Tomographic analysis of photometric galaxy clustering with Euclid

CPPM seminar, October 30th 2023

Vincent Duret

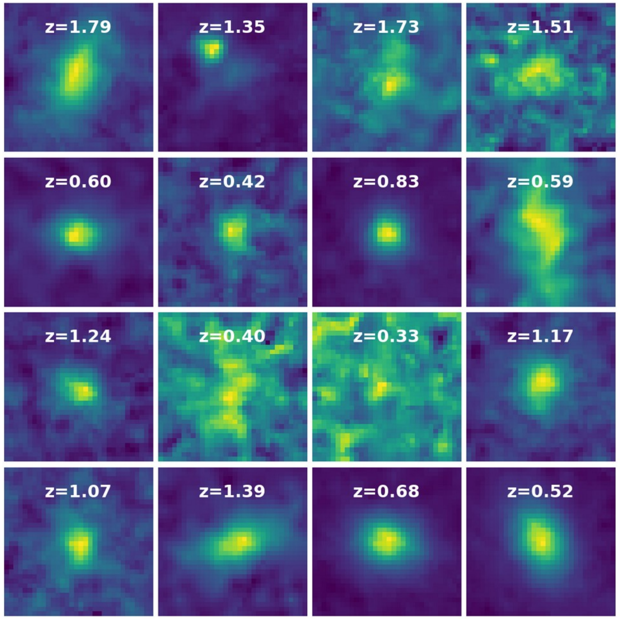
Supervisors : Stéphanie Escoffier, William Gillard

Early commissioning test image, VIS instrument

Early commissioning test image, NISP instrument

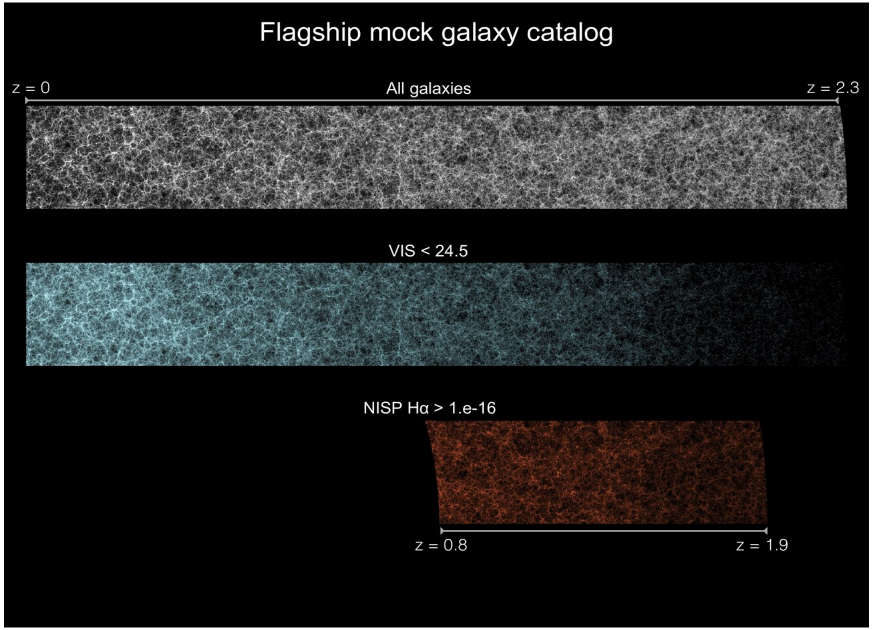
Thesis subject

1) Photometric redshifts calibration of Euclid data using deep learning and multi-band images :



Simulated galaxies
(Science Challenge 8)

2) Tomographic analysis of photometric galaxy clustering with the angular two-point correlation function :

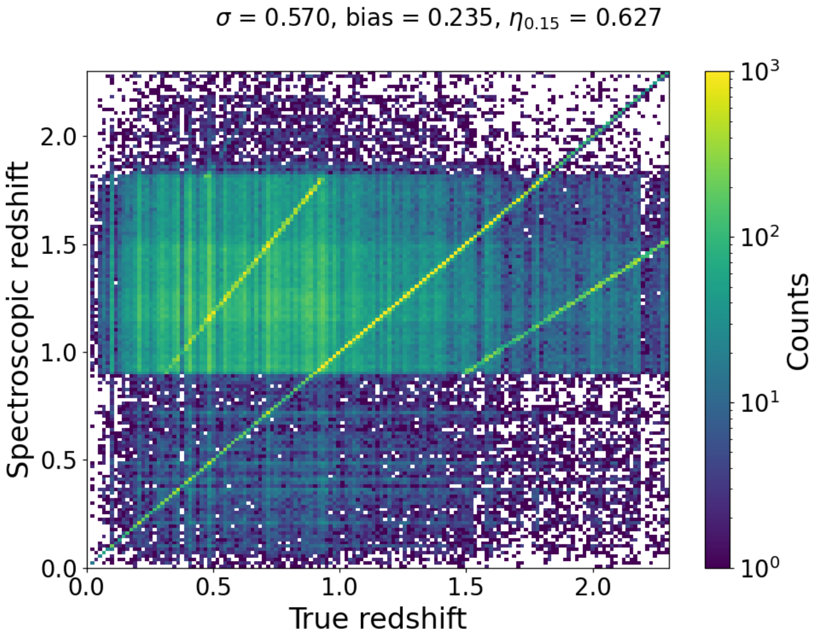


Flagship simulation of the Euclid survey

Photometric redshifts calibration

- Goal : find the relationship between the input galaxy images and their redshift.

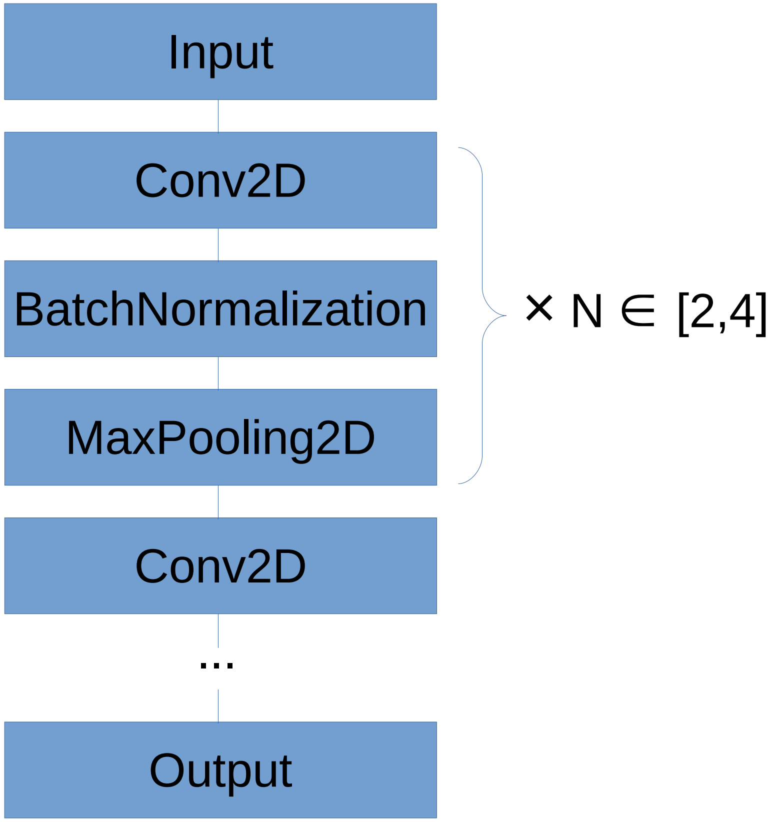
- Method : neural networks + optimization framework (Optuna)
- Networks : CNNs, inception CNNs, ResNet and variants with additional inputs.
- Data : 500000 galaxies simulated from the Euclid Science Challenge 8 + simulated spec-zs.



Spectroscopic vs true redshifts from Flagship

Photometric redshifts calibration

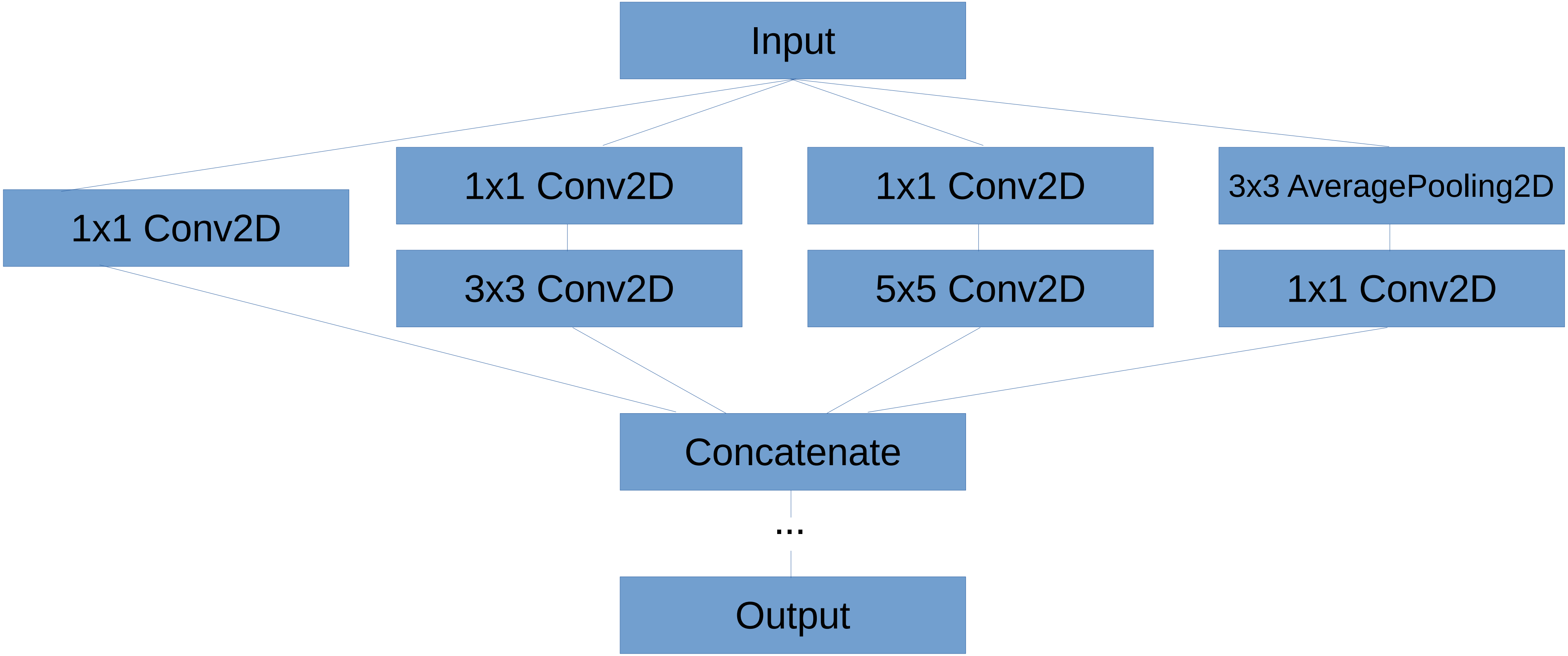
Tested neural networks : sequential CNN



Sequential CNN architecture

Photometric redshifts calibration

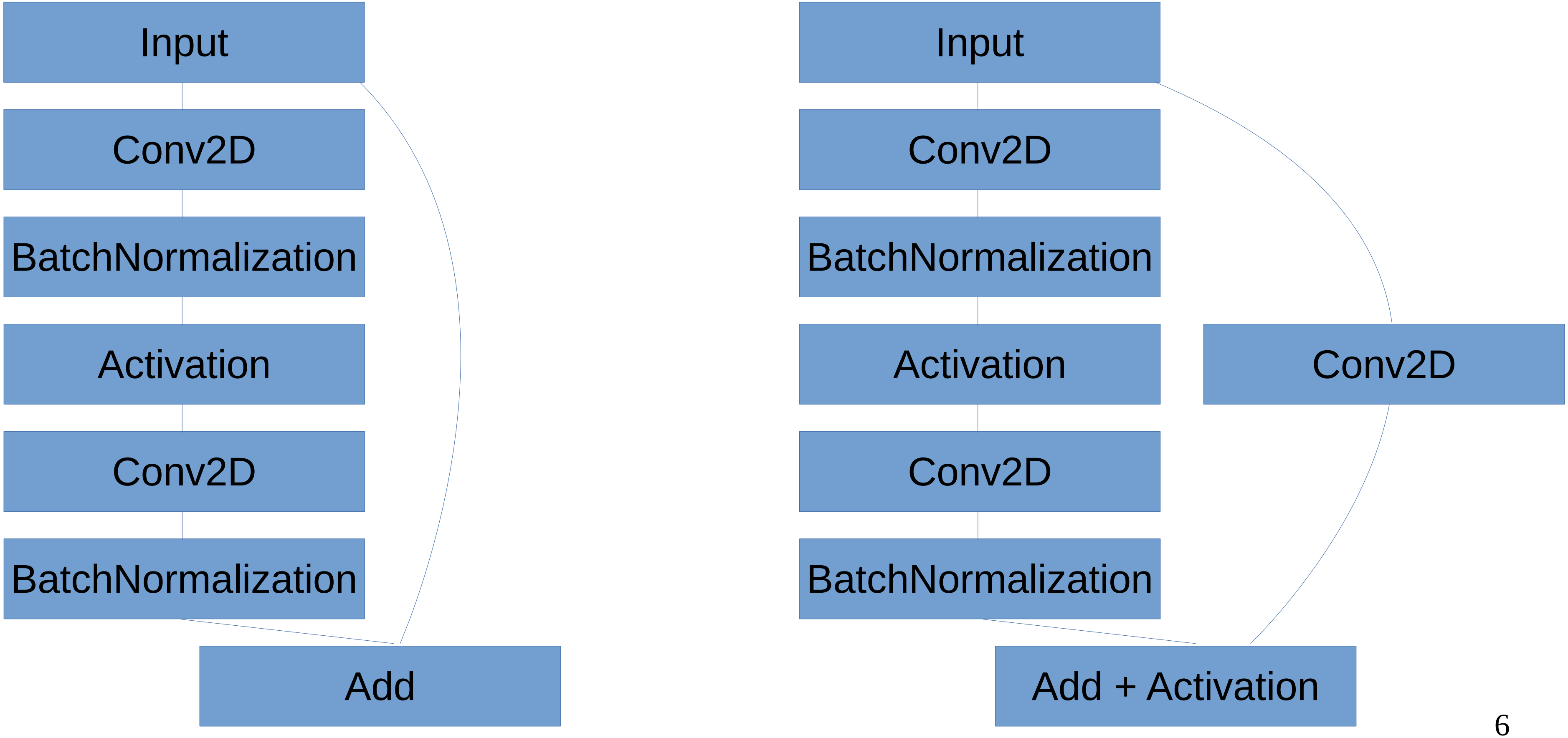
Tested neural networks : sequential CNN, inception CNN



Inception block architecture (arXiv:1512.00567)

Photometric redshifts calibration

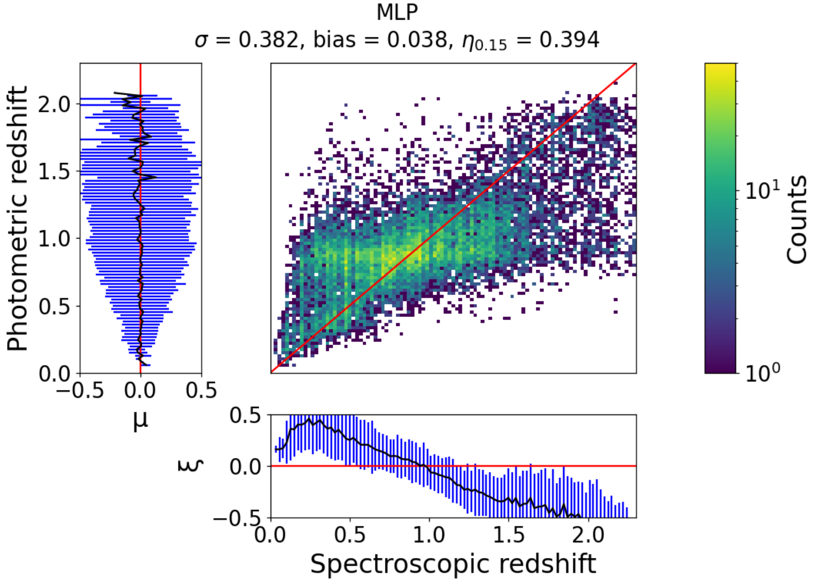
Tested neural networks : sequential CNN, inception CNN, ResNet34

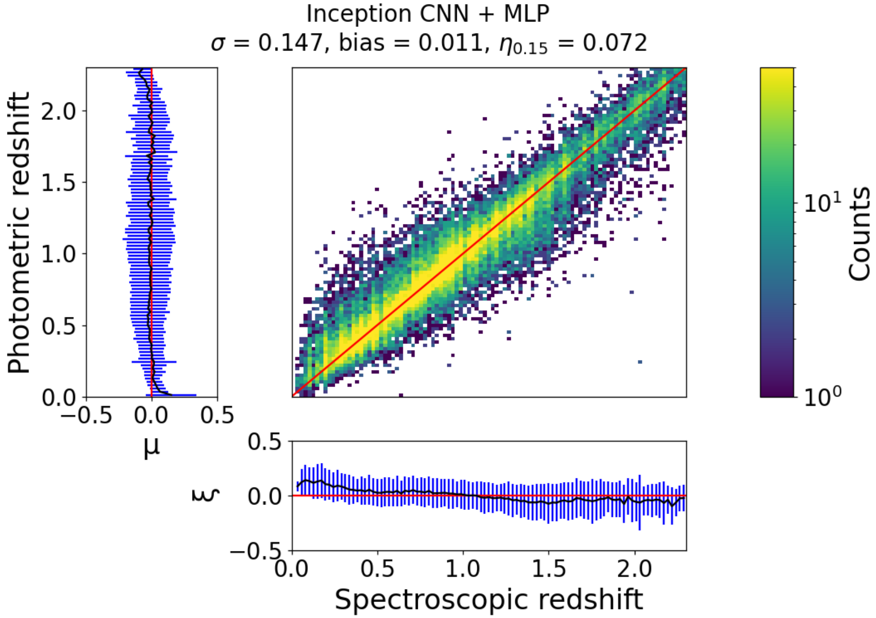


Photometric redshifts calibration

- Goal : find the relationship between the input galaxy images and their redshift.

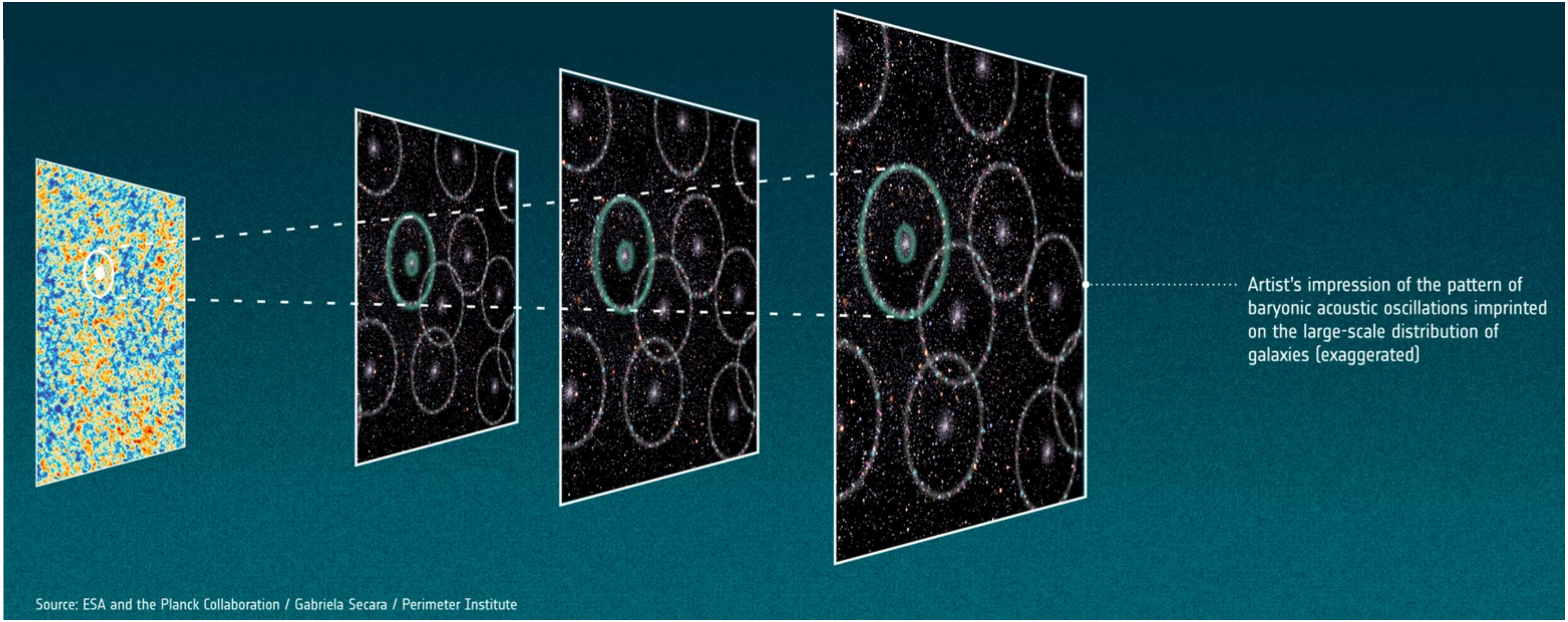
- Method : neural networks + optimization framework (Optuna)
- Networks : CNNs, inception CNNs, ResNet and variants with additional inputs.
- Comparison for MLP (magnitudes) and inception CNN+MLP (images + magnitudes) :





Baryonic Acoustic Oscillations

- Specific scale where galaxies are more often found
- Result of the opposition between radiation pressure and the gravitational pull of matter in the early Universe



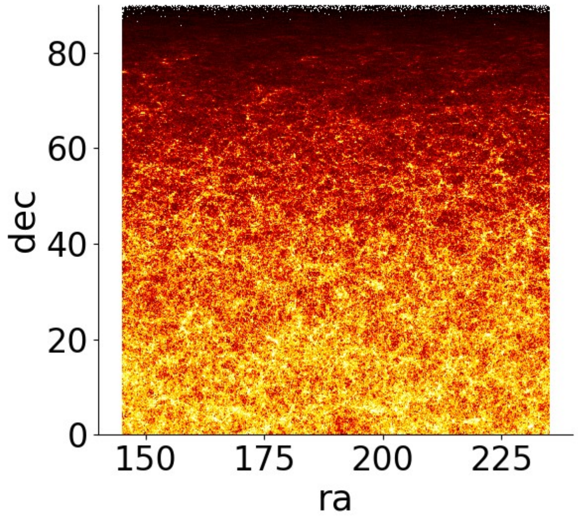
2pcf measurement

- Landy-Szalay estimator $w(\theta) = \frac{DD - 2DR + RR}{RR}$
- Code : TreeCorr
- Flagship simulation :

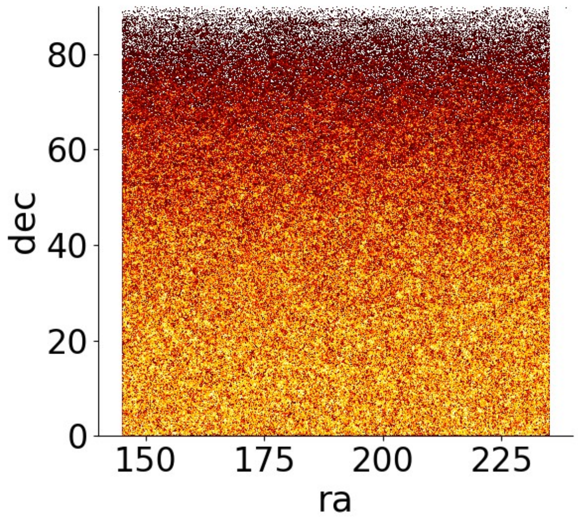
- one octant of the sky (5157 sq.deg)

- 500×10^6 galaxies with VIS < 24.5

- fiducial cosmology : $\Omega_b = 0.049$
$\Omega_c = 0.27$
$h = 0.67$
$A_s = 2.1\times10^9$
$n_s = 0.96$



0.200 < z < 0.303



1.677 < z < 2.500

- 13 bins between 0.2 < z < 2.5

2pcf measurement

- Landy-Szalay estimator $w(\theta) = \frac{DD - 2DR + RR}{RR}$

- Code : TreeCorr

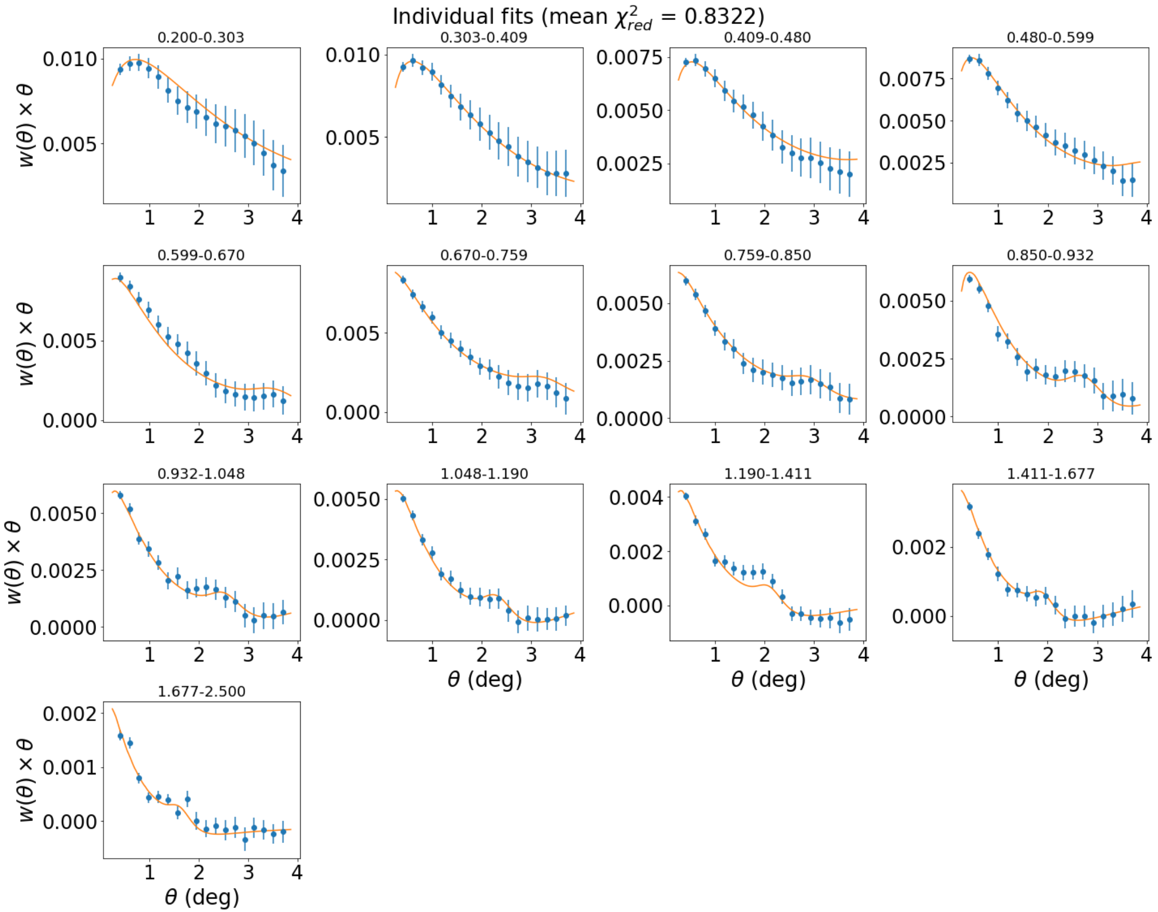
- Flagship simulation :

- one octant of the sky (5157 sq.deg)

- 500 × 10^6 galaxies with VIS < 24.5

- fiducial cosmology : $\Omega_b = 0.049$
$\Omega_c = 0.27$
$h = 0.67$
$A_s = 2.1\times10^9$
$n_s = 0.96$

- 13 bins between 0.2 < z < 2.5



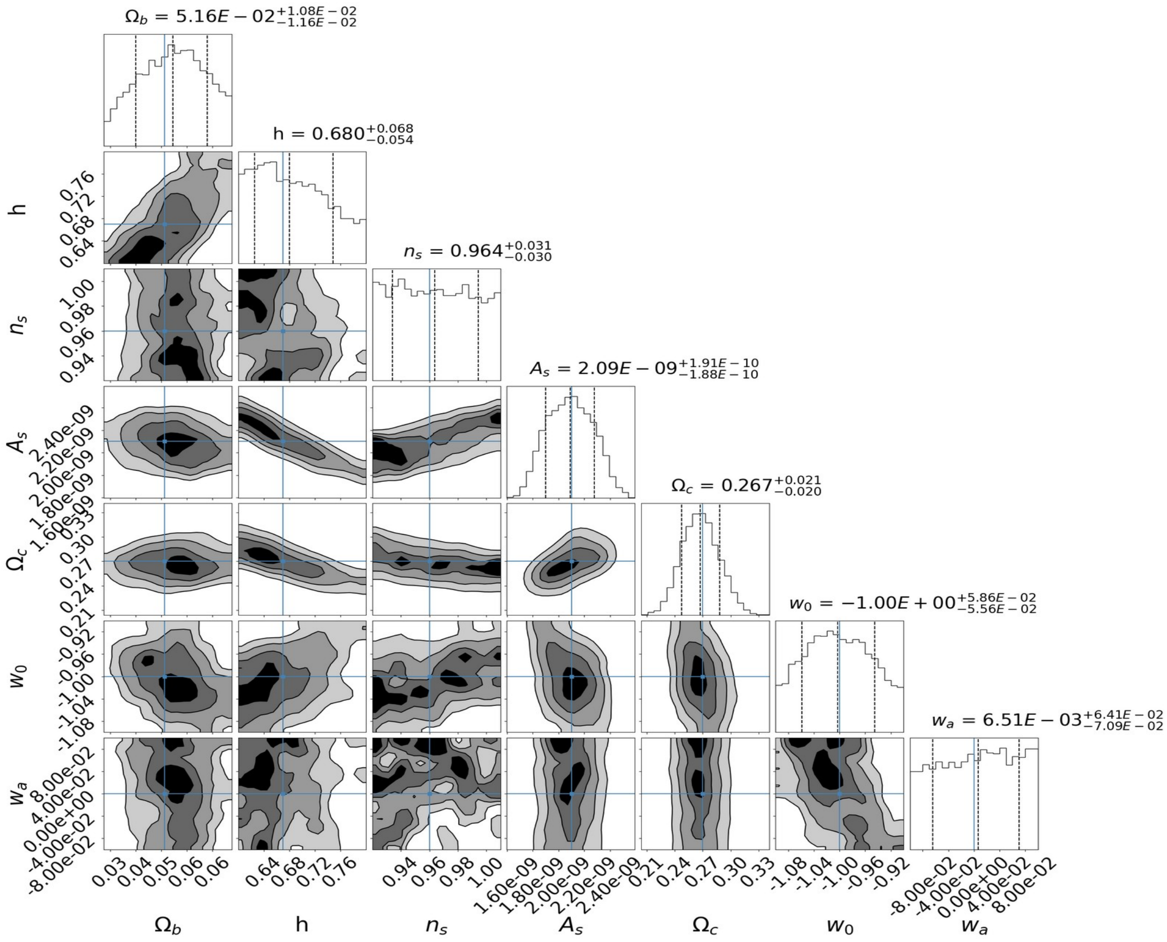
Full-shape analysis

- Scale cuts : 0.12° < θ < 1.7°
- Dark energy equation of state :

$$w(z) = w_0 + w_a \frac{z}{1+z}$$

- MCMC on Ω_b, Ω_c, h, A_s, n_s, w_0 and w_a

At each step of the MCMC, a new 2-point correlation function is computed using the cosmology defined by these parameters



Theoretical 2pcf

Computed with the Core Cosmology Library (arXiv:1812.05995)

$$\xi^{ab}(\theta) = \sum_l \frac{2l+1}{4\pi} C_l^{ab} P_l(cos\theta)$$

$$C_l^{ab} = 4\pi \int_0^\infty \frac{dk}{k} \mathcal{P}_\Phi(k) \Delta_l^a(k) \Delta_l^b(k)$$

$$\Delta_l^D(k) = \int dz\, \boldsymbol{n_z(z)} b(z) T_\delta(k,z) j_l(k\chi(z))$$

n(z) : normalized distribution of sources in redshift

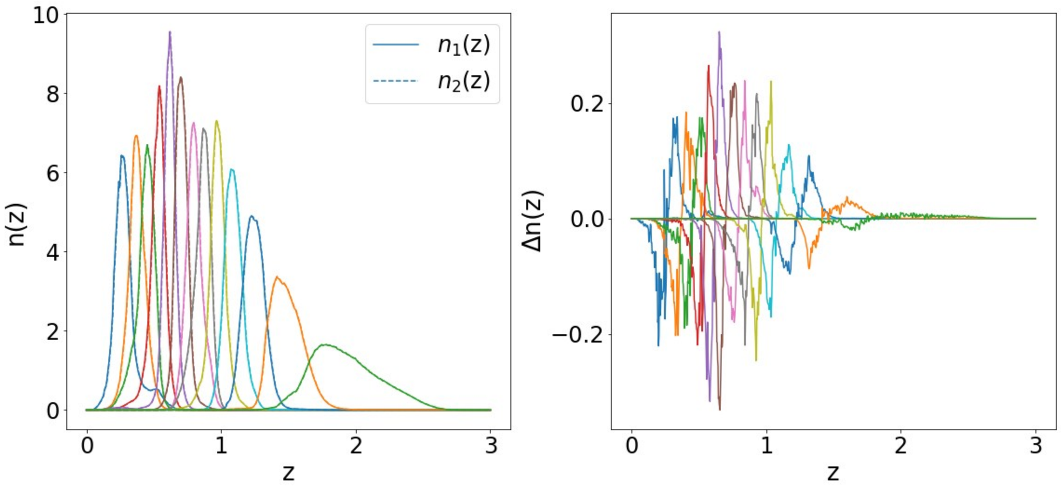
→ used to compute the model in the likelihood.

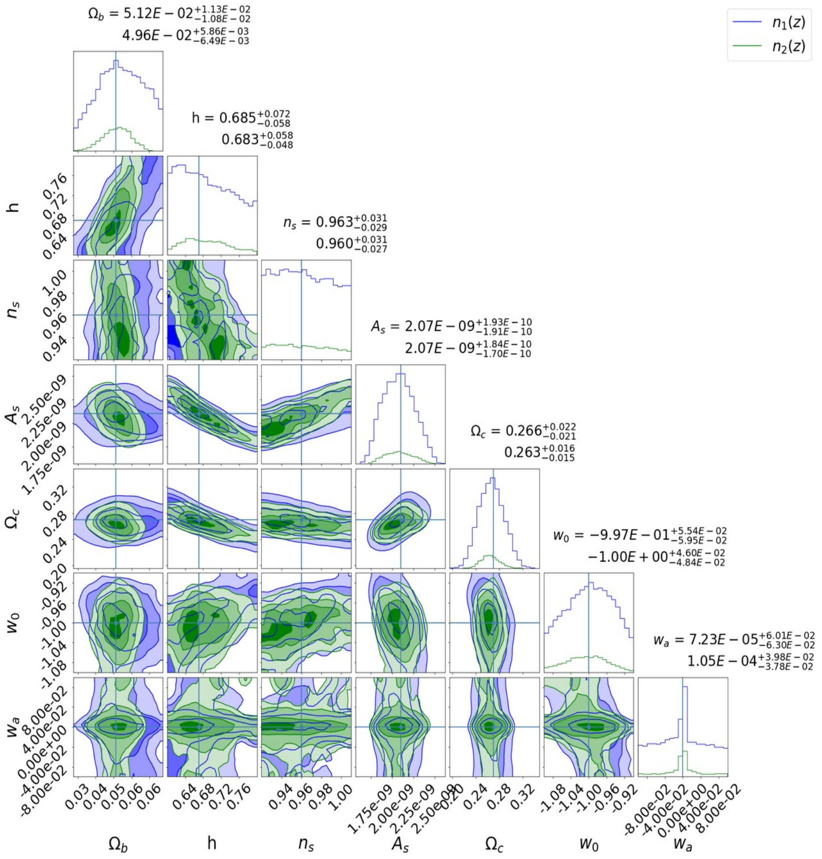
Full-shape analysis with modified n(z) :

Small scales only : θ < 1.7 °

Bias of n(z) :

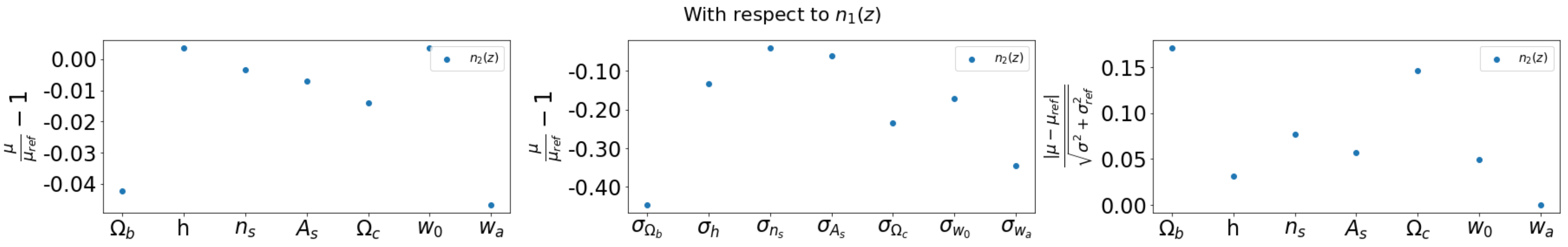
$n_2(z) = n_1(z + 0.002)$





Shift < 0.2 σ

With respect to $n_1(z)$



BAO analysis

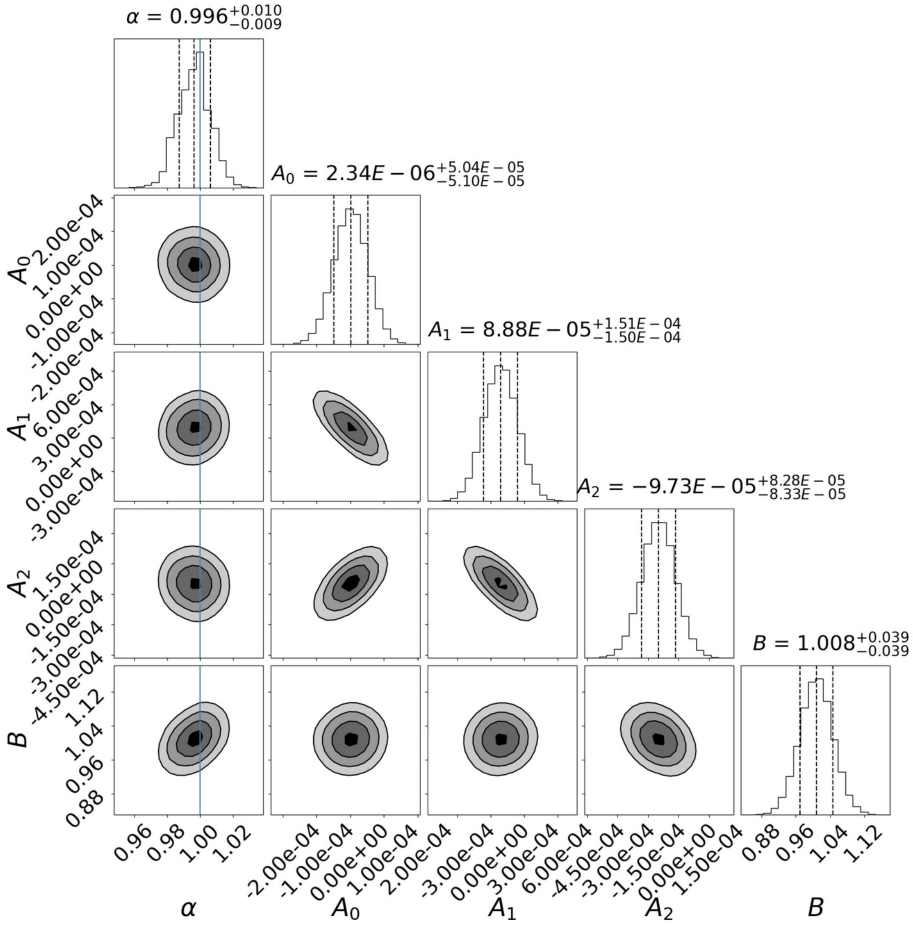
- No restriction to small scales since we're interested in the BAO peak (≠ full-shape).

- Template : $B \times w(\alpha\theta) + A_0 + \frac{A_1}{\theta} + \frac{A_2}{\theta^2}$

- The cosmological parameters are fixed to the fiducial cosmology (≠ full-shape)

α quantifies an eventual shift of the BAO peak in the data with respect to the fiducial cosmology. Since the 2pcf is measured on Flagship, we expect α = 1.

B is a nuisance parameters accounting for corrections of the amplitude.



BAO analysis

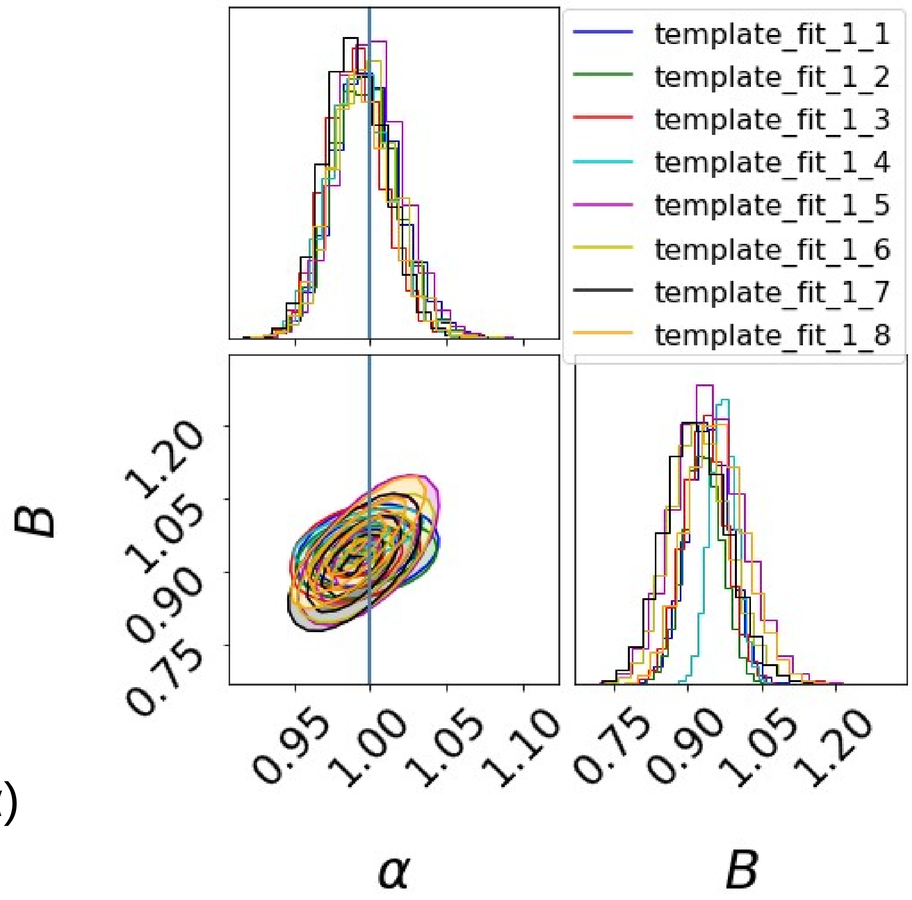
- No restriction to small scales since we want to extract the BAO peak (≠ full-shape).

- Several templates were tested to identify the one providing the best constraints :

Templates 1-4 :

$$B \times w(\alpha\theta) + A_0 + \frac{A_1}{\theta} + \frac{A_2}{\theta^2}$$
$$B \times w(\alpha\theta) + A_0 + A_1\theta + \frac{A_2}{\theta}$$
$$B \times w(\alpha\theta) + A_0 + A_1\theta + \frac{A_2}{\theta^2}$$
$$B \times w(\alpha\theta) + A_0 + A_1\theta + A_2\theta^2$$

Templates 5-8 : $B \rightarrow \frac{B}{\alpha^2}$



- Comparison templates 3 and 1 : $\sigma_3(\alpha) = 0.88\ \sigma_1(\alpha)$

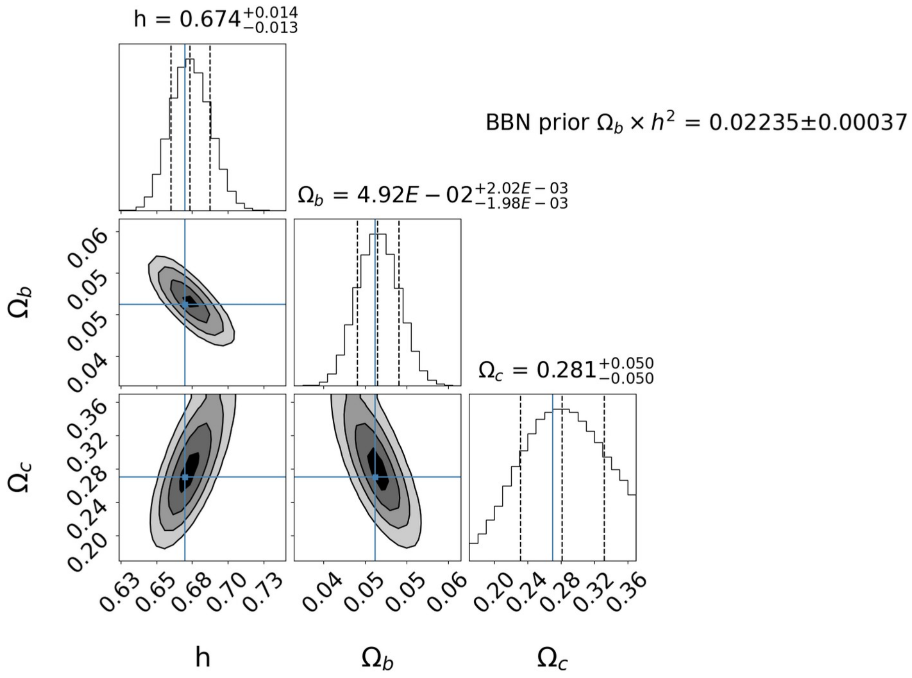
BAO analysis

- Extracting α in each redshift bin allows us to constrain the Hubble parameter $h = H_0/100$

$$\alpha = \frac{d_A}{r_{drag}} \frac{r_{drag,fid}}{d_{A,fid}}$$

$$d_A = \begin{cases} \frac{sin(H_0\sqrt{|\Omega_k|r})}{(1+z)H_0\sqrt{|\Omega_k|}} & \text{if } \Omega_k < 0 \\ \frac{r}{1+z} & \text{if } \Omega_k = 0 \\ \frac{sinh(H_0\sqrt{|\Omega_k|r})}{(1+z)H_0\sqrt{|\Omega_k|}} & \text{if } \Omega_k > 0 \end{cases}$$

$$\text{with } r = \int_0^z \frac{c}{\sqrt{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}}$$



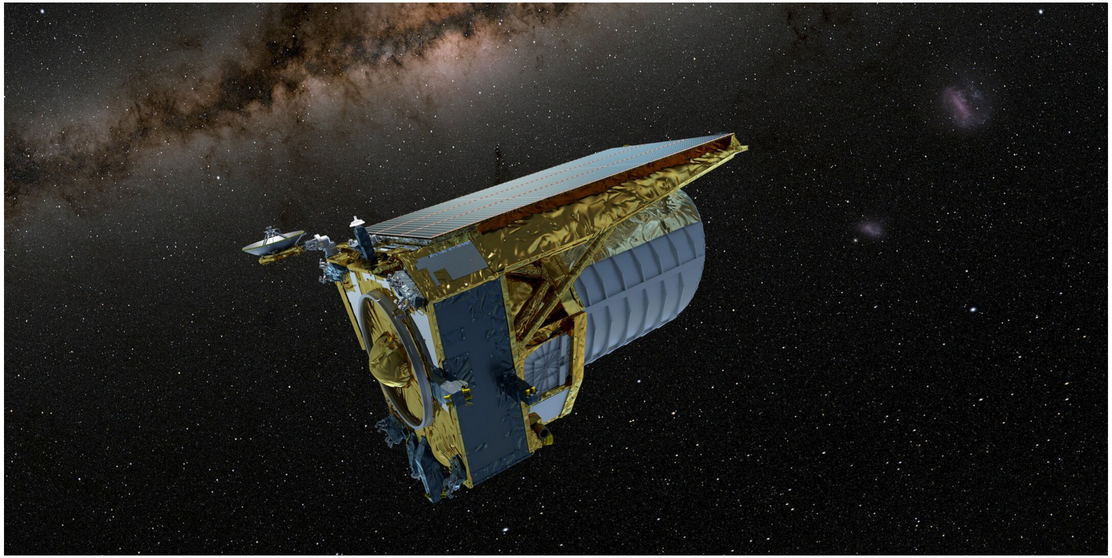
Conclusion

- Promising results are obtained for photo-zs with the inception CNN + MLP model.

→ additional optimizations will be conducted to study the benefit from including other inputs.

- The pipeline for full-shape and BAO analyses with photometric galaxy clustering is ready and will be used to check the influence of scale cuts, priors and other systematic effects.

- Next year : application of this work to the first Euclid data



Thank you for your attention !

Questions ?

Early commissioning test image, VIS instrument

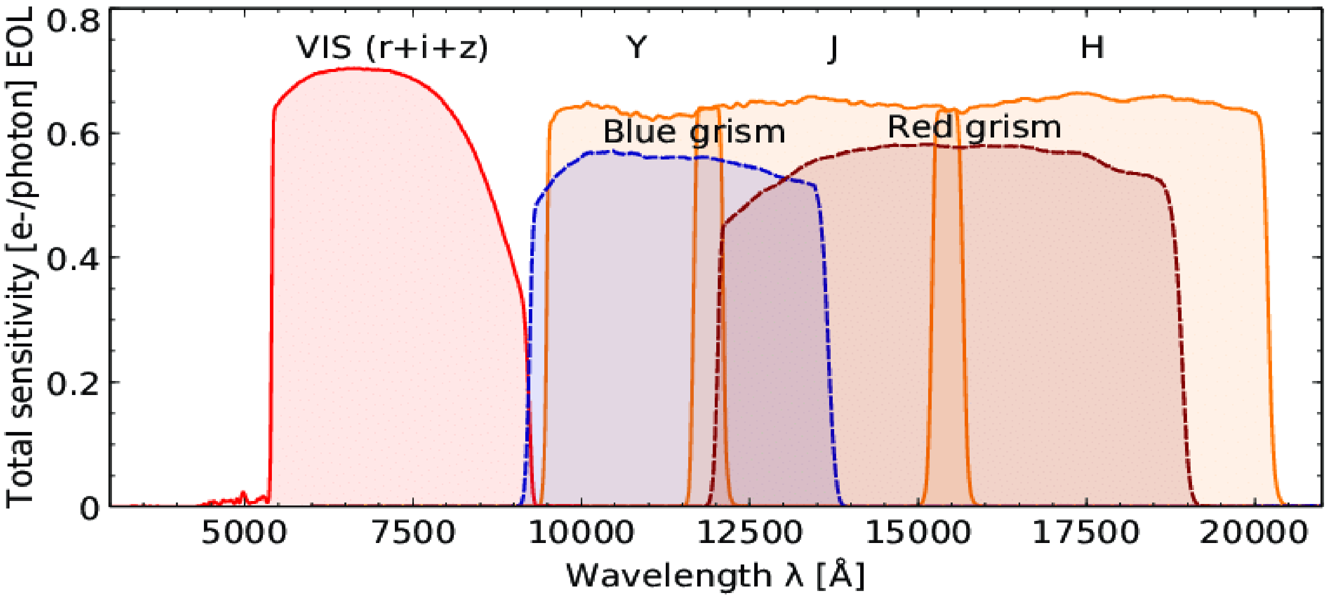
Early commissioning test image, NISP instrument

Back-up

Photometric redshifts

Euclid bands :

VIS 550-900 nm
Y 920-1146 nm
J 1146-1372 nm
H 1372-2000 nm



Euclid preparation: I. The Euclid Wide Survey (arXiv:2108.01201)

Photometric redshifts

Loss function used to train : mean squared error

Metrics :

- Standard deviation of residuals σ = std(Δz) wtih $\Delta z = z_{phot} - z_{spec}$

- Bias : mean(| Δz | / (1 + z_{spec}))

- Outlier fraction at 15 % : #(bias > 0.15) / #(test set)
  \+ fractions at 10 % and 5 %

- σ_{NMAD} = 1.4826 × median(| Δz | - median(Δz))

- σ_{MAD} = 1.48 × median(| Δz |)

Photometric redshifts

Side plots :

Learning error $\xi = p(z_{phot} - z_{spec} \mid z_{spec})$
→ in each bin of the histogram, I compute the mean and standard deviation of the $z_{predicted,i} - z_{bin}$ for all $z_{spec,i}$ falling into that bin

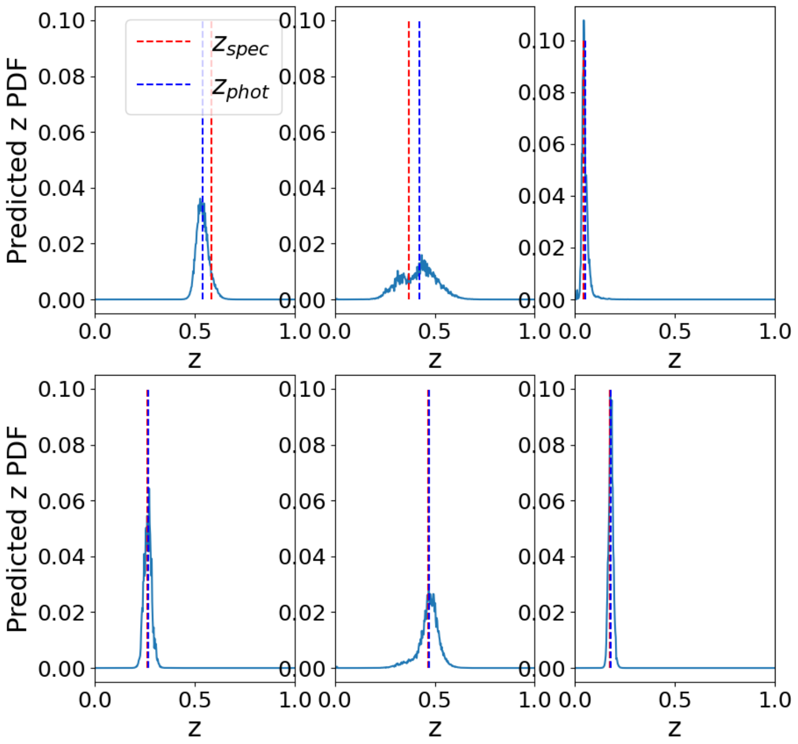
Prediction uncertainty $\mu = p(z_{phot} - z_{spec} \mid z_{phot})$
→ in each bin of the histogram, I compute the mean and standard deviation of the $z_{spec,i} - z_{bin}$ for all $z_{predicted,i}$ falling into that bin

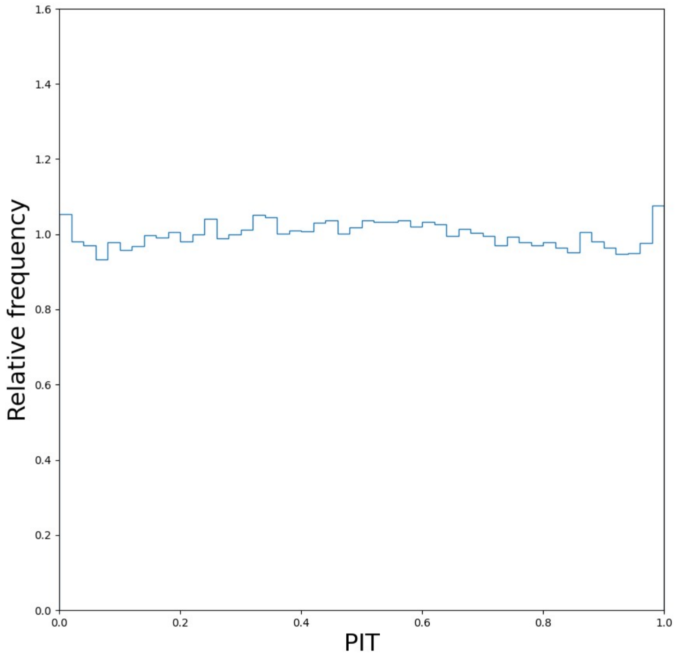
Additional statistics on ξ and μ :

Avg % error	Min % error	Max % error	Avg % error without high z bins	Min % error without	Max % error without
16.05	5.26	90.56	19.23	10.97	90.56

Photometric redshifts

Example of PDFs produced after adaptation of the networks :





PIT distribution of the PDFs

Photometric redshifts

Characterization of the PDFs :

Probability Integral Transform (PIT), for a galaxy i of redshift $z_{spec} = z_i$

$$CDF_i(z_i) = \int_0^{z_i} PDF_i(z)\, dz$$

If PDFs are often too narrow then the z_{spec} will more often be under/overestimated and the PIT value will be close to 0 or 1.
If they are too wide then z_{spec} will often be in the PDF, which favors intermediate PIT values

→ study of the PITs distribution :
- if PDFs have inadequate shapes then the distribution will either be concave or convex.
- if there is a bias between the predicted redshifts and z_{spec} then it creates a slope

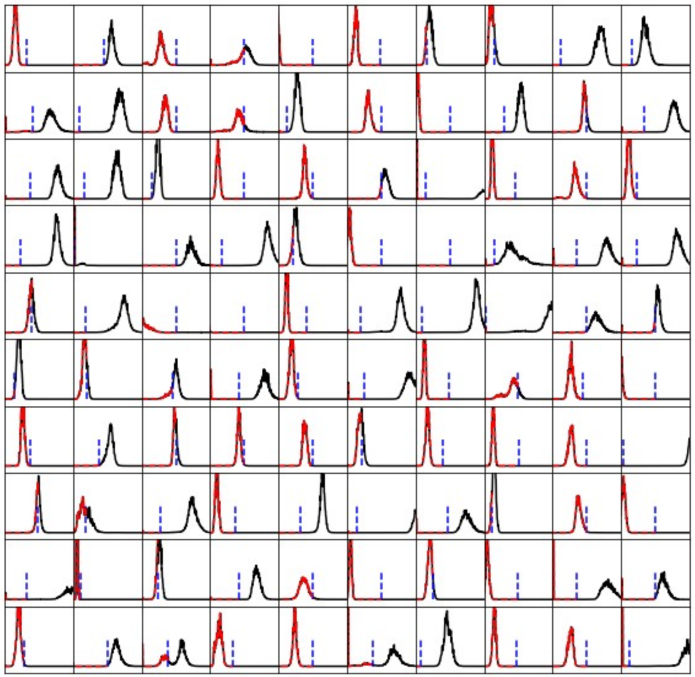
→ an ideal PIT distribution is horizontal and has no curvature.

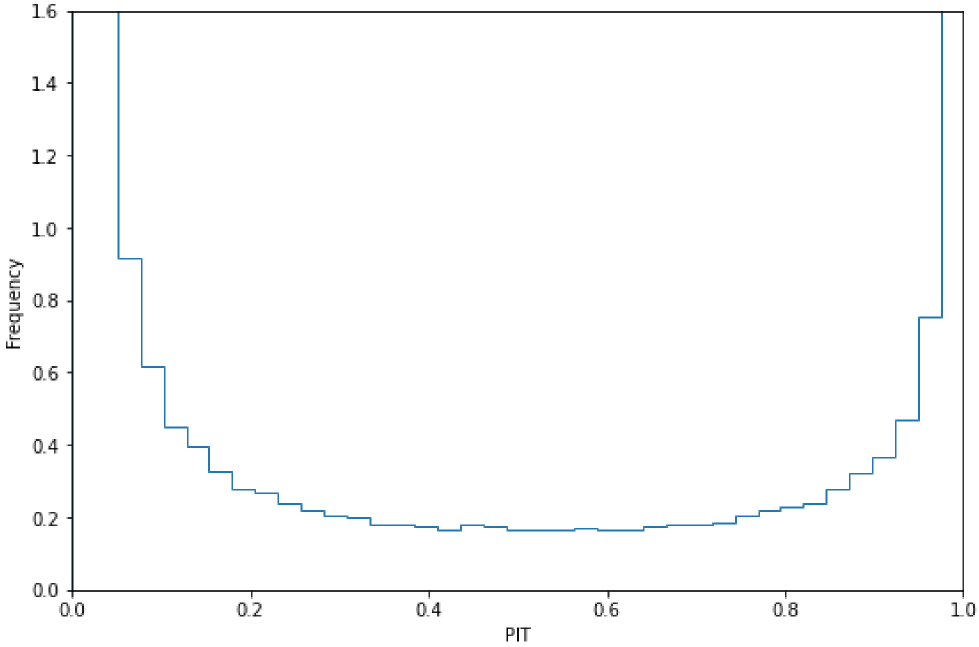
Photometric redshifts

Example of a bad PIT distribution :

Many PDFs miss z_{spec}

The PIT distribution is convex

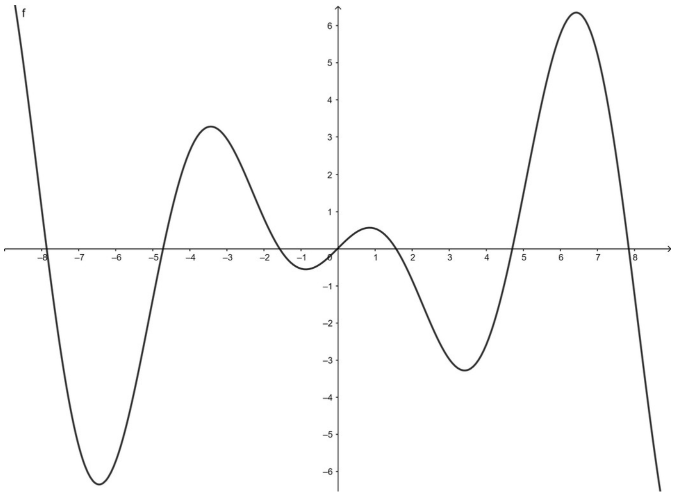




Vanishing gradients

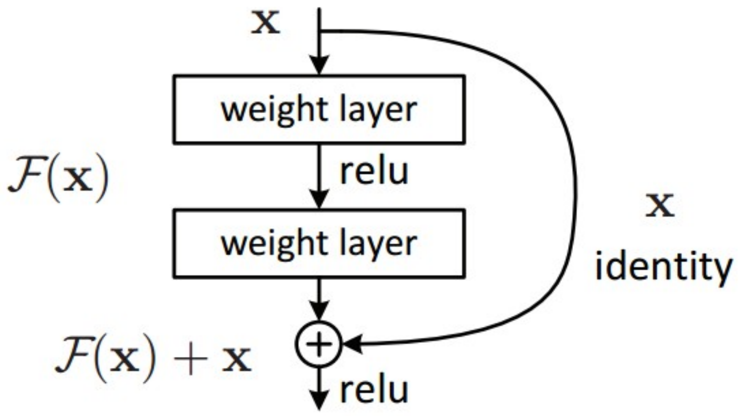
The update of weights is proportional to the gradient of the loss function with respect to current weights. In the backpropagation, the chain rule for partial derivatives is used, which implies that we can end up multiplying very small gradients in chain. This entails the death of some neurons because their weights no longer change.

As for exploding gradients, Rectified activation functions like ReLu limit this issue because they can only saturate by negative values but the issue can still appear. Some oscillating functions can be used to counter this problem
like the Growing Cosine Unit



https://commons.wikimedia.org/wiki/File:Growing_Cosine_Unit_(GCU)_activation_function.png

Residual blocks



https://arxiv.org/abs/1512.03385

The layer n give its output to layer n+1 and layer n+5 (in ResNet34) or n+3,… depending on the architecture

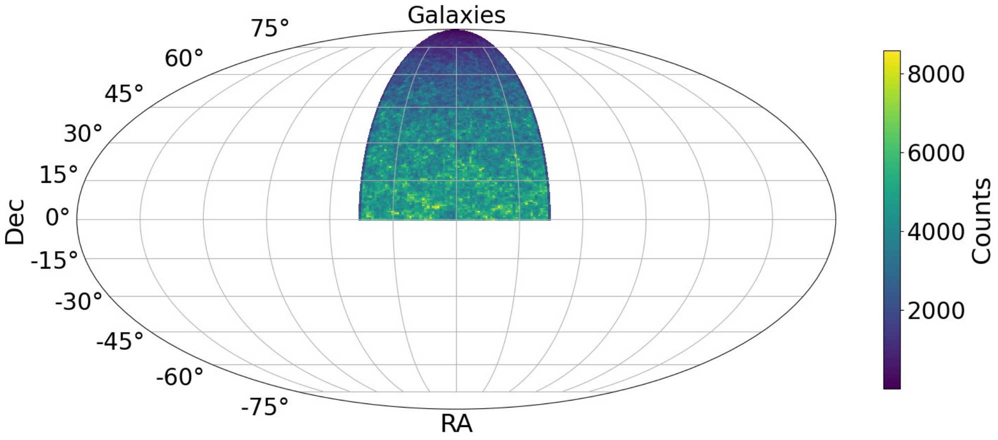
Benefit : when the number of layers is increased in a neural network, results improve before reaching a maximum and then degrade (vanishing gradients).

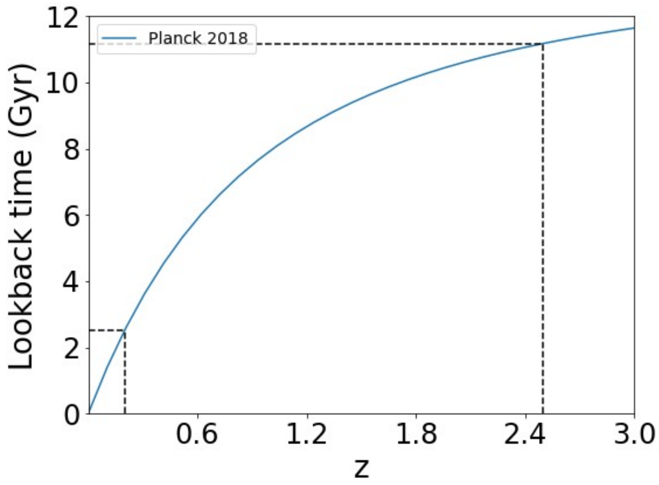
Idea :
residual = output – input ↔ output = residual + input

This enables the identity operation when the residual is fixed to 0. This is useful since the identity can’t be the output of a neural network if there is no skip connection (non linear activation functions) → the least useful layers have weights close to 0 but won’t make gradients vanish because the skip connection will have larger weights.

Flagship 2.1

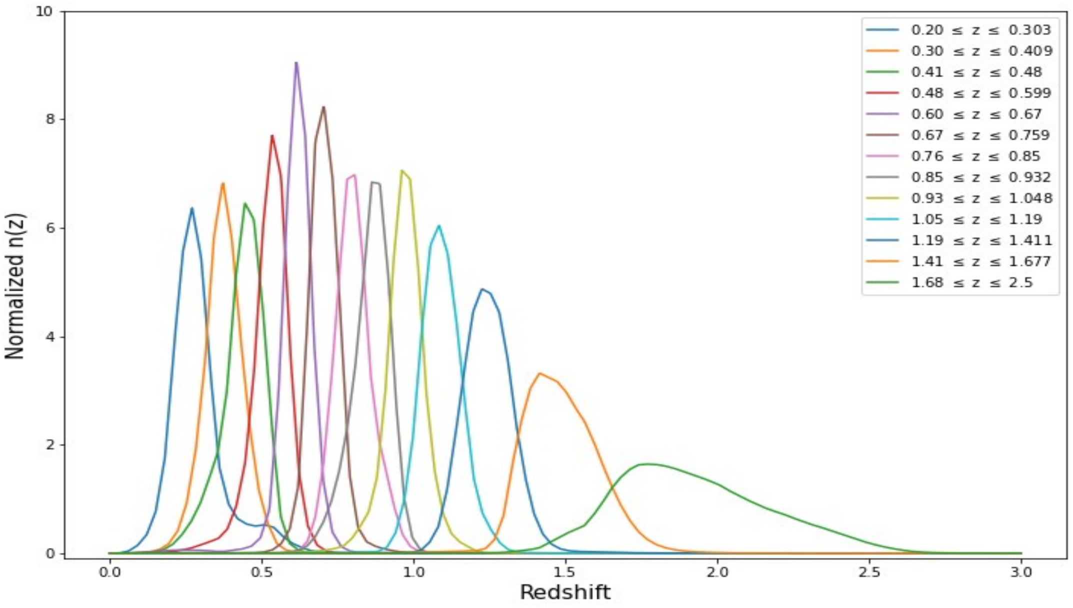
- one octant of the sky, 145 < ra < 235 deg, 0 < dec < 90 deg
- 500 × 10^6 galaxies with VIS < 24.5 and photo-zs.
- fiducial cosmology : $\Omega_b = 0.049$
  $\Omega_c = 0.27$
  $h = 0.67$
  $A_s = 2.1\times10^9$
  $n_s = 0.96$
- 13 bins between 0.2 < z < 2.54



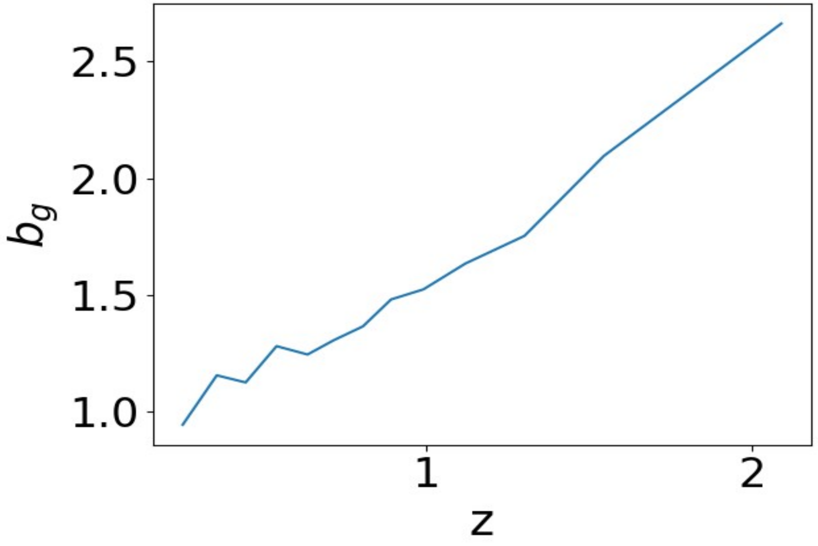


Flagship 2.1

Equipopulated bins n(z) :



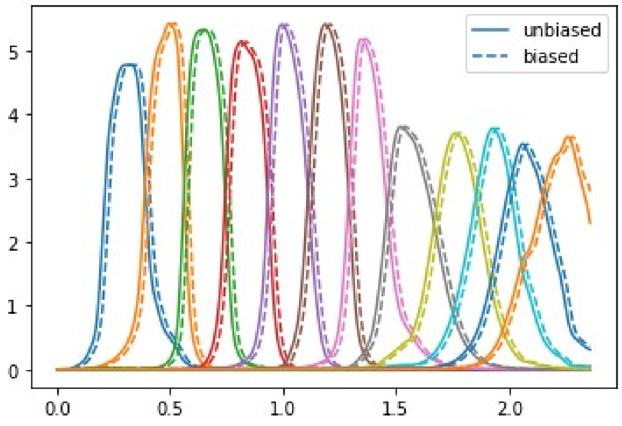
Measured galaxy bias:



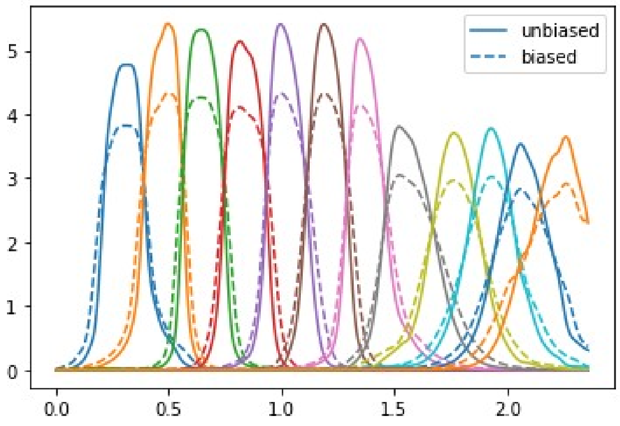
Full-shape analysis with modified n(z)

Goal of GCPHz WP paper 3 : study systematic uncertainties like n(z) model misspecifications

Modifications of n(z) :



Additive bias

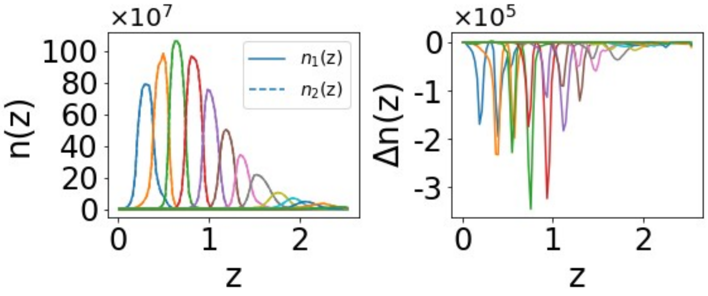


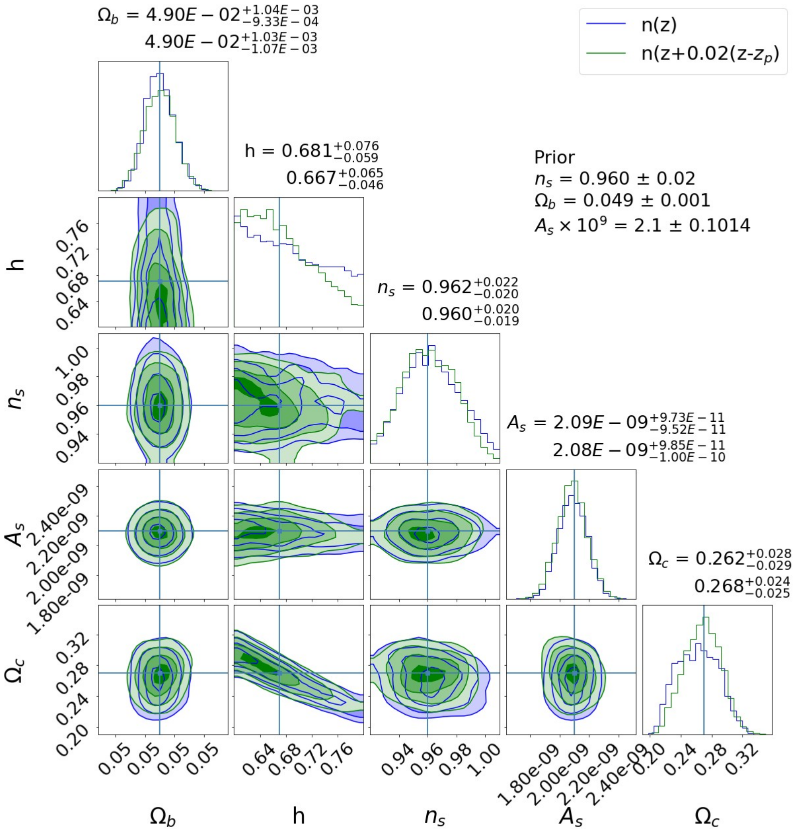
Broadening

Full-shape analysis with modified n(z)

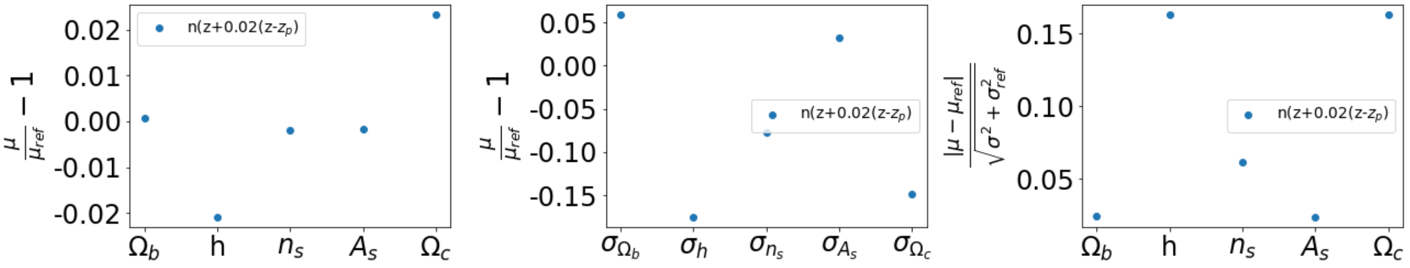
Broadening of n(z) :

$$n_2(z) = n_1(z + 0.02(z - z_p))$$





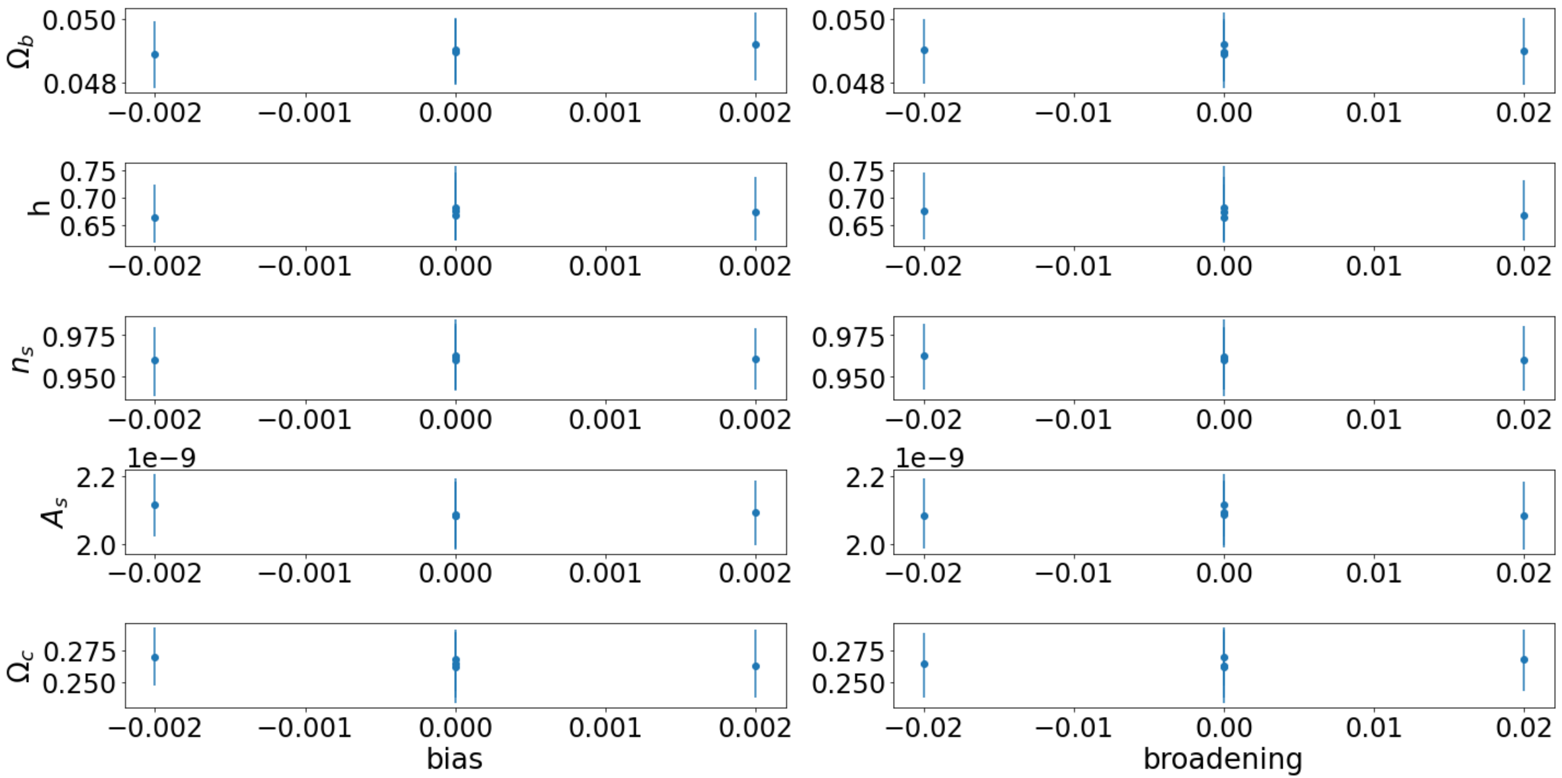
With respect to n(z)



Shift of **0.15 σ** on h and Ω_c

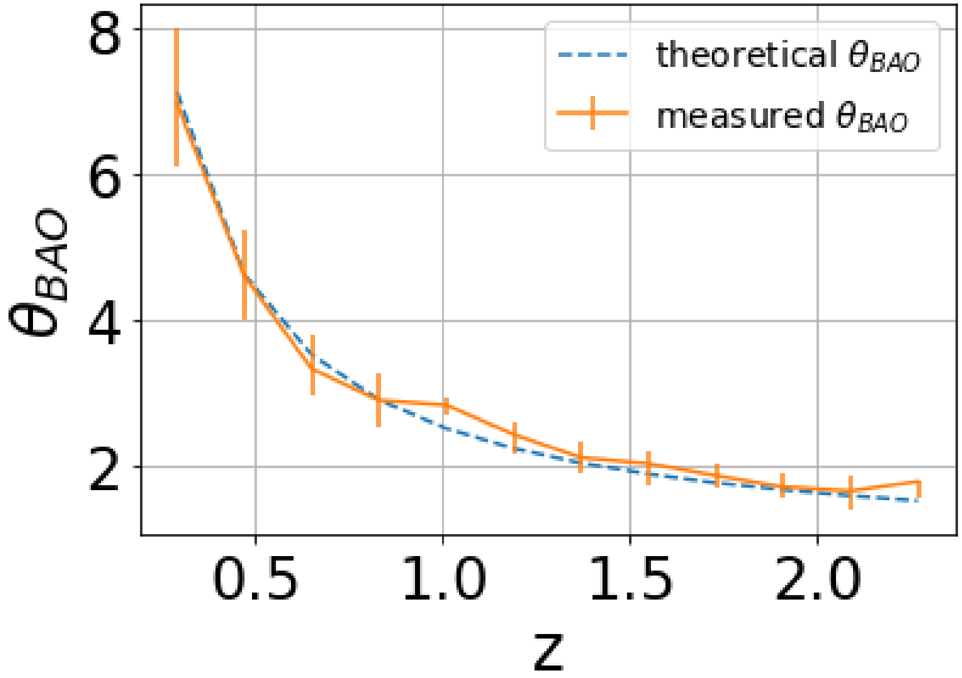
Full-shape analysis with modified n(z)

Influence of n(z) model misspecifications



BAO analysis

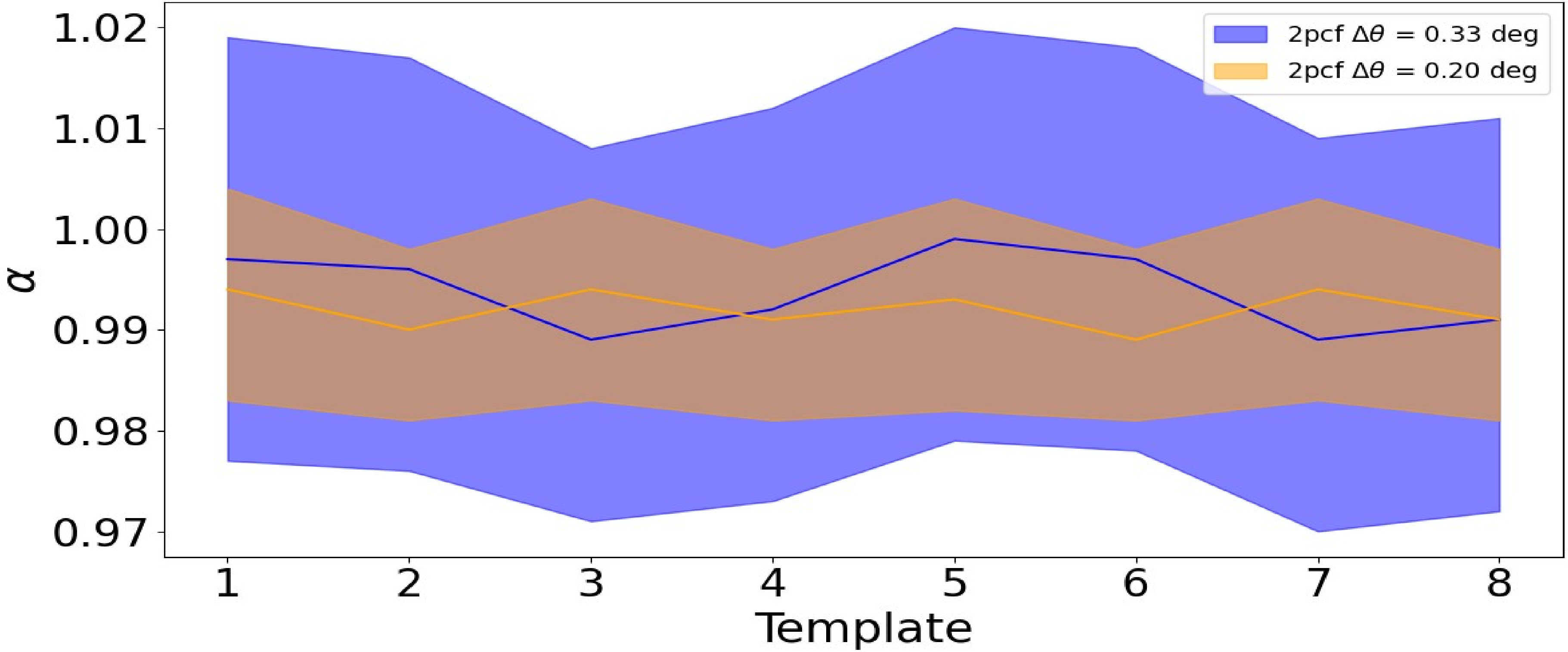
BAO extracted from the 2pcf measured on Flagship, in each bin of redshift



θ_{BAO} and its error are obtained by MCMC.

BAO analysis

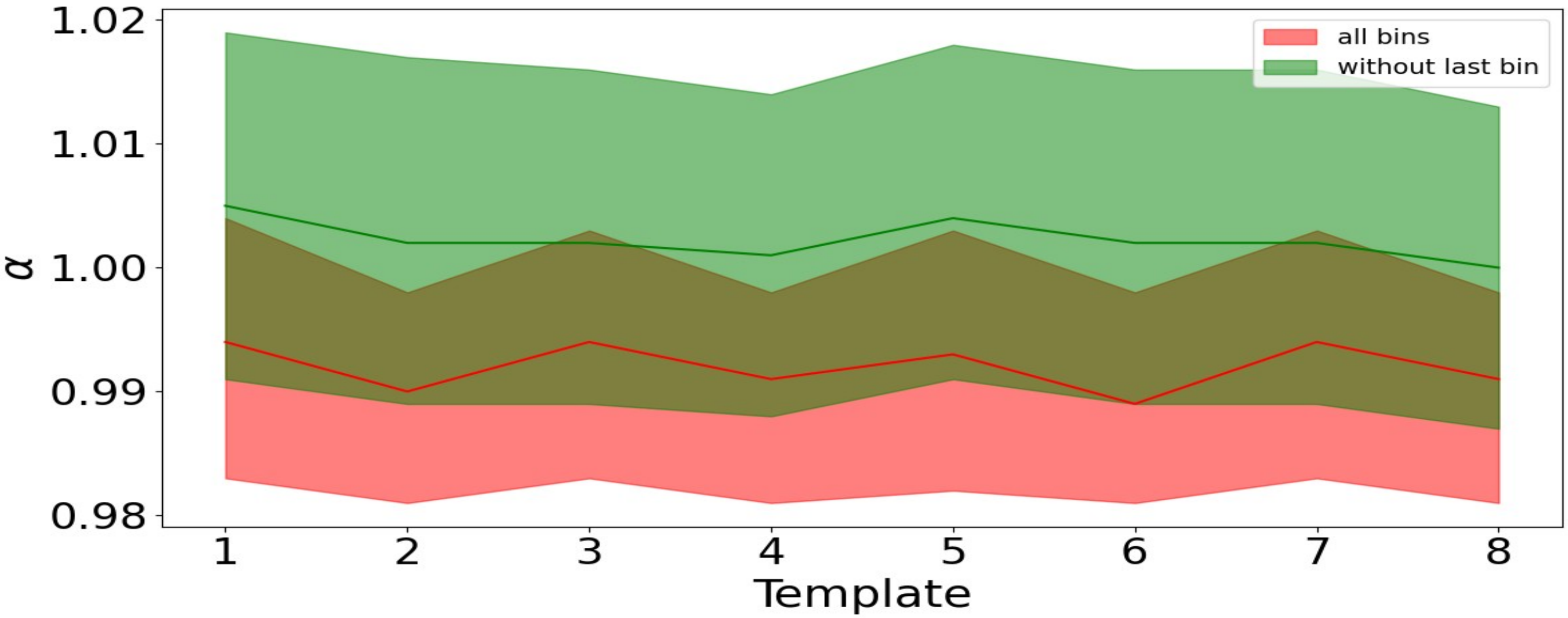
MCMC with the previous measurement (left) and the new one (right) :



The error on α is divided by 2 with the new measurement.

BAO analysis

Comparison including or excluding the last redshift bin :



In agreement at 1σ but there is an obvious systematic shift towards larger α and errors.
The robustness of the results with respect to the redshift bins used should be checked.