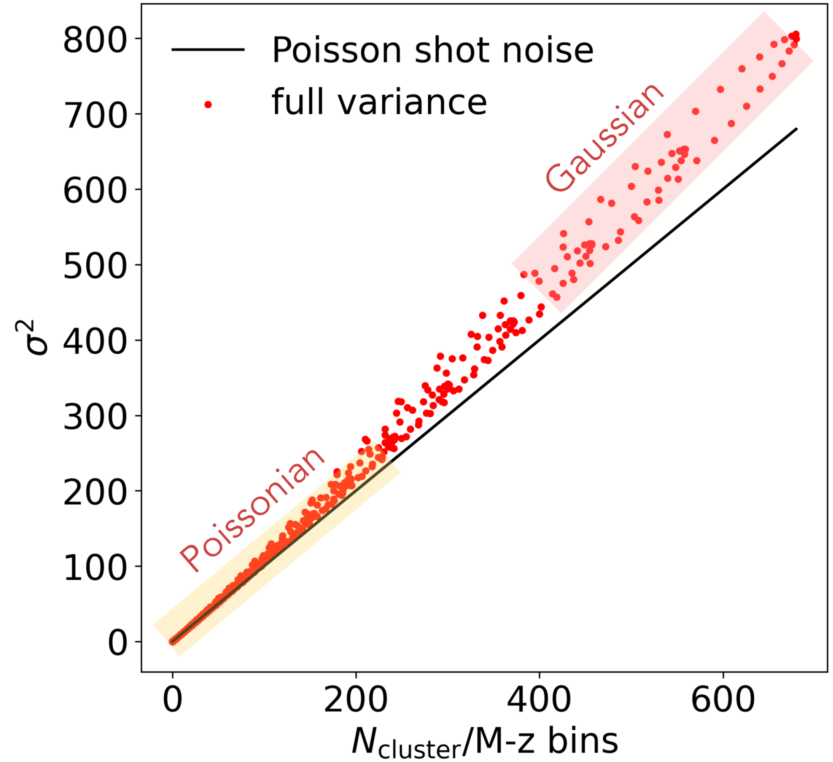
estimate posteriors $p(\overrightarrow{\theta} \mid \widehat{N}) = \pi(\overrightarrow{\theta})\mathscr{L}(\widehat{N} \mid \overrightarrow{\theta})$

	Poissonian
Likelihood	$\frac{N(\overrightarrow{\theta})^{\widehat{N}} e^{-N(\overrightarrow{\theta})}}{\widehat{N}!}$
Condition	$N \gg \sigma^2_{\text{sample}}(N)$
Pros	Discrete Unbinned framework
Cons	No sample variance



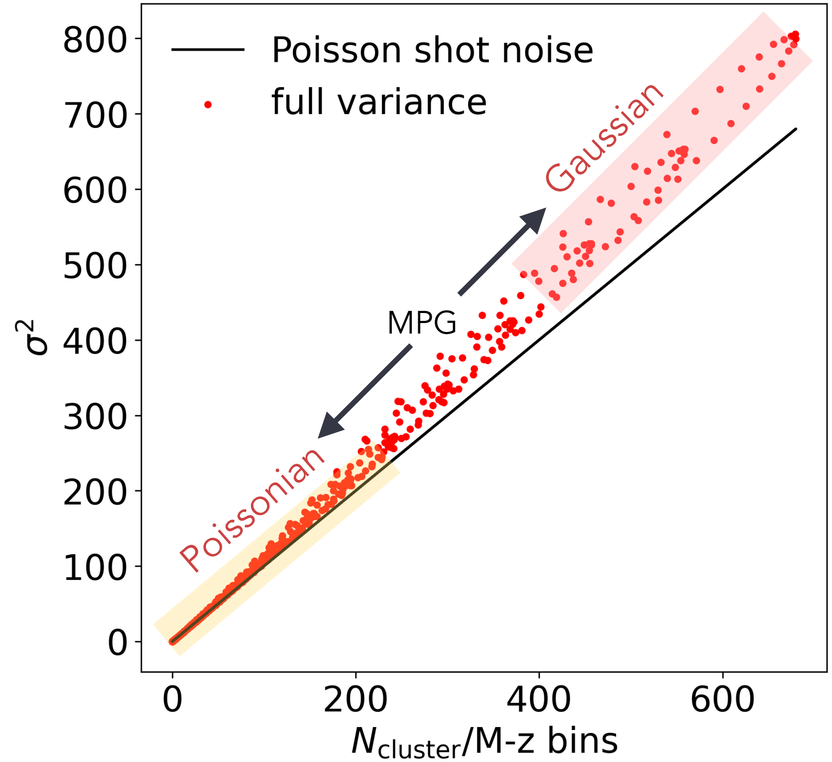
estimate posteriors $p(\overrightarrow{\theta} \mid \widehat{N}) = \pi(\overrightarrow{\theta})\mathscr{L}(\widehat{N} \mid \overrightarrow{\theta})$

	Poissonian	Gaussian
Likelihood	$\frac{N(\overrightarrow{\theta})^{\widehat{N}} e^{-N(\overrightarrow{\theta})}}{\widehat{N}!}$	$\propto e^{-\frac{1}{2}[\overrightarrow{\widehat{N} - N(\overrightarrow{\theta})}]^T \Sigma^{-1} [\overrightarrow{\widehat{N} - N(\overrightarrow{\theta})}]}$
Condition	$N \gg \sigma^2_{\mathrm{sample}}(N)$	$N \sim \sigma^2_{\mathrm{sample}}(N)$
Pros	Discrete Unbinned framework	Sample variance
Cons	No sample variance	No discrete sampling No unbinned framework



estimate posteriors $p(\vec{\theta}\,|\,\widehat{N}) = \pi(\vec{\theta})\mathscr{L}(\widehat{N}\,|\,\vec{\theta})$

	Poissonian	Gaussian
Likelihood	$\dfrac{N(\vec{\theta})^{\widehat{N}}\, e^{-N(\vec{\theta})}}{\widehat{N}\,!}$	$\propto e^{-\frac{1}{2}[\overrightarrow{\widehat{N}-N(\vec{\theta})}]^T\Sigma^{-1}[\overrightarrow{\widehat{N}-N(\vec{\theta})}]}$
Condition	$N \gg \sigma^2_{\rm sample}(N)$	$N \sim \sigma^2_{\rm sample}(N)$
Pros	Discrete Unbinned framework	Sample variance
Cons	No sample variance	No discrete sampling No unbinned framework



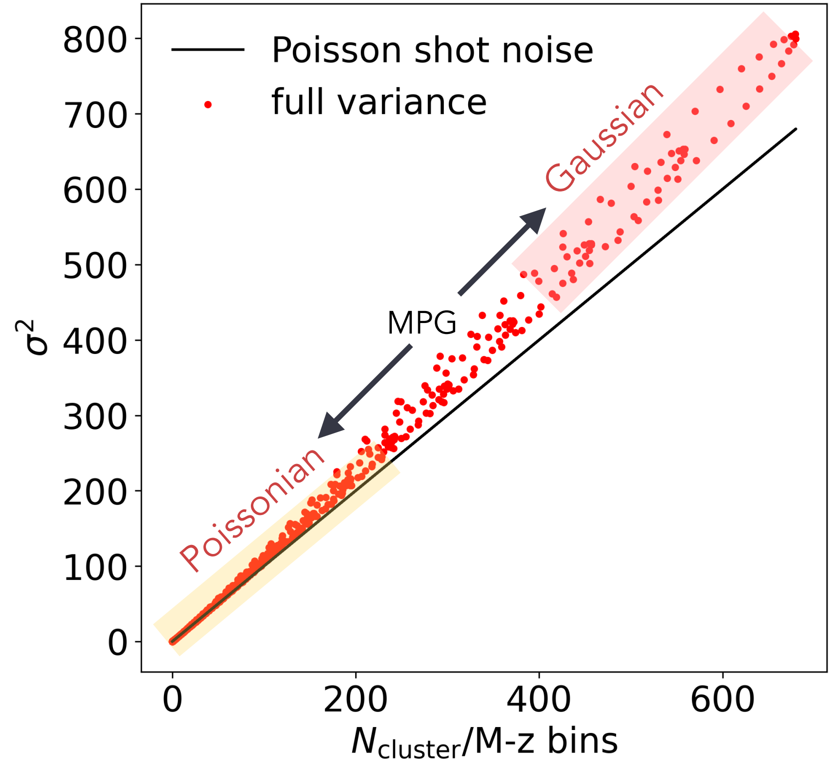
Multivariate Poisson-Gaussian (Hu & Kravtsov 2003)

$$\mathscr{L}(\widehat{N}\,|\,\vec{\theta}) \propto \int d\vec{x}\;\underbrace{\mathscr{G}[\vec{x}\,|\,\vec{N}(\theta)]}_{\text{Gaussian matter density field}} \times \underbrace{\prod_{k=1}^{n} \mathscr{P}[\widehat{N}_k\,|\,x_k]}_{\text{Poisson sampling}}$$

Cosmological constraints: which likelihood to use ?

estimate posteriors $p(\vec{\theta}\mid\widehat{N}) = \pi(\vec{\theta})\mathscr{L}(\widehat{N}\mid\vec{\theta})$

	Poissonian	Gaussian
Likelihood	$\dfrac{N(\vec{\theta})^{\widehat{N}}\, e^{-N(\vec{\theta})}}{\widehat{N}\,!}$	$\propto e^{-\frac{1}{2}[\overrightarrow{\widehat{N}-N(\vec{\theta})}]^T\Sigma^{-1}[\overrightarrow{\widehat{N}-N(\vec{\theta})}]}$
Condition	$N \gg \sigma^2_{\rm sample}(N)$	$N \sim \sigma^2_{\rm sample}(N)$
Pros	Discrete Unbinned framework	Sample variance
Cons	No sample variance	No discrete sampling No unbinned framework



Multivariate Poisson-Gaussian (Hu & Kravtsov 2003)

$$\mathscr{L}(\widehat{N}\mid\vec{\theta}) \propto \int d\vec{x}\ \mathscr{G}[\vec{x}\mid\vec{N}(\theta)] \times \prod_{k=1}^{n} \mathscr{P}[\widehat{N}_k\mid x_k]$$

Gaussian matter density field Poisson sampling

Gaussian: $N \sim \sigma^2_{\rm sample}$

Poissonian: $N \gg \sigma^2_{\rm sample}$

$\boxed{\text{Upcoming Rubin LSST} \sim 10^5 \text{ clusters}}$

Contribution of sample variance will be important for future cosmological analysis

Choose MPG to use all possible cosmological information

- Poisson sampling
- Sample variance

$\boxed{\text{Upcoming Rubin LSST} \sim 10^5 \text{ clusters}}$

Contribution of sample variance will be important for future cosmological analysis

Choose MPG to use all possible cosmological information

- Poisson sampling
- Sample variance

- Are constraints stronger with MPG instead of Gaussian/Poissonian?
- Is there an optimal binning?
- Given a likelihood, are the errors correct? (are the likelihoods accurate?)

We present a framework to quantify accuracies of Poisson and Gaussian likelihoods relative to MPG

Standard cosmological analysis: posterior $p(\overrightarrow{\theta}\mid\widehat{N}) = \pi(\overrightarrow{\theta})\mathscr{L}(\widehat{N}\mid\overrightarrow{\theta})$

- Posterior variance must provide wide enough confidence region
- comparable to the spread of best fits obtained from multiple realisations of the data
- Criteria to choose a likelihood instead of another

Standard cosmological analysis: posterior $p(\overrightarrow{\theta} \mid \widehat{N}) = \pi(\overrightarrow{\theta})\mathscr{L}(\widehat{N} \mid \overrightarrow{\theta})$

- Posterior variance must provide wide enough confidence region
- comparable to the spread of best fits obtained from multiple realisations of the data
- Criteria to choose a likelihood instead of another

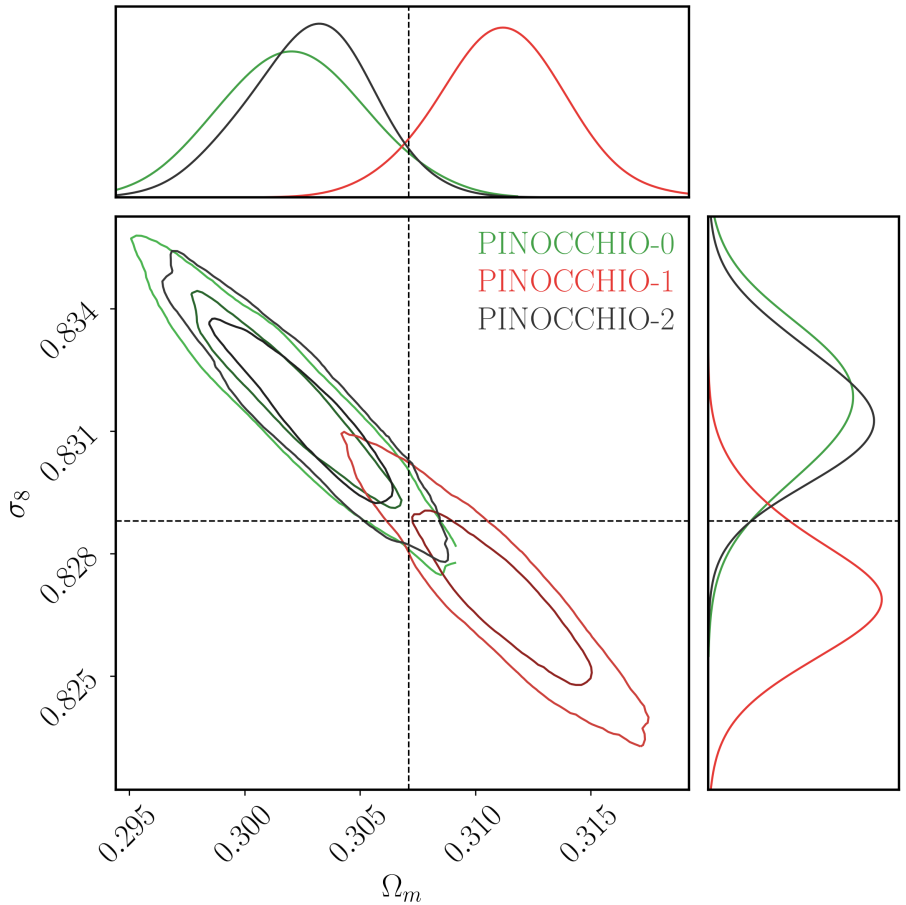
Dataset

1000 simulated dark matter halo catalogs (Euclid collaboration)

- PINOCCHIO algorithm (Monaco et al., 2013)
- Planck cosmology
- Masses calibrated on known halo mass function
- Euclid-like sky area ~ ¼ of full-sky
- 10^5 halos/simulation
- $M > 10^{14}\ M_\odot$

Method

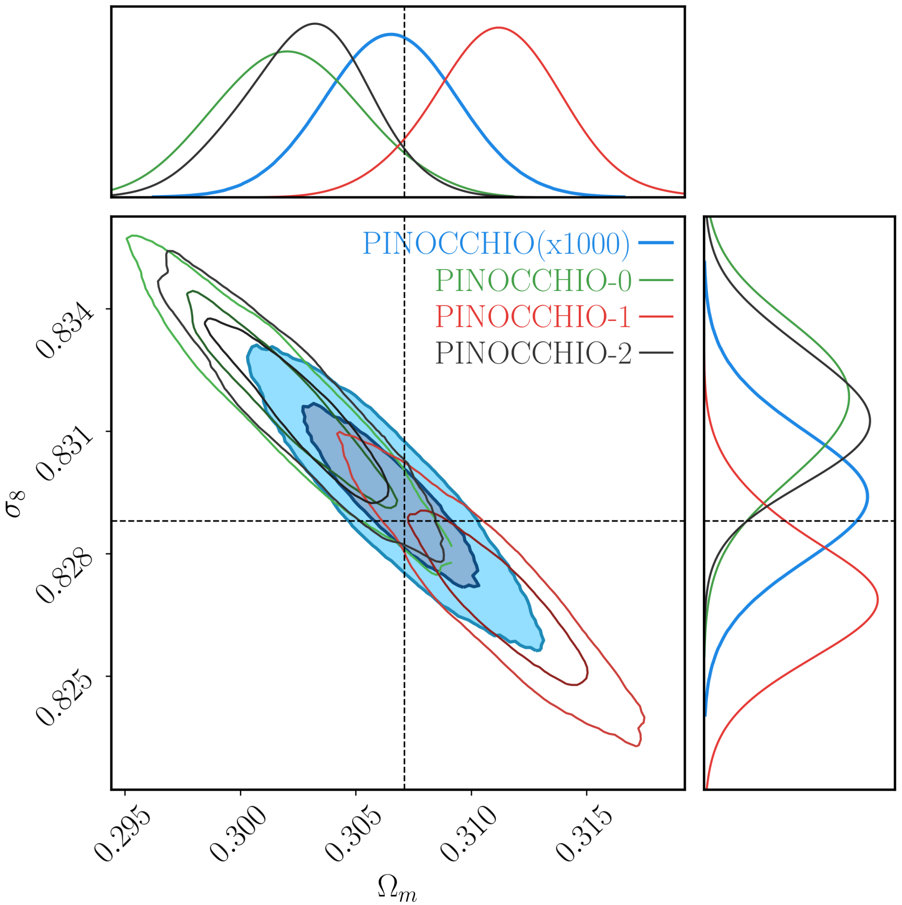
- For each simulation, posterior of ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)



Un-filled: Posterior distributions contours for ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)

Method

- For each simulation, posterior of ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)
- x1000 times over the 1000 simulations
- Look at the distribution of best fits



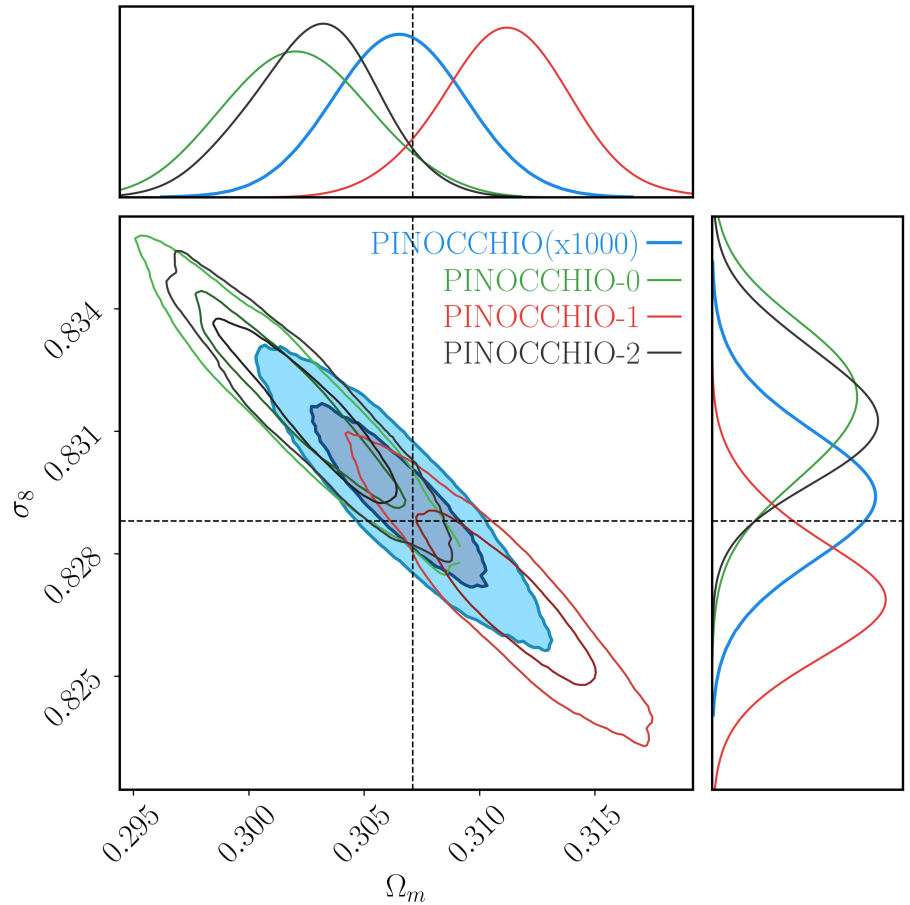
Un-filled: Posterior distributions contours for ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)
Filled: Histogram of the 1000 ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$) individual means

Method

- For each simulation, posterior of ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)
- x1000 times over the 1000 simulations
- Look at the distribution of best fits

Different error definitions

- Individual errors $\sigma(\Omega_{\rm m})_i, \sigma(\sigma_8)_i$
    - Obtained on each simulation
    - Depends on input likelihood



Un-filled: Posterior distributions contours for ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)
Filled: Histogram of the 1000 ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$) individual means

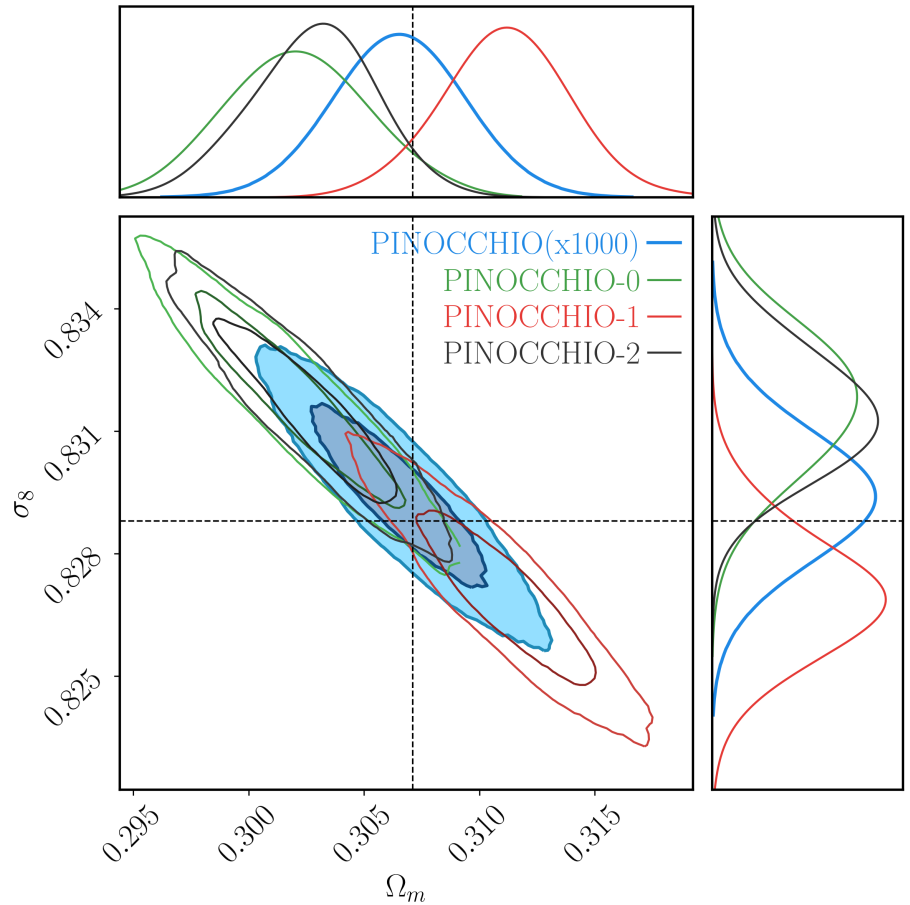
Method

- For each simulation, posterior of ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)
- x1000 times over the 1000 simulations
- Look at the distribution of best fits

Different error definitions

- Individual errors $\sigma(\Omega_{\rm m})_i, \sigma(\sigma_8)_i$
  - Obtained on each simulation
  - Depends on input likelihood
- Standard deviation of 1000 best fits
  - depends on underlying true likelihood
  - Accessed with means over the 1000 simulations

- Compare individual errors to the spread of best fits
- Test if a given likelihood gives robust constraints



Un-filled: Posterior distributions contours for ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$)
Filled: Histogram of the 1000 ($\boldsymbol{\Omega}_m$,$\boldsymbol{\sigma}_8$) individual means

Strategy

Set up

- Redshift $0.2 < z < 1.2$
- Mass $14.2 < \log_{10}(M) < 15.6$
- 3 different binning set-ups for each of 3 likelihoods:

	Redshift bins	Mass bins	# of bins	Average # cluster/bin
#1	4	4	16	5000
#2	20	30	600	150
#3	100	100	10 000	10

=> browse a variety of regimes from shot noise to sample variance

Set up

- Redshift $0.2 < z < 1.2$
- Mass $14.2 < \log_{10}(M) < 15.6$
- 3 different binning set-ups for each of 3 likelihoods:

	Redshift bins	Mass bins	# of bins	Average # cluster/bin
#1	4	4	16	5000
#2	20	30	600	150
#3	100	100	10 000	10

=> browse a variety of regimes from shot noise to sample variance

Methodology

For each likelihood (Poisson, Gaussian, MPG) (x3)

1. (Ω_m, σ_8) posteriors for each simulation **(x1000)**
2. For 3 binning set-ups **(x3)**

$\boxed{\rightarrow \text{9 000 cosmological constraints !}}$

: standard choice is MCMC, **it's too slow**

-> we used Importance Sampling

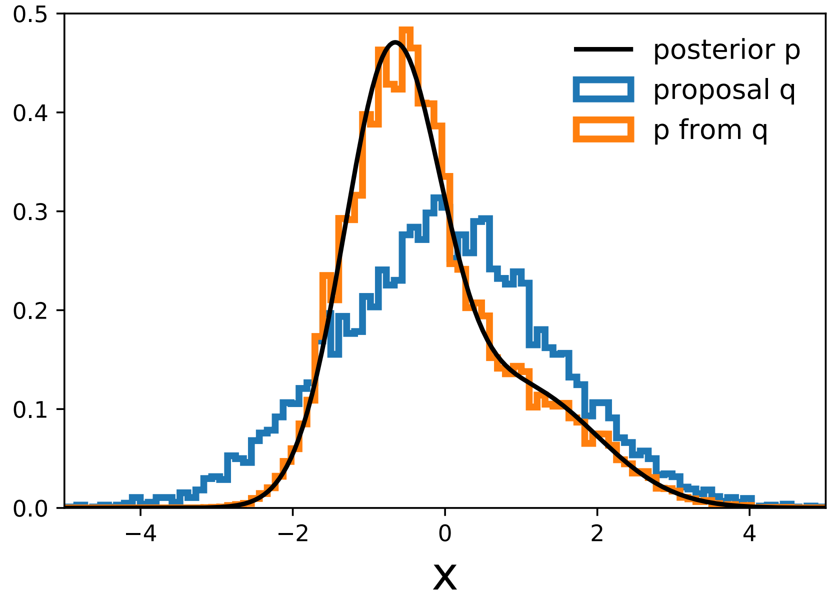
Used to estimate properties of p (posterior) from a known distribution q (proposal)

Method

- q is known, and fast to evaluate
- Draw random q-sample $\{X_i\} \sim q$
- Compute individual weights $w_i = \dfrac{p(X_i)}{q(X_i)}$

Outputs

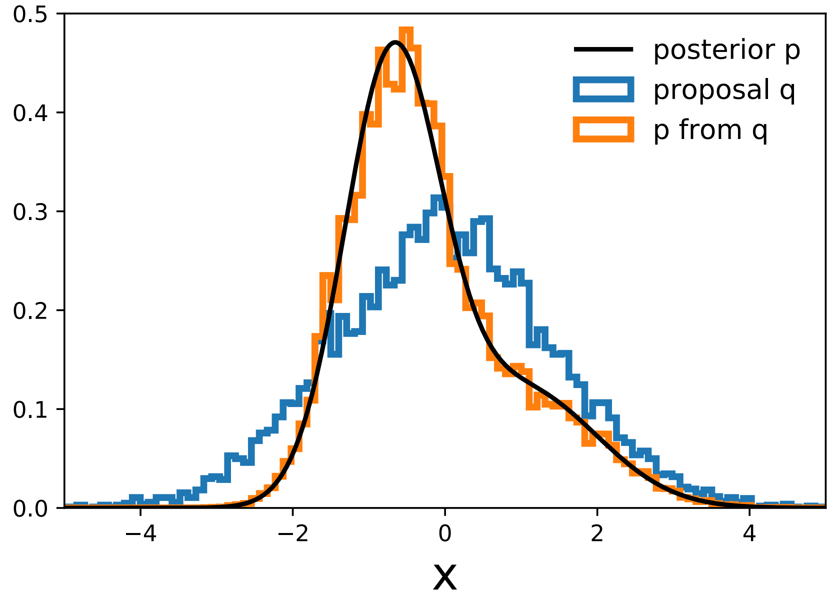
- Moments of p are given $\mathrm{E}(X^n)_p = \dfrac{1}{N_q} \sum_{i=1}^{N_q} w_i \ x_i^n$
- Posterior p are weighted histograms



Used to estimate properties of p (posterior) from a known distribution q (proposal)

Method

- q is known, and fast to evaluate
- Draw random q-sample $\{X_i\} \sim q$
- Compute individual weights $w_i = \dfrac{p(X_i)}{q(X_i)}$



Outputs

- Moments of p are given $\mathrm{E}(X^n)_p = \dfrac{1}{N_q}\sum_{i=1}^{N_q} w_i\ x_i^n$
- Posterior p are weighted histograms

Requirement

- Make appropriate choice of q to "contain" the posterior p

Advantages

- Model pre-computed (long to compute)
- Only limited by likelihood computation time $\mathscr{L}[\,\widehat{N} \mid N(\theta)]$

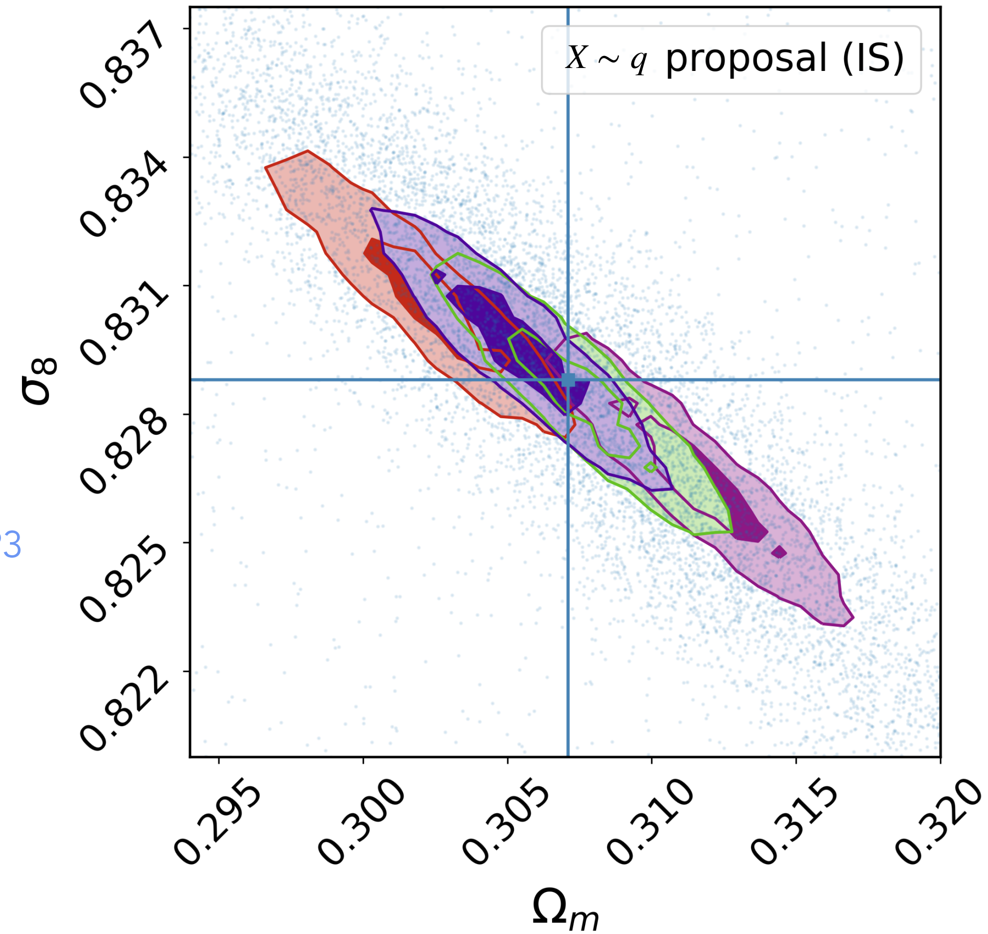
Timescales with Importance Sampling

- $\sim$ 1 sec to 1 min per posterior Poissonian, Gaussian
- $\sim$ 1 to 15 h for MPG

→ repeatability in reasonable timescales*

*multitasks jobs @ CC-IN2P3

Posteriors on (Ω_m, σ_8)

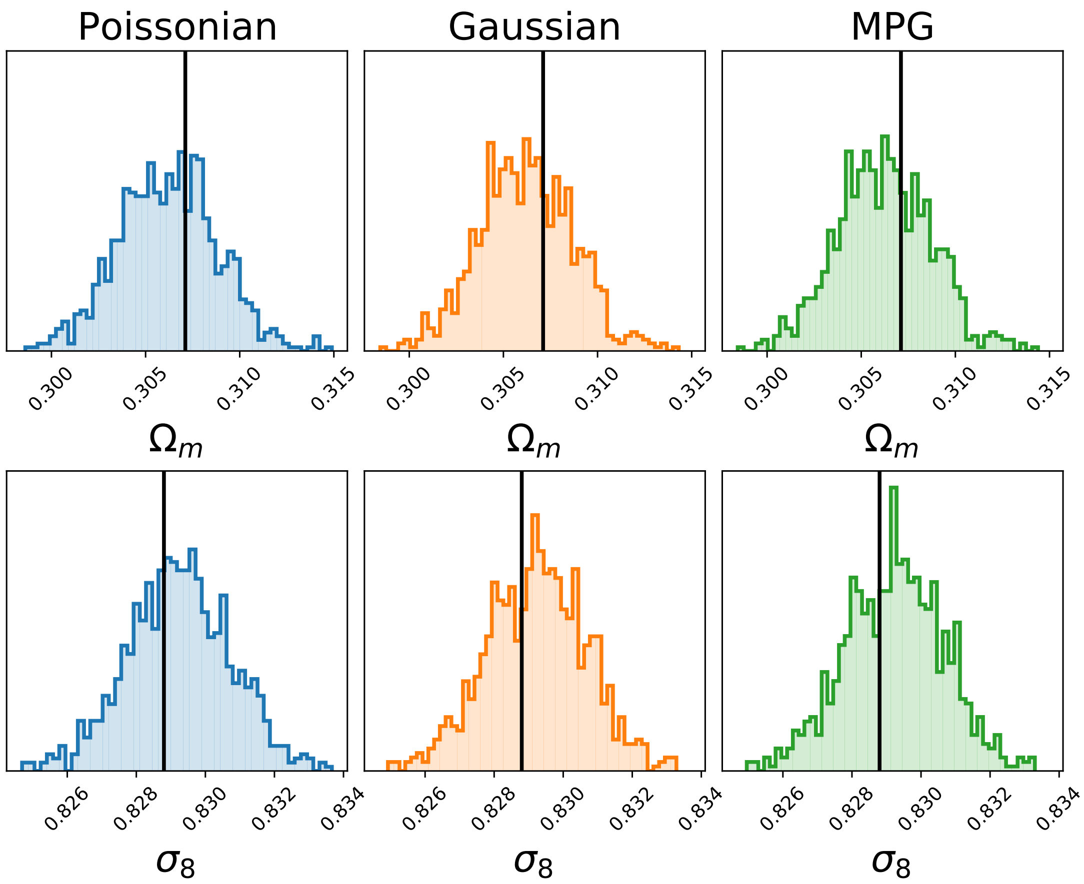


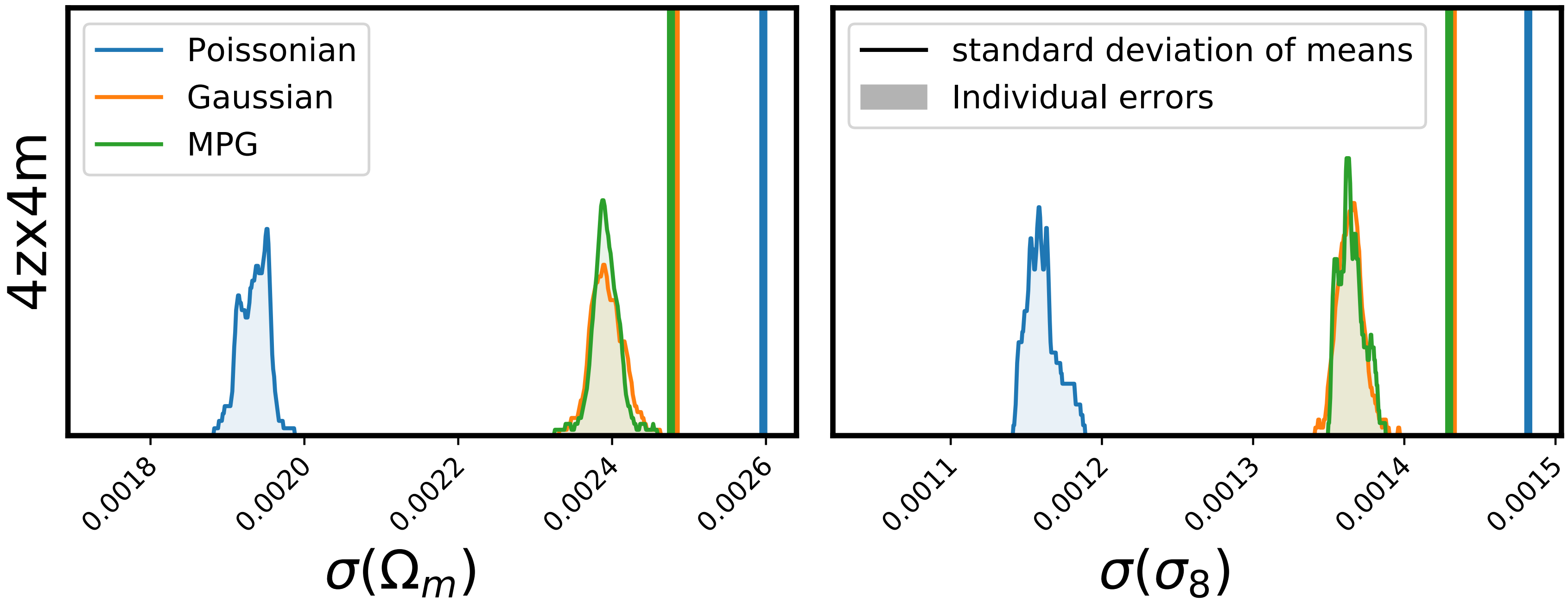
Blue points: sample (importance sampling)
Coloured 2σ contours: posterior distributions of 4 simulations

Results: (4 redshift bins)x(4 mass bins) case

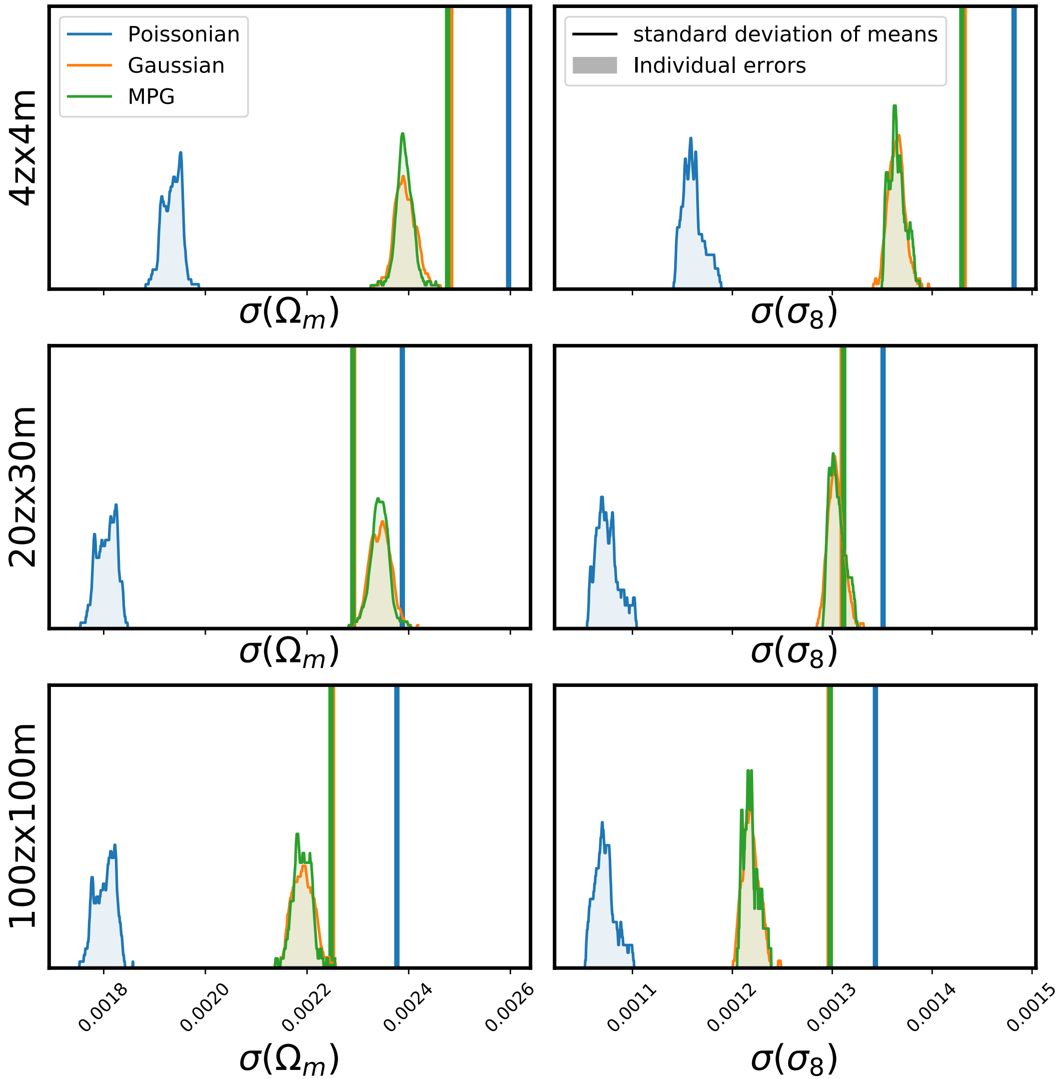
distribution of 1000 means

- Binning setup #1
- Scatter around input cosmology
- Validate the modelling input
- No significant bias on the cosmology





- Individual Poisson errors are underestimated compared to std(1000 means)
- Gaussian & MPG individual errors are slightly underestimated
- Gaussian approximates MPG correctly



- Individual errors decrease

Error comparison

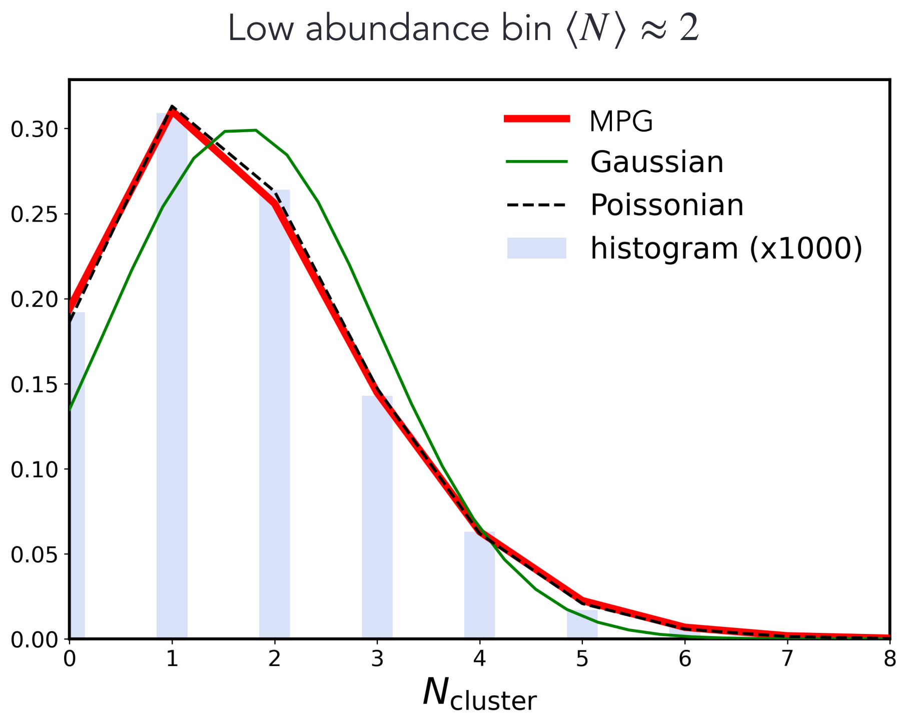
- Poisson individual errors are always underestimated
- Gaussian captures the MPG behaviour

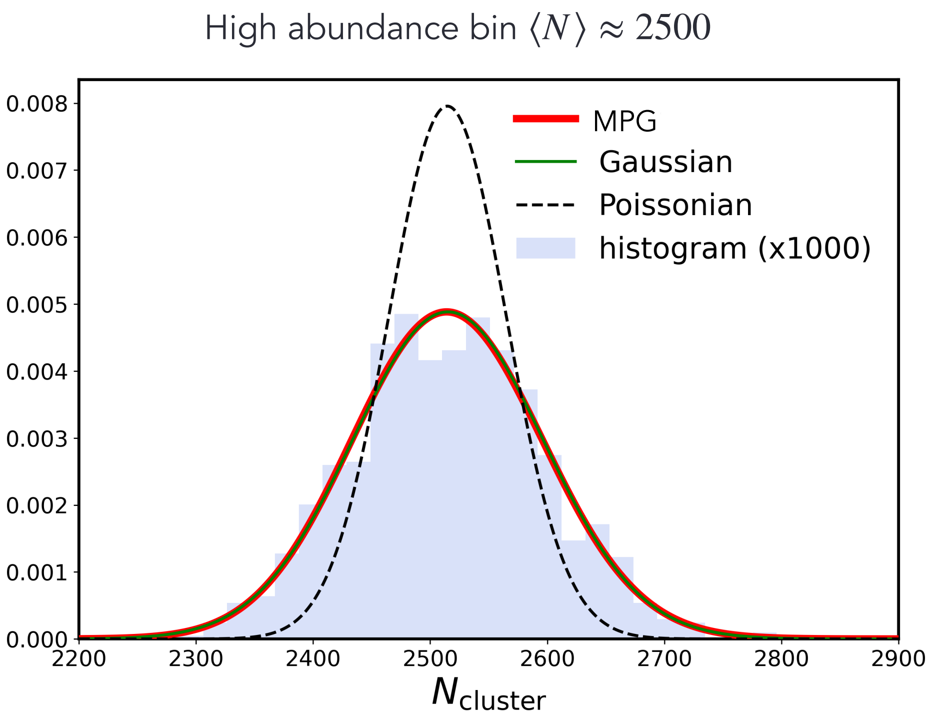
- Increase the number of bins improves constraints (10%)
- Gaussian likelihood is still valid up to 10^4 bins

- We tested the accuracy of likelihoods on cluster abundance with simulations
  - Posterior variance vs spread of best fits
- For wide survey with $f_{\rm sky} = 1/4$
  - Poisson likelihood underestimates errors compared to true error
  - the Gaussian likelihood describes MPG correctly
  - Errors decrease by 10% by increasing the number of bins (16 to 10^4 bins)
- Paper in prep.

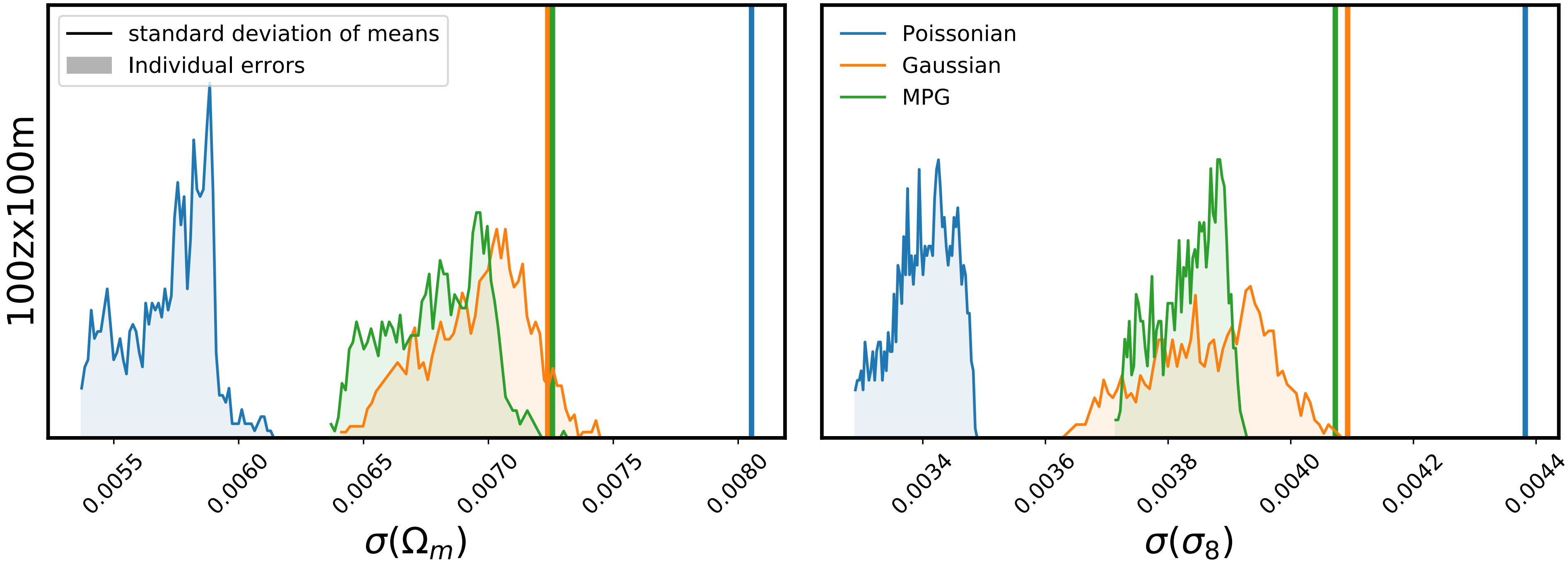
Perspectives

- Switch to unbinned likelihood
  - DESC project on unbinned likelihood with sample variance (M. Penna-Lima)

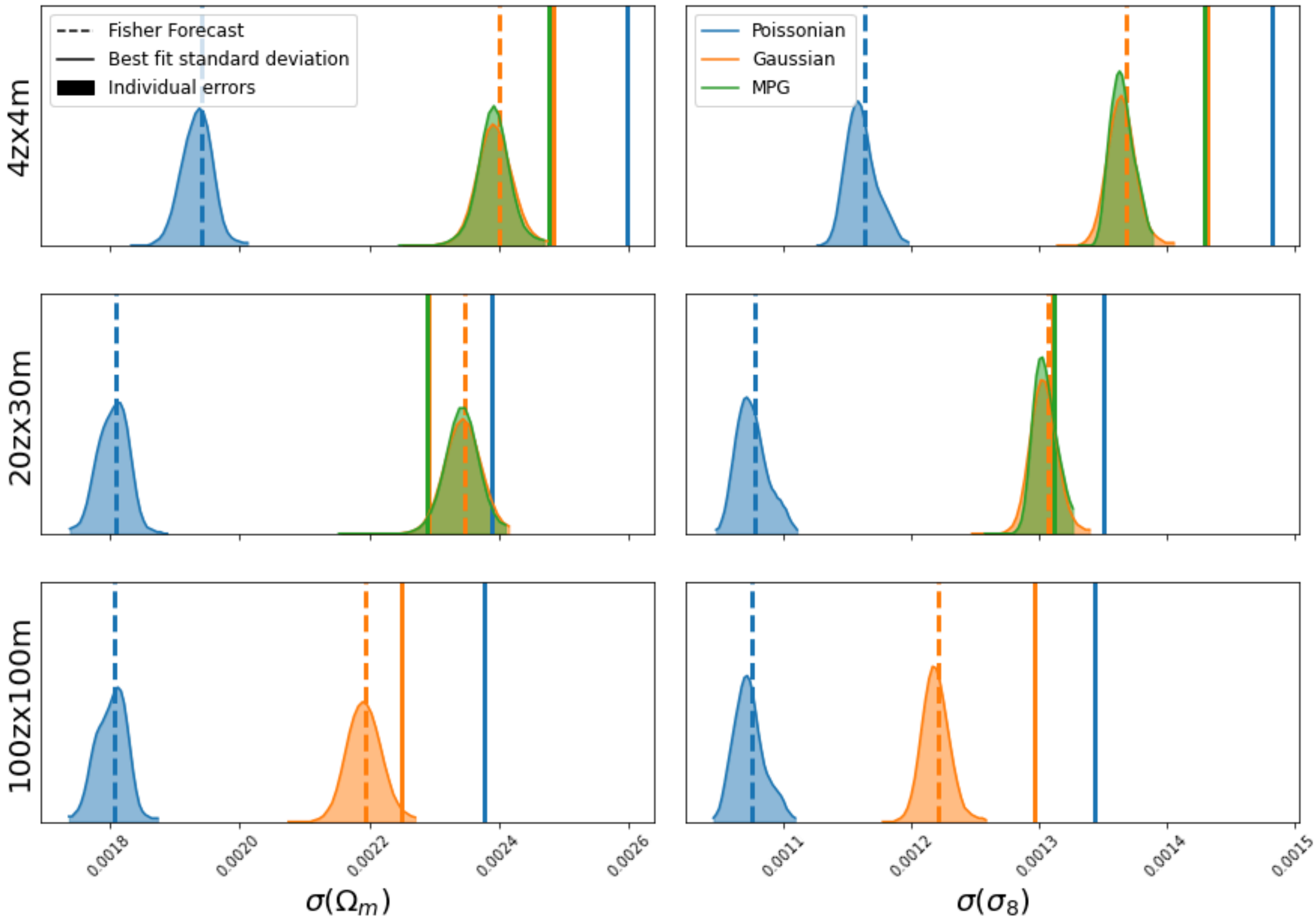

Low abundance bin $\langle N \rangle \approx 2$
MPG
Gaussian
Poissonian
histogram (x1000)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
2
3
4
5
6
7
8
N_{cluster}


High abundance bin $\langle N \rangle \approx 2500$
MPG
Gaussian
Poissonian
histogram (x1000)
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
2200
2300
2400
2500
2600
2700
2800
2900
N_{cluster}

$f_{\text{sky}} = 1/40$

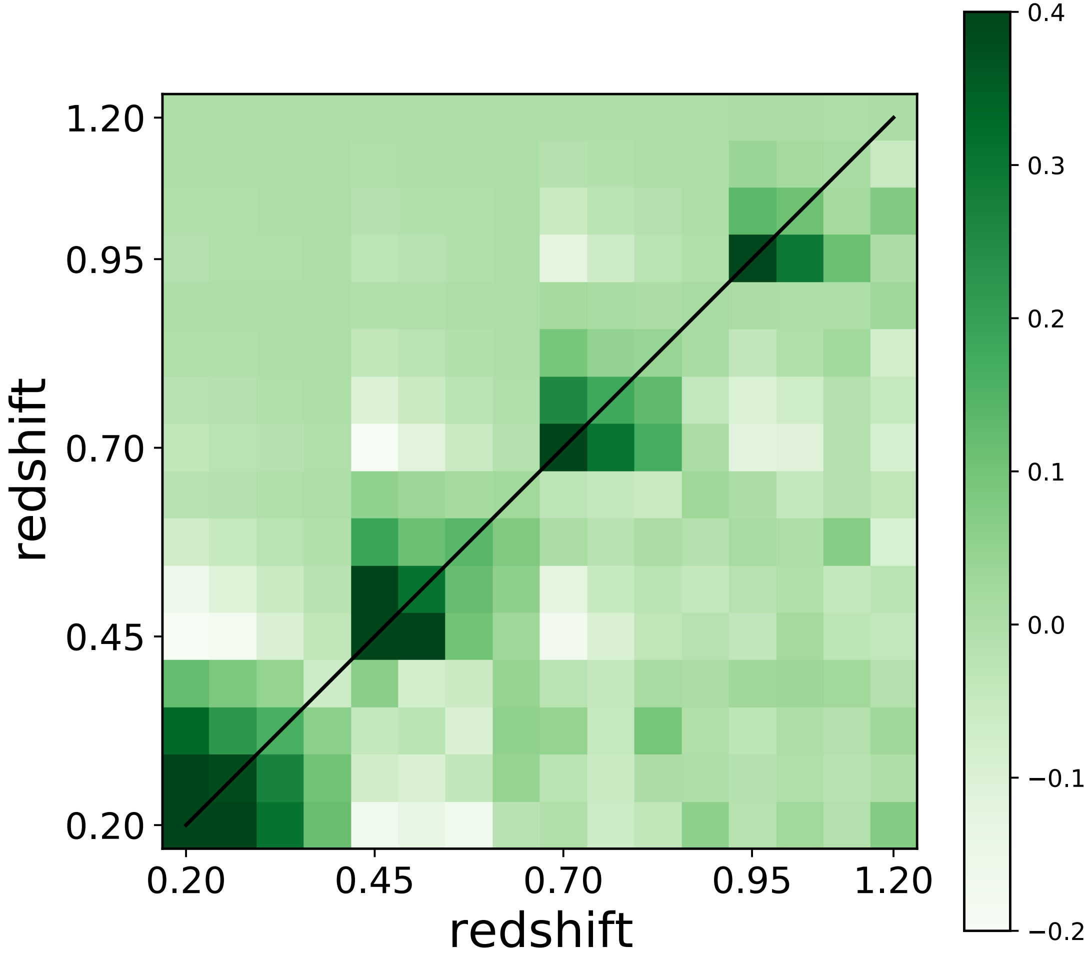


Fisher forecast



$$\mathrm{Cov}(N_{\alpha_1}, N_{\alpha_2}) = N_{\alpha_1}\delta_K^{\alpha_1,\alpha_2} + \langle bN_{\alpha_1}\rangle\langle bN_{\alpha_1}\rangle S_{\alpha_1\alpha_2}$$

$$R = \frac{\text{Sample covariance}}{\text{Shot noise}}$$



$$N(\theta) = \Omega_s \int_{z_1}^{z_2} dz \frac{d^2V(z)}{dz d\Omega} \int_{M_1}^{M_2} dM \frac{dn(M,z)}{dM}$$

Differential comoving volume (cosmology)

Halo mass function $(\Omega_{\mathrm{m}}, \sigma_8)$