Likelihood-free large-scale structure inference with robustness to model misspecification

Cosmology & Gravitation seminar,
Stockholm University

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

In collaboration with:
Wolfgang Enzi (MPA), Alan Heavens (Imperial College), Tristan Hoellinger (IAP), Jens Jasche (Stockholm University)

and the Aquila Consortium

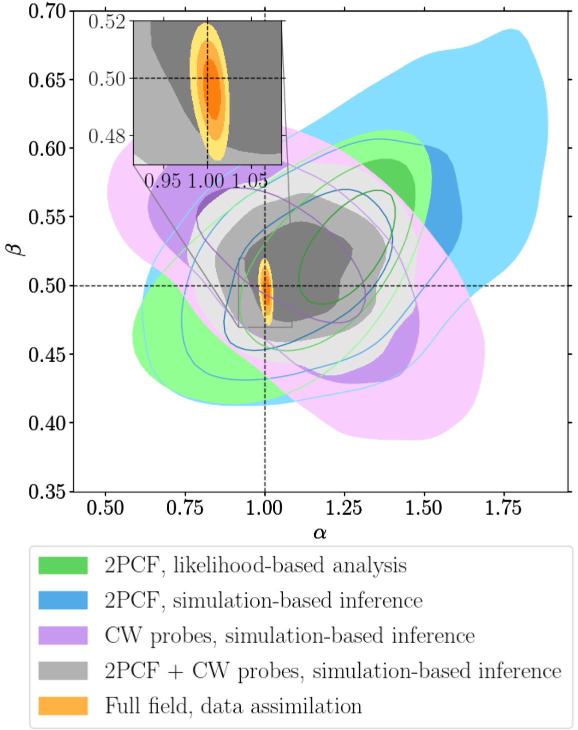
www.aquila-consortium.org

15 June 2023

Why I decided to go “likelihood-free” for galaxy clustering additional probes

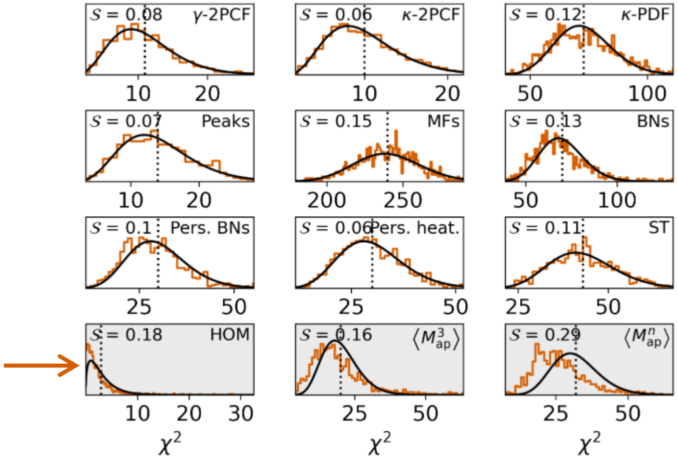
Note: likelihood-free inference ≈ simulation-based inference ≈ implicit likelihood inference

- A question of accuracy: first, avoid biases.



FL & Heavens, 2103.04158

- Some WL additional probes also have a non-Gaussian distribution.



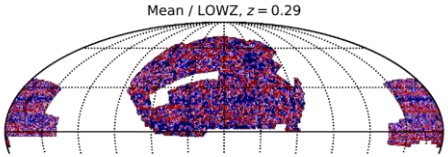
- A question of precision: can numerical forward models be used to push further than $k \gtrsim 0.15\ h/\mathrm{Mpc}$? The full field contains much more information.

Euclid HOWLS-KP paper 1, Ajani *et al.*, 2301.12890

The issue of model misspecification in Bayesian inference and in simulation-based inference (SBI)

- Model misspecification arises when model differs from actual data-generating process.
- Field-based inference techniques have a successful track record at handling model misspecification, e.g. automatically reporting unknown data contaminations.

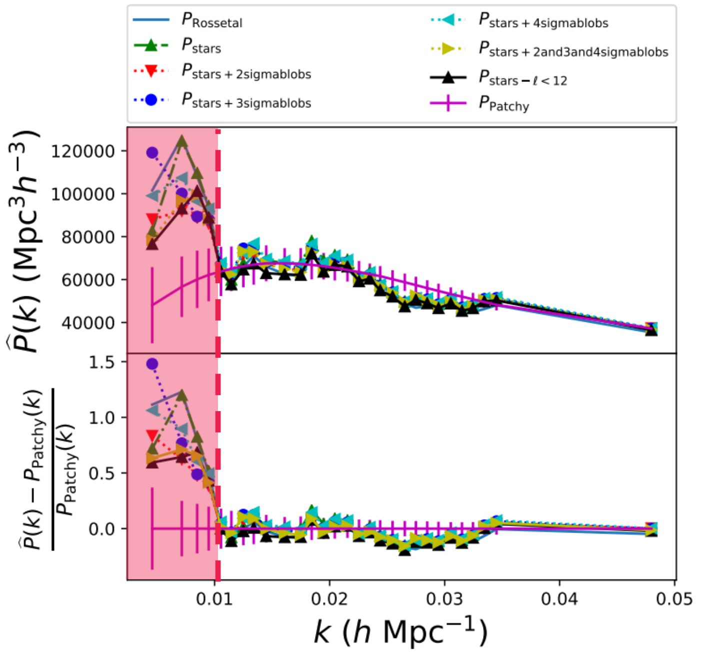
Map of unknown foreground contaminant



- Model misspecification: a major challenge particularly for approaches that marginalise over latent variables, such as simulation-based inference (SBI).

Porqueres, Ramanah, Jasche & Lavaux, 1812.05113
Lavaux, Jasche & FL, 1909.06396

- Typical cosmological example: the galaxy *power spectrum* at large scales.



Kalus, Percival *et al.*, 1806.02789

A general class of Bayesian hierarchical models (BHMs): Complex observations of a latent function controlled by top-level parameters

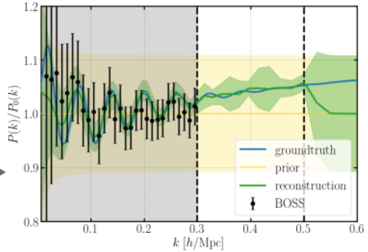
$\mathcal{P}(\omega)$

Example from cosmology:

Target parameters → ω ← Cosmological parameters → $\Omega_\mathrm{m}, \Omega_\mathrm{b}, H_0, n_\mathrm{S}, \sigma_8, \tau$

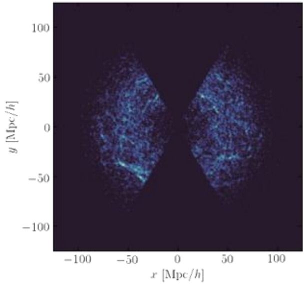
Cheap (deterministic) simulator → $\mathcal{T}$ ← Solver of the Boltzmann equation (CAMB, CLASS)

Latent function → θ ← Matter power spectrum →



Complex probabilistic observational process → $\mathcal{P}(\Phi|\theta)$ ← *N*-body simulation, galaxy formation, survey modelling…

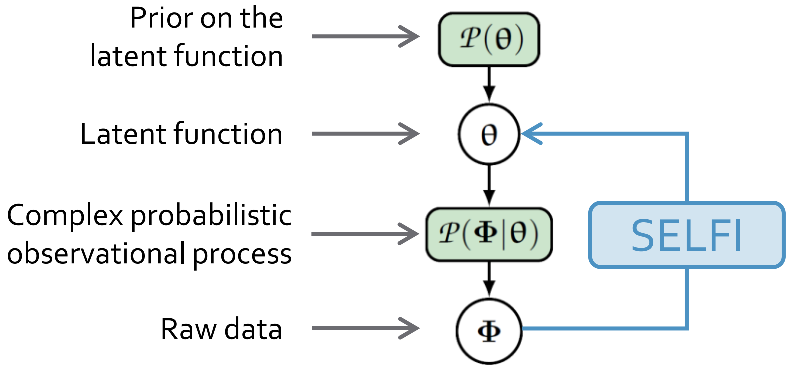
Raw data → Φ ← Galaxy survey data →



Compressor → $\mathcal{C}$ ← Score compression

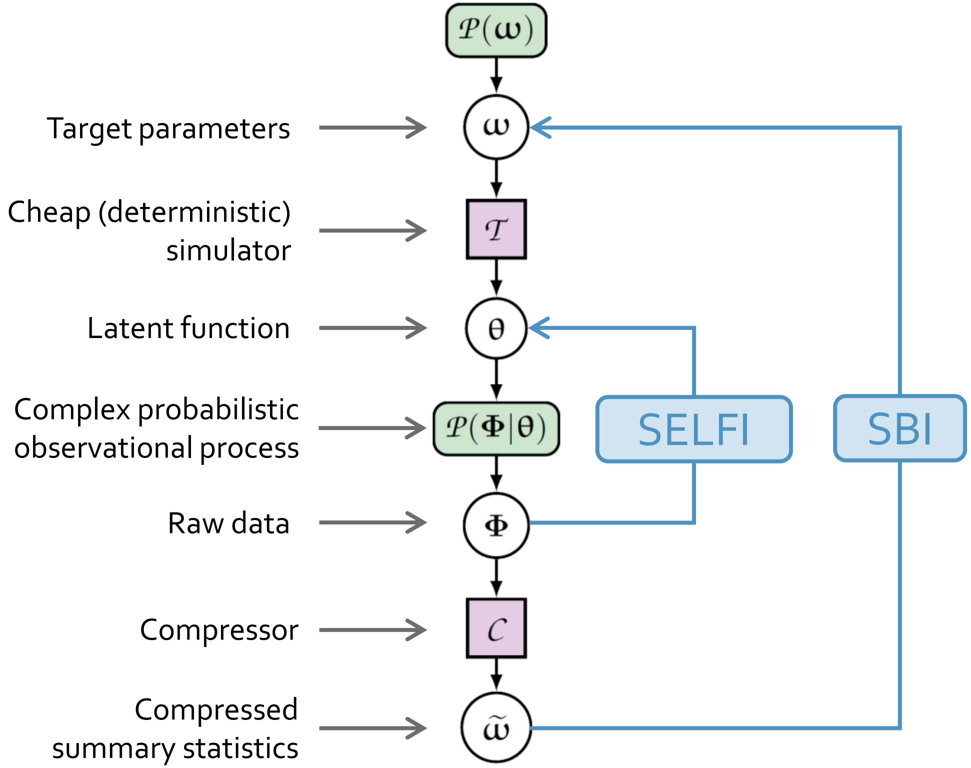
Compressed summary statistics → $\tilde{\omega}$ ← Quasi maximum-likelihood estimators → $\widetilde{\Omega_\mathrm{m}}, \widetilde{\Omega_\mathrm{b}}, \widetilde{H_0}, \widetilde{n_\mathrm{S}}, \widetilde{\sigma_8}, \widetilde{\tau}$

Key idea: a two-step SBI process that recycles simulations



1. Inference of the latent function θ, to check for model misspecification:
   - SELFI algorithm

Key idea: a two-step SBI process that recycles simulations

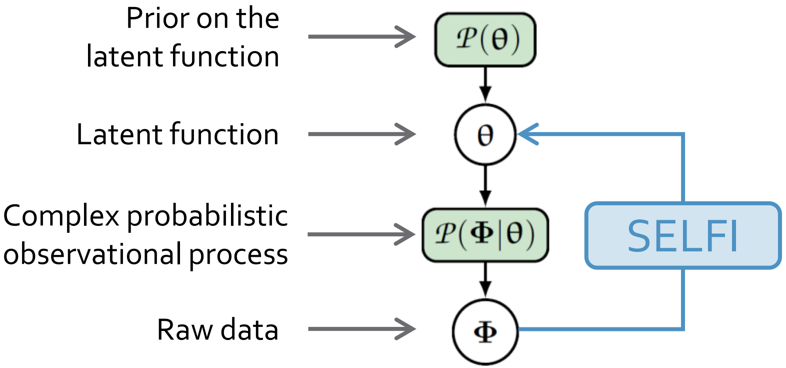


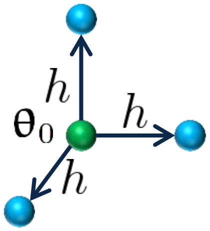
1. Inference of the latent function θ, to check for model misspecification:
   - SELFI algorithm

2. Simulation-based inference of ω:
   - Approximate Bayesian Computation (ABC), Likelihood-Free Rejection Sampling
   - Density/ratio estimation (DELFI / NRE)
   - Bayesian optimisation (BOLFI)
   - others…

Important: the simulations necessary for step 1. are recycled for data compression, which is required for step 2.

Latent function inference: the SELFI approach (*Simulator Expansion for Likelihood-Free Inference*)





- We aim at inferring the latent function θ, which usually contains most/all of the information on ω.

  (initial power spectrum in cosmology, prey/predator population functions in ecology)
- This requires doing SBI in $d = \mathcal{O}(100) - \mathcal{O}(1,000)$
- If we trust the results of earlier experiments, we can Taylor-expand the black-box around an expansion point θ_0 :

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}} \approx \mathbf{f}_0 + \nabla\mathbf{f}_0 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \boxed{\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \cdot \mathbf{H} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)} + \ldots$$

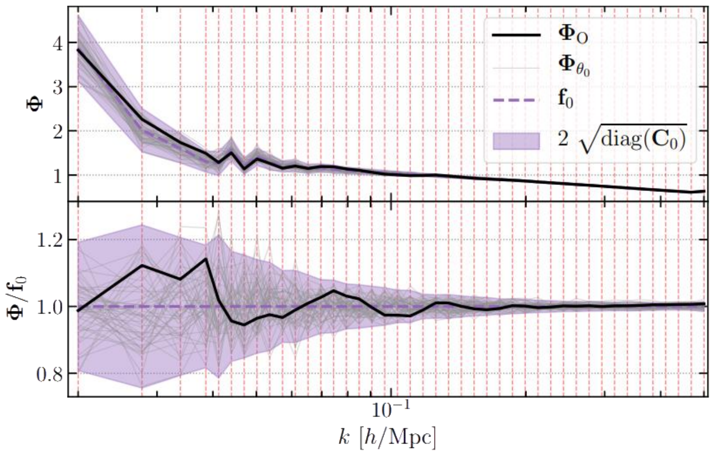
SELFI-2 (second order): coming soon!

- Gradients, Hessian matrix, etc. of the black-box can be evaluated via finite differences in parameter space.

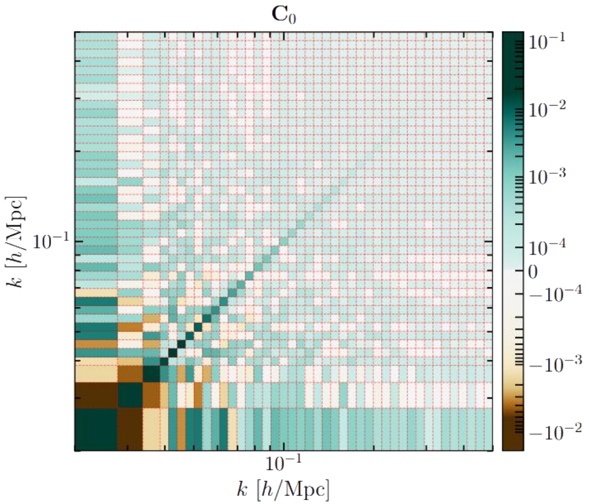
FL, Enzi, Jasche & Heavens, 1902.10149

Galaxy clustering additional probes pipeline: diagnostics of the linearised black-box data model

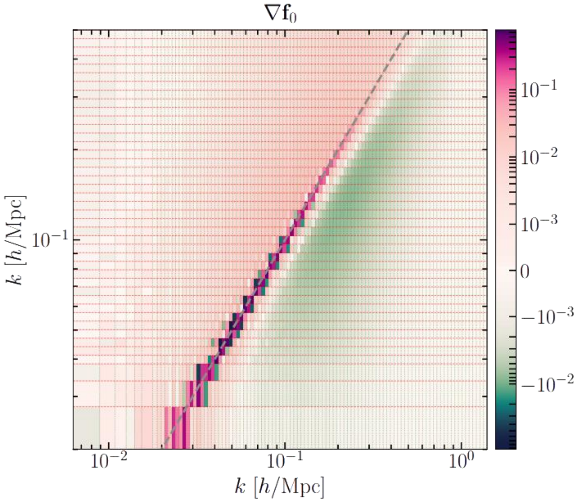
Black-box mean



Black-box covariance



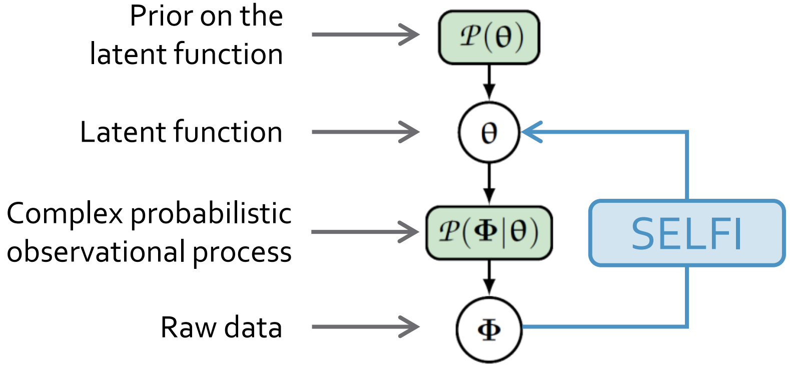
Black-box gradient



- Using only here the (final, non-linearly evolved) power as summary statistics.
- Any additional probe can go in the data vector, since we need the simulations anyway!

FL, Enzi, Jasche & Heavens, 1902.10149

Latent function inference:
the SELFI approach (*Simulator Expansion for Likelihood-Free Inference*)



- Linearisation of the black-box data model:

$$\hat{\Phi}_\theta \approx f_0 + \nabla f_0 \cdot (\theta - \theta_0)$$

- Further assume:
  - Gaussian prior: $\mathcal{P}(\theta) = \mathcal{G}(\theta_0, \mathbf{S})$
  - Gaussian effective likelihood: $\mathcal{P}(\Phi|\theta) = \mathcal{G}\left[\mathbf{f}(\theta), \mathbf{C}_0\right]$

- The posterior is Gaussian and analogous to a Wiener filter:

expansion point

observed summaries

mean: $\gamma \equiv \theta_0 + \Gamma\,(\nabla f_0)^\intercal\, \mathbf{C}_0^{-1} (\Phi_\mathrm{O} - f_0)$

covariance: $\Gamma \equiv \left[(\nabla f_0)^\intercal\, \mathbf{C}_0^{-1} \nabla f_0 + \mathbf{S}^{-1}\right]^{-1}$

prior covariance

covariance of summaries

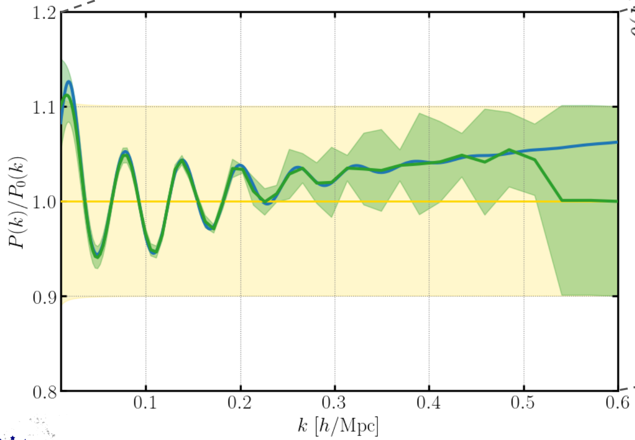
gradient of the black-box

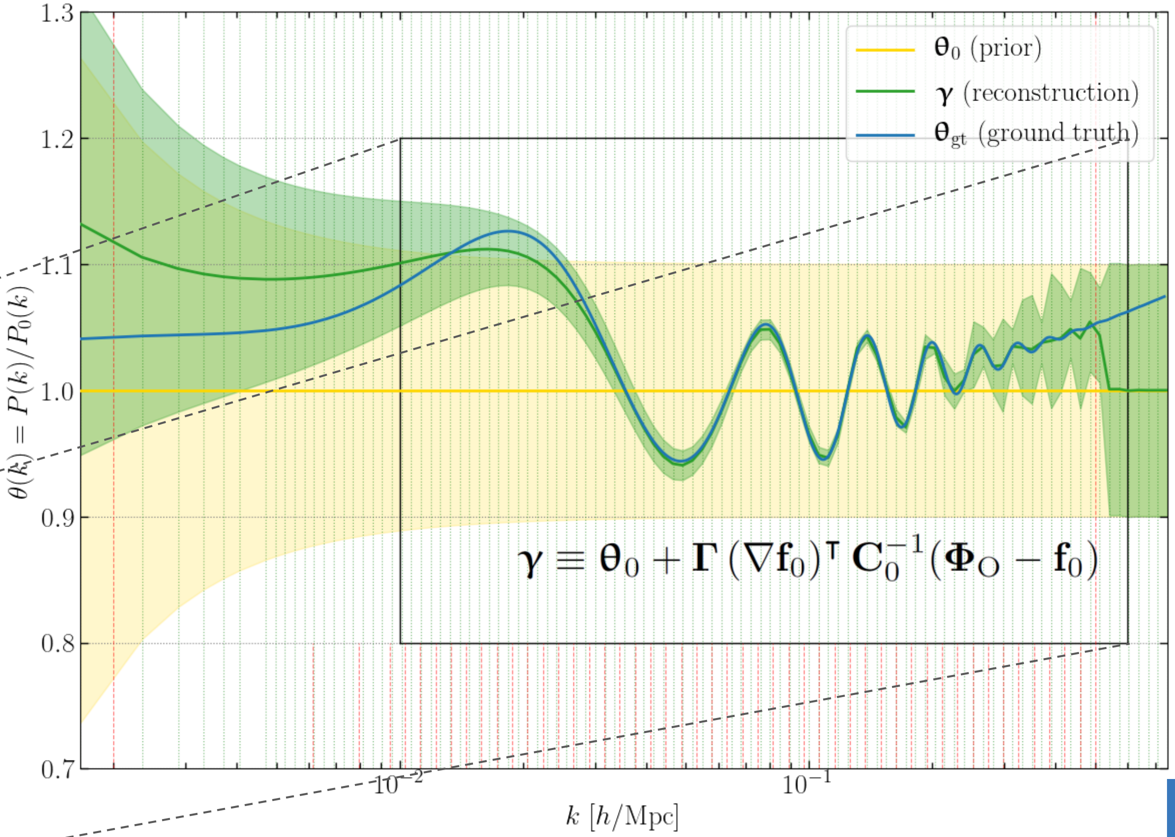
- f_0, $\mathbf{C}_0$ and ∇f_0 can be evaluated through simulations only.
- The number of required simulations is fixed *a priori* (contrary to MCMC).
- The workload is perfectly parallel.

FL, Enzi, Jasche & Heavens, 1902.10149

SELFI-1 Euclid forecast (cosmic variance limit)

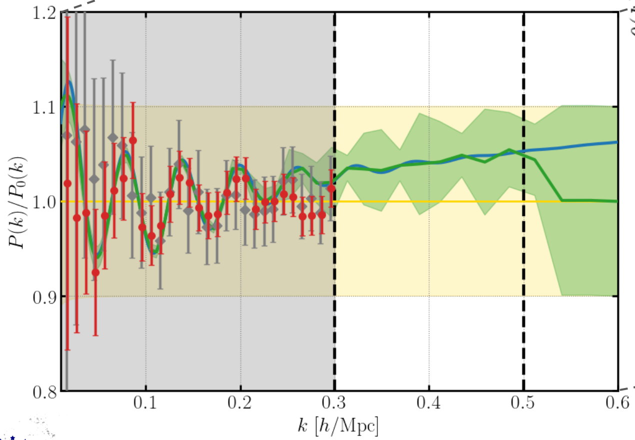
- $V = (3780\ \mathrm{Mpc}/h)^3$ (volume of the Euclid flagship simulation)
- Gaussian random field data model; 6,060 simulations
- 100 parameters are simultaneously inferred

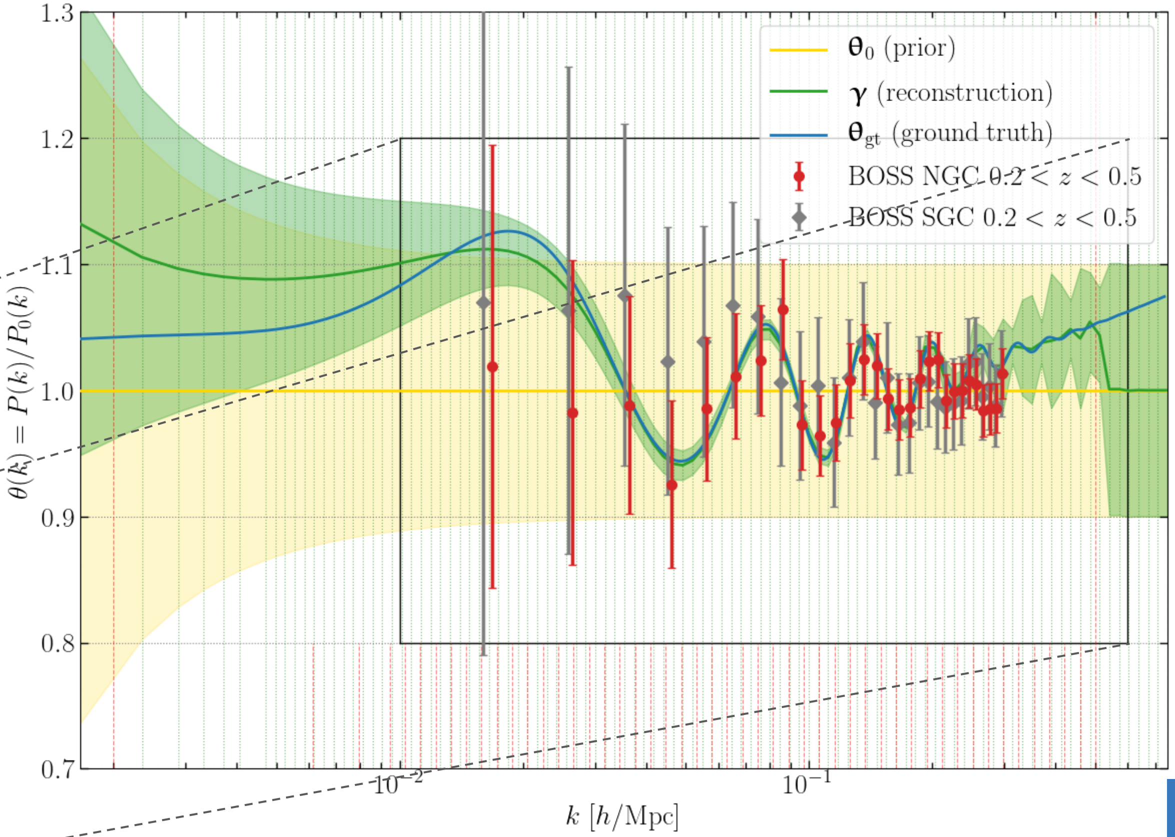




SELFI-1 Euclid vs BOSS

- Numerical data models allow using the galaxy power spectrum as summary statistics up to at least $k \gtrsim 0.5\ h/\mathrm{Mpc}$ safely
- $N_{\mathrm{modes}} \propto k^3$: 5 times more modes are used in the analysis.

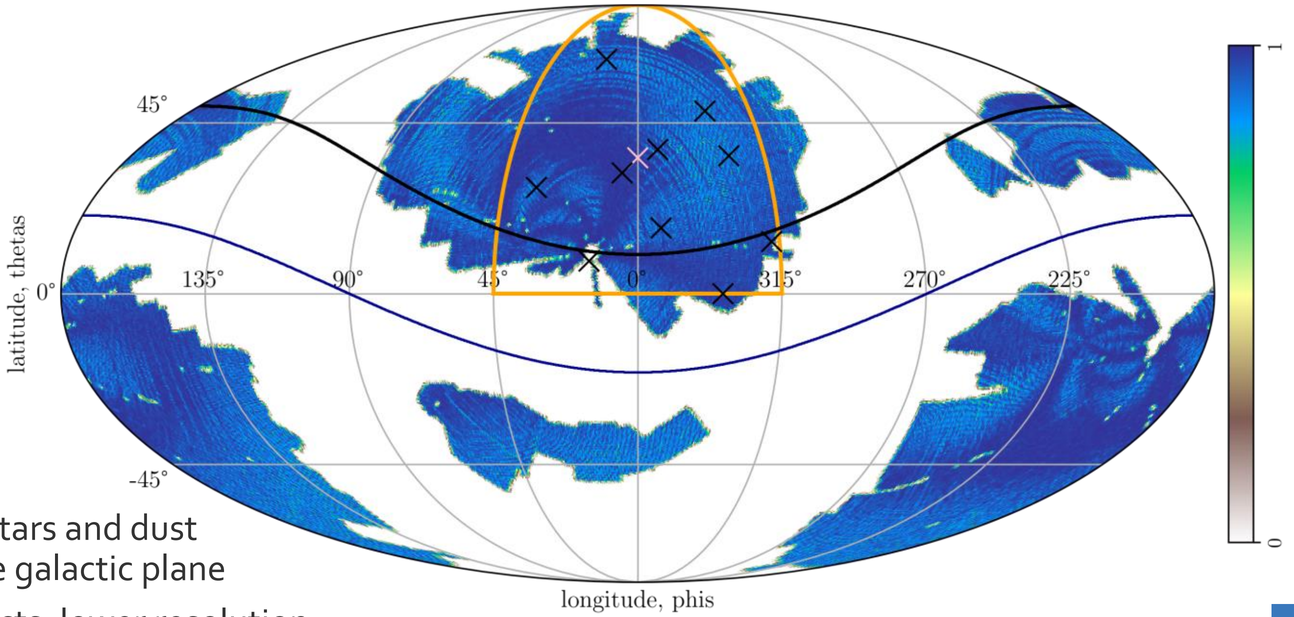




Data points from Beutler *et al.*, 1607.03149

Systematic effects at large scales: mask and selection functions

- $V = (3780\ \mathrm{Mpc}/h)^3$ cubic box, covering one octant of the sky

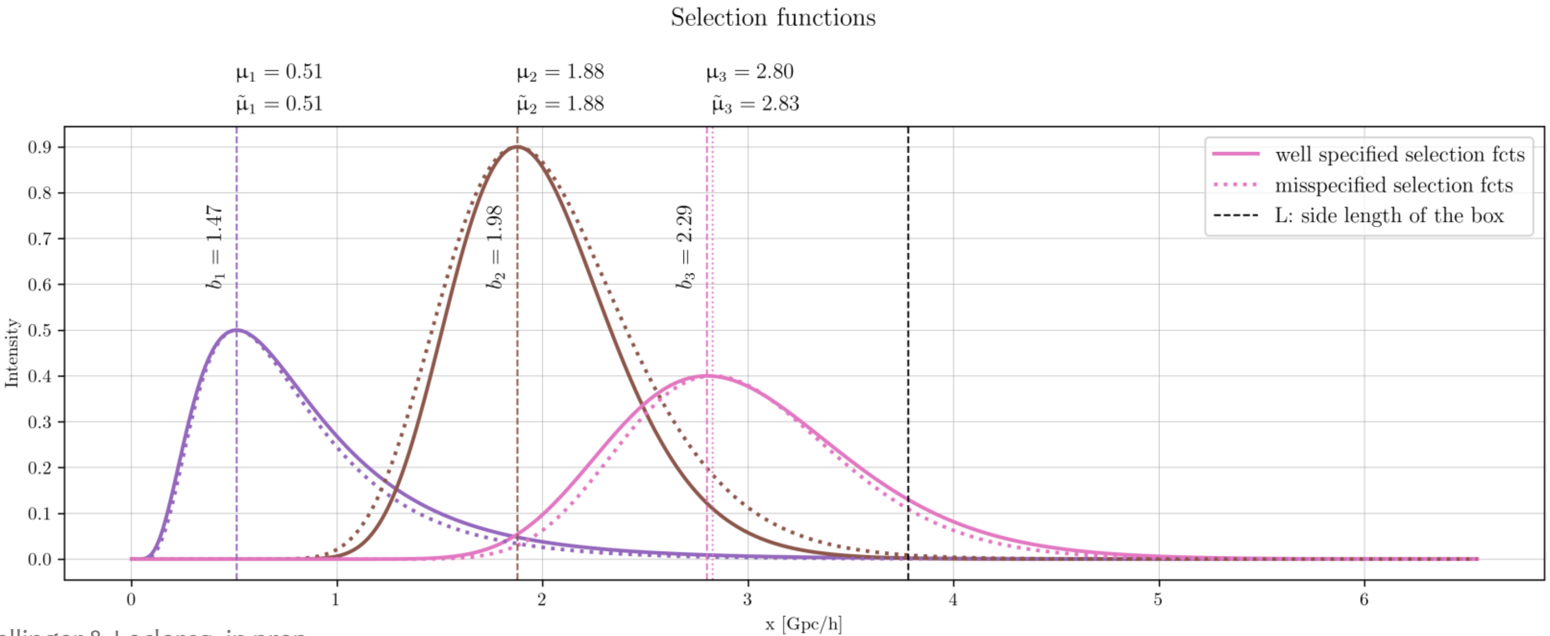


- Two models:
  - Model A: 10 masked stars and dust extinction close to the galactic plane
  - Model B: no such effects, lower resolution

Hoellinger & Leclercq, in prep.

Systematic effects at large scales: mask and selection functions

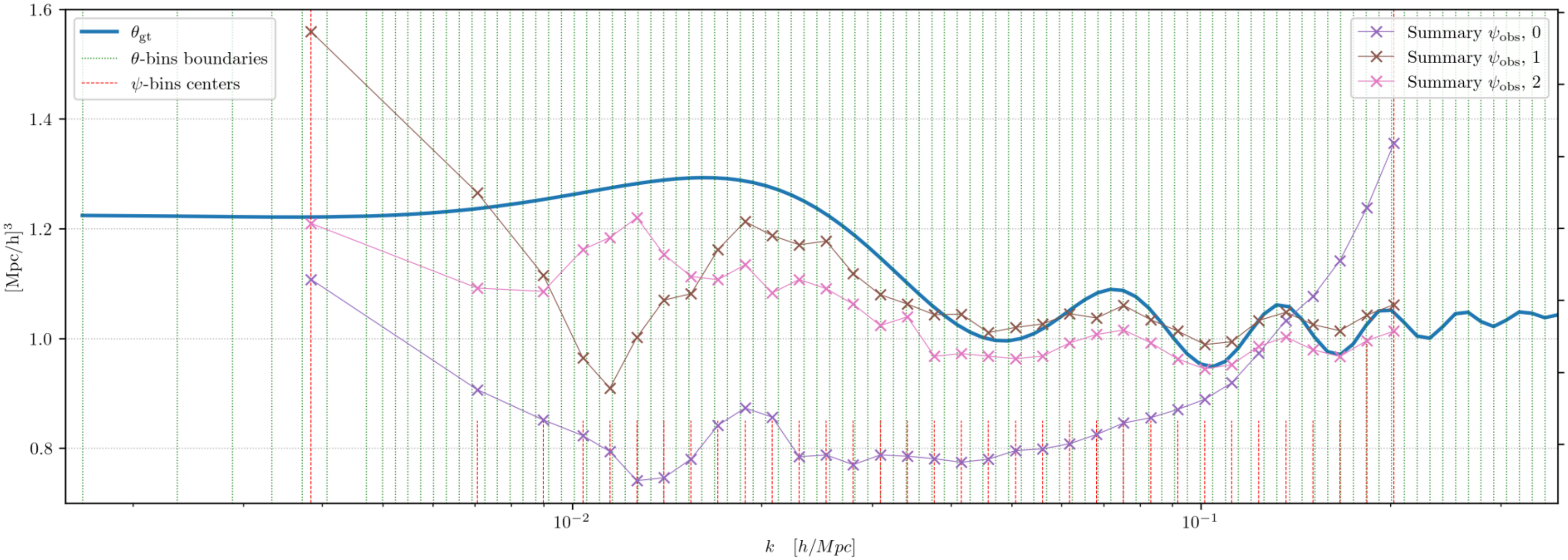
- Two models:
  - Model A: lognormal selection functions, luminosity-dependent galaxy bias
  - Model B: misspecified selection functions and galaxy biases



Hoellinger & Leclercq, in prep.

Synthetic observations

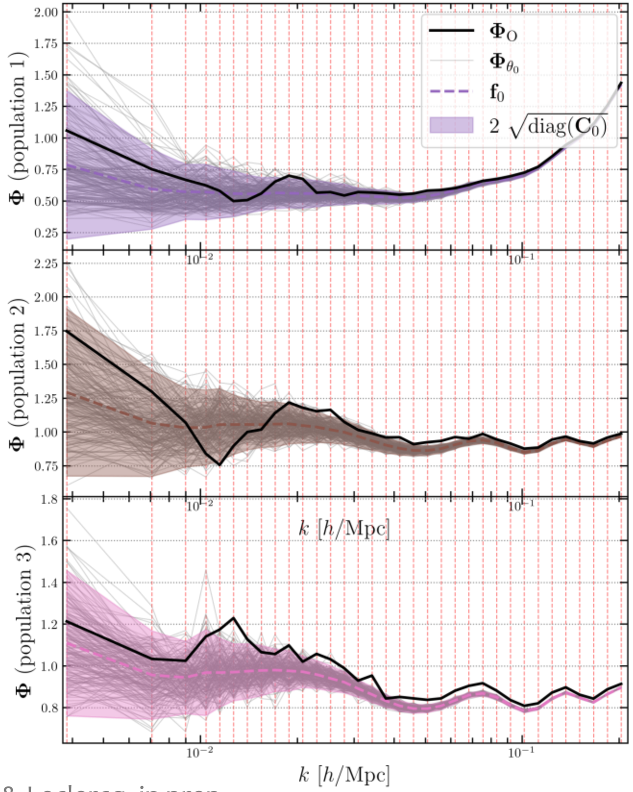
- Generated using model A:



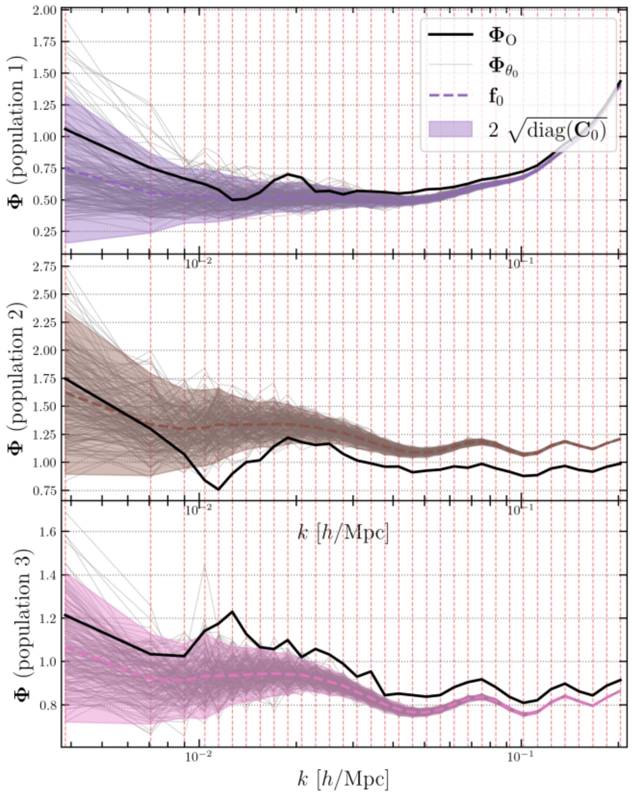
Hoellinger & Leclercq, in prep.

Synthetic observations versus simulations and linearised data models

Model A



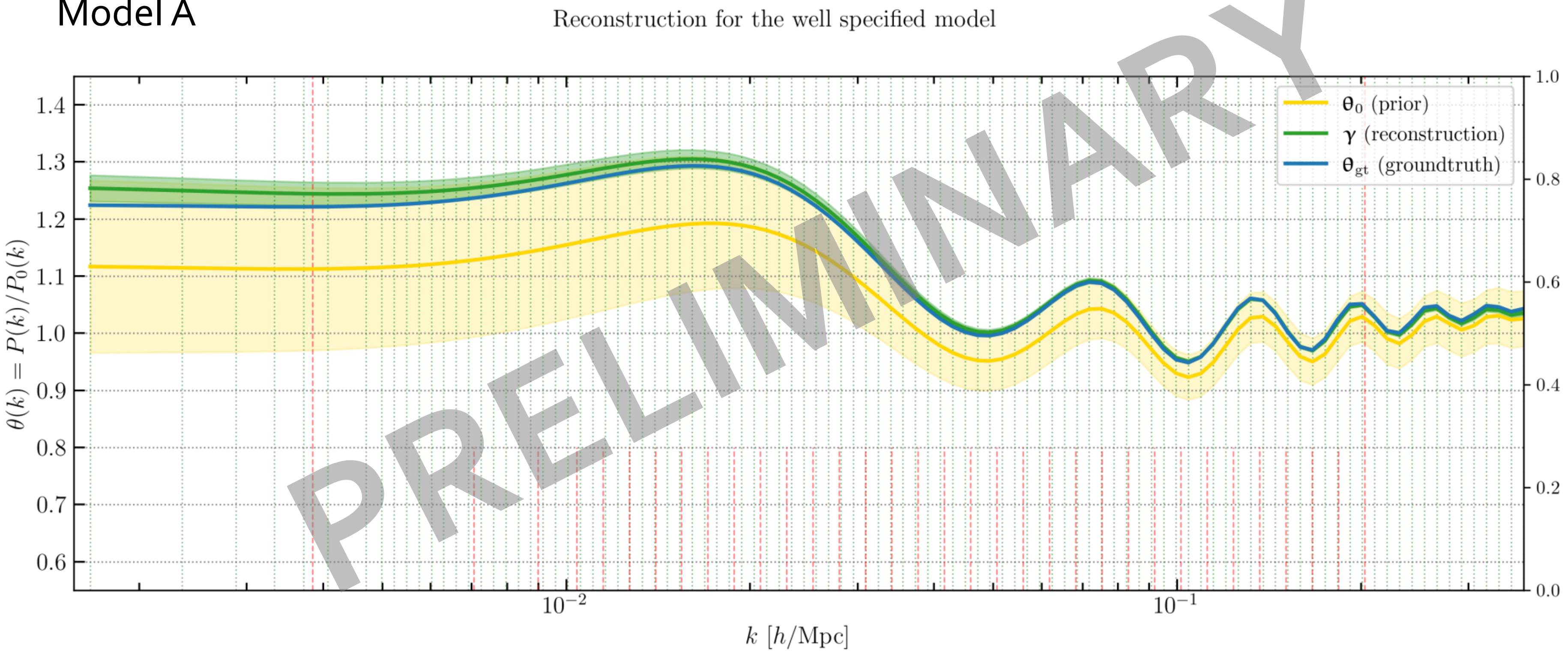
Model B



Hoellinger & Leclercq, in prep.

SELFI posterior: reconstructed initial matter power spectrum

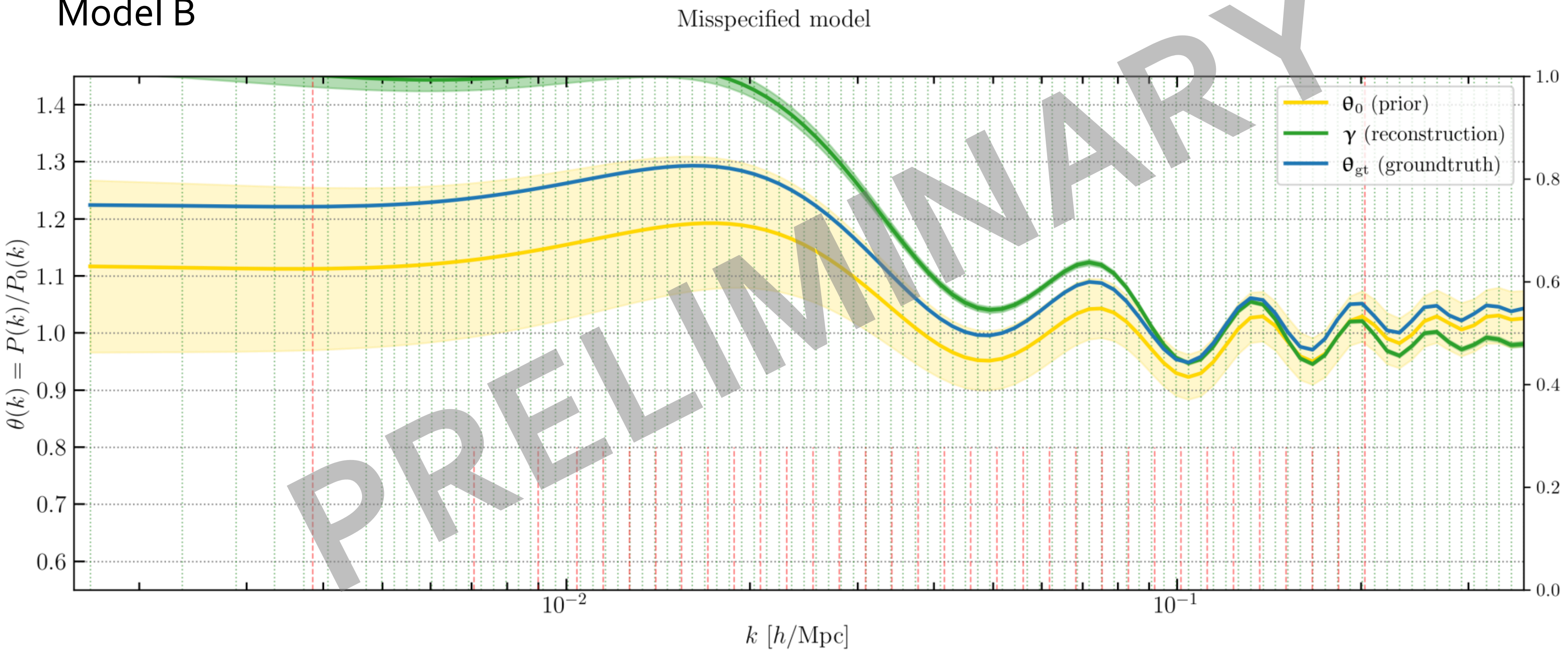
Model A



Hoellinger & Leclercq, in prep.

SELFI posterior: reconstructed initial matter power spectrum

Model B

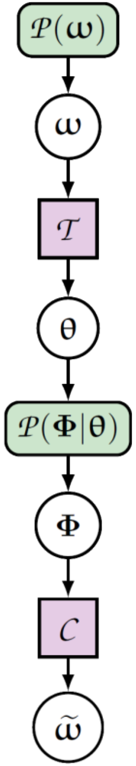


Hoellinger & Leclercq, in prep.

Check for model misspecification and data compression for SBI

- Qualitatively: the shape of the reconstructed θ is useful as a check for model misspecification (independent theoretical understanding).
- Quantitatively: we can use the Mahalanobis distance between the reconstruction γ and the prior distribution $\mathcal{P}(\theta)$:

$$d_\mathrm{M}(\boldsymbol{\gamma}, \boldsymbol{\theta}_0|\mathbf{S}) \equiv \sqrt{(\boldsymbol{\gamma}-\boldsymbol{\theta}_0)^\intercal\, \mathbf{S}^{-1} (\boldsymbol{\gamma}-\boldsymbol{\theta}_0)}$$



- The score function $\nabla_{\boldsymbol{\omega}} \hat{\ell}_{\boldsymbol{\omega}0}$ is the gradient of the log-likelihood at fiducial point $\boldsymbol{\omega}_0$ in parameter space.
- A quasi maximum-likelihood estimator for the parameters is

$$\mathcal{C}(\boldsymbol{\Phi}) = \tilde{\boldsymbol{\omega}} \equiv \boldsymbol{\omega}_0 + \mathbf{F}_0^{-1} \left[(\nabla_{\boldsymbol{\omega}} \mathbf{f}_0)^\intercal \boxed{\mathbf{C}_0^{-1}} (\boldsymbol{\Phi} - \boxed{\mathbf{f}_0})\right]$$

Fisher matrix: $\mathbf{F}_0 = (\nabla_{\boldsymbol{\omega}} \mathbf{f}_0)^\intercal \boxed{\mathbf{C}_0^{-1}} \nabla_{\boldsymbol{\omega}} \mathbf{f}_0$

$$\nabla_{\boldsymbol{\omega}} \mathbf{f}_0 = \boxed{\nabla \mathbf{f}_0} \cdot \boxed{\nabla_{\boldsymbol{\omega}} \mathcal{T}_0}$$

Already computed for SELFI

Cheap via finite differences

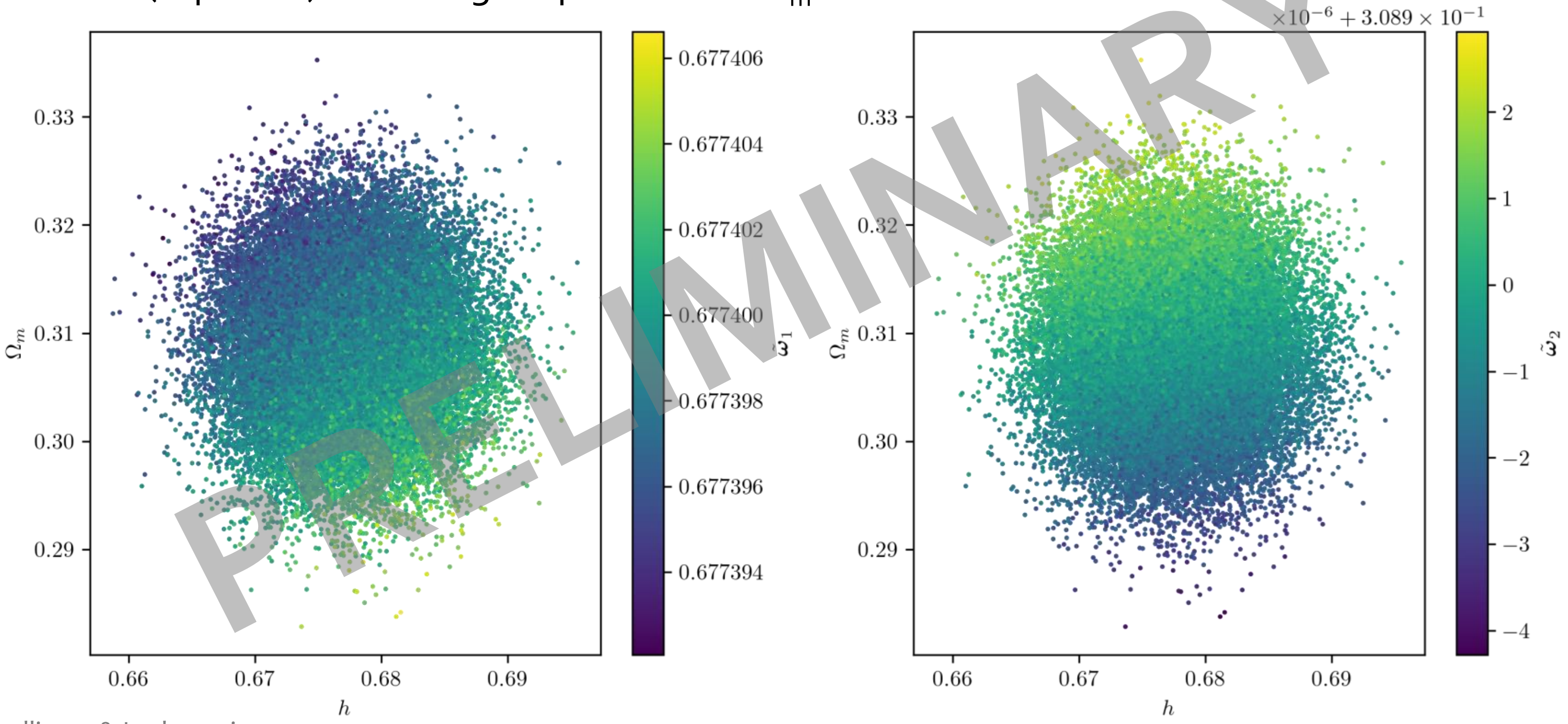
- Score compression is optimal in the sense that it preserves the Fisher information content of the data.

FL, 2209.11057

Alsing & Wandelt, 1712.00012

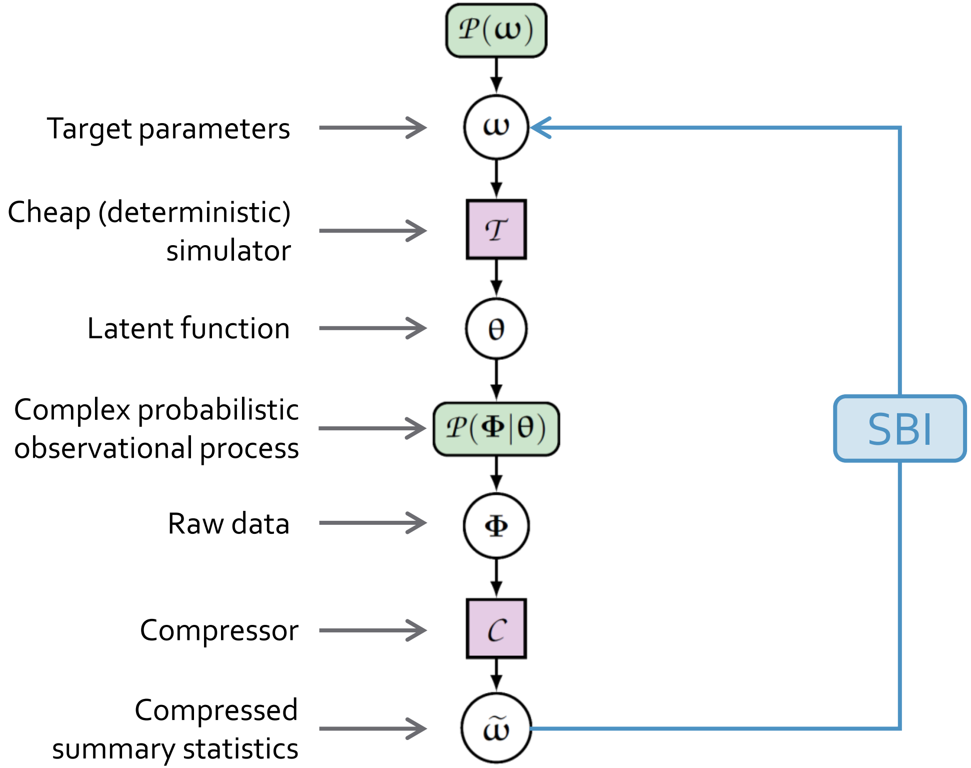
Score compression of the observed and simulated statistical summaries

- For two (top-level) cosmological parameters Ω_m and h:



Hoellinger & Leclercq, in prep.

Simulation-based inference of top-level target parameters

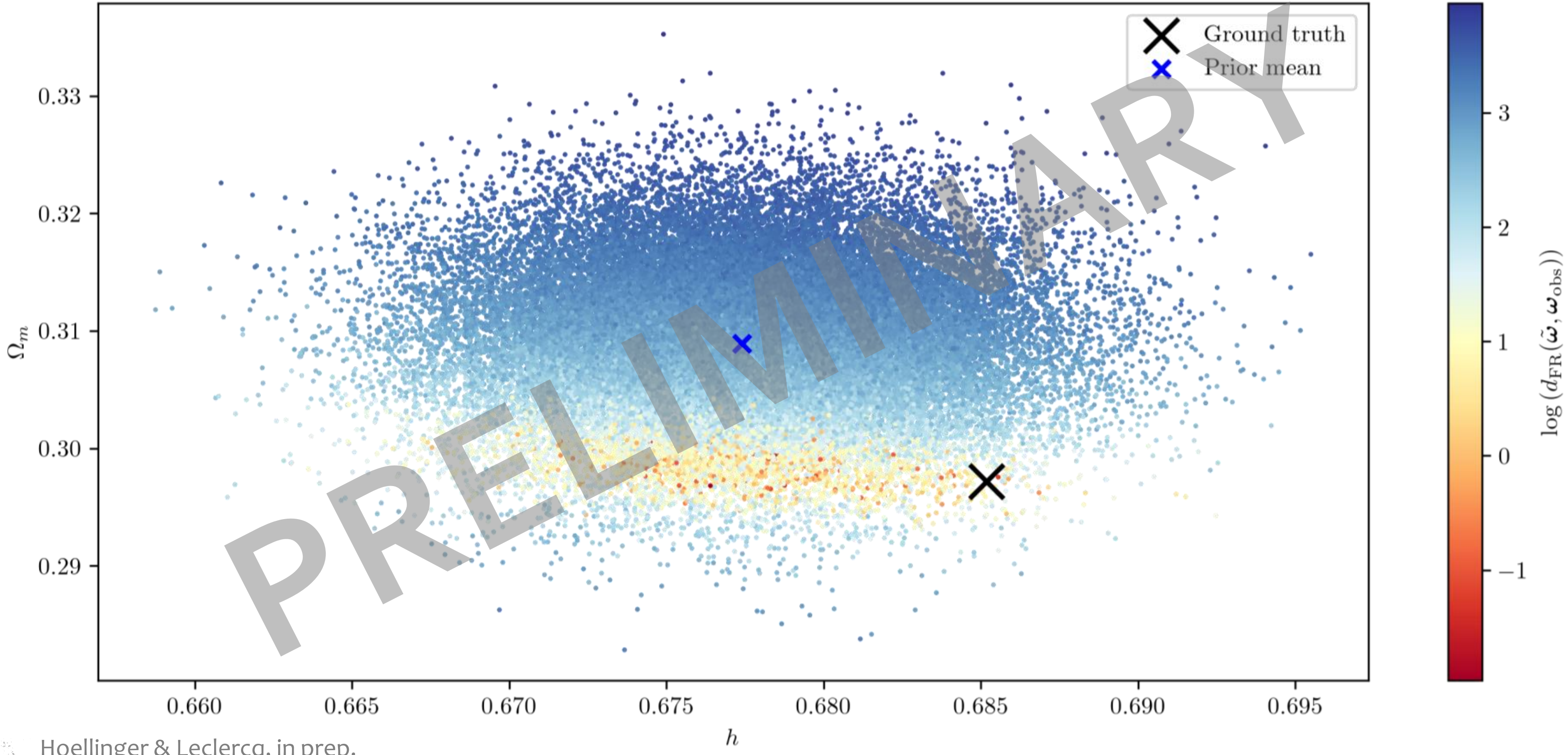


- Any SBI algorithm can be used to obtain the posterior $\mathcal{P}(\omega|\tilde{\omega}_\mathrm{O})$.
- Final inference:
  - does not depend on the assumptions made to check for model misspecification,
  - is unbiased (only more conservative) in case data compression is lossy.
- Non-parametric approaches can use the Fisher-Rao distance between simulated summaries $\tilde{\omega}$ and observed summaries $\tilde{\omega}_\mathrm{O}$:

$$d_\mathrm{FR}(\tilde{\omega}, \tilde{\omega}_\mathrm{O}) \equiv \sqrt{(\tilde{\omega} - \tilde{\omega}_\mathrm{O})^\intercal \, \mathbf{F}_0 (\tilde{\omega} - \tilde{\omega}_\mathrm{O})}$$

FL, 2209.11057

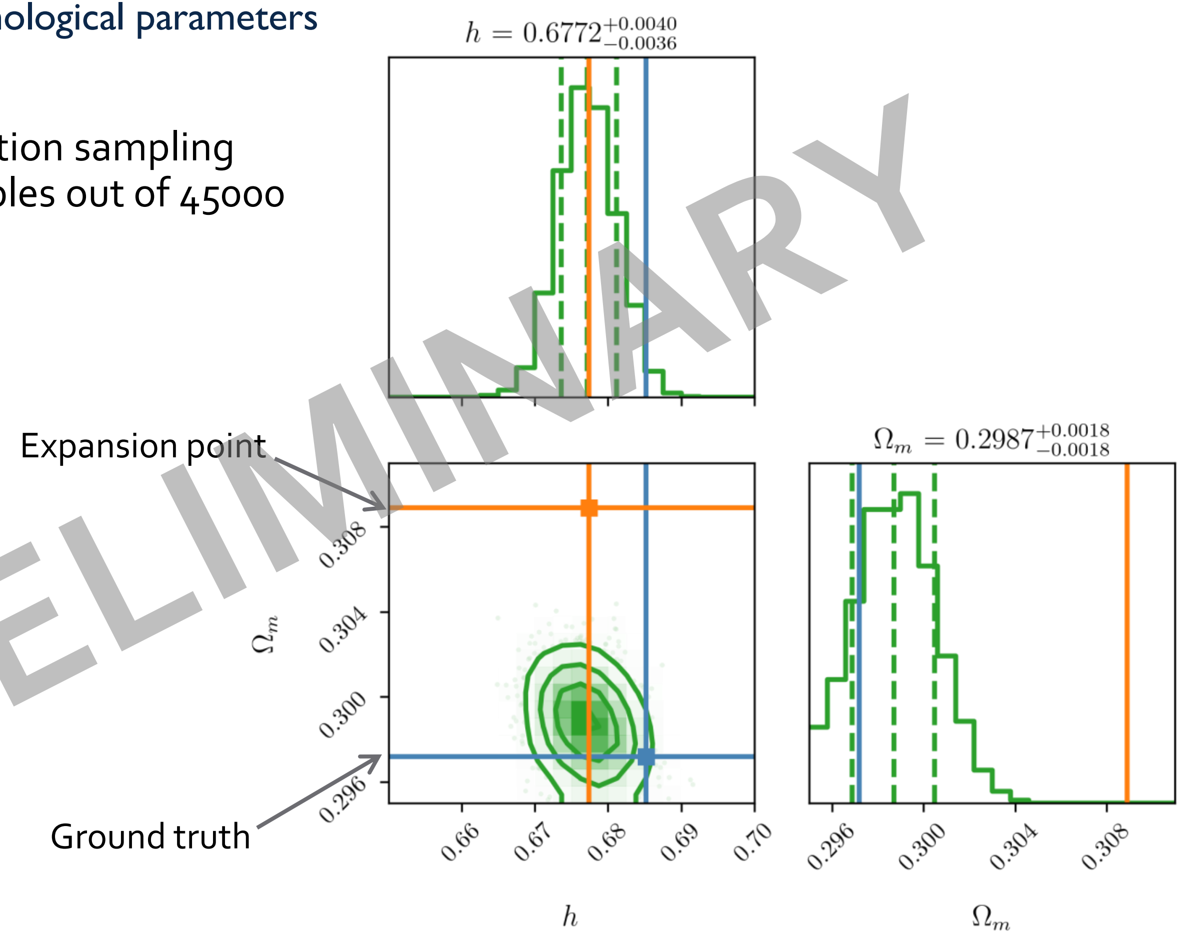
Fisher-Rao distance between simulated and observed summaries



Hoellinger & Leclercq, in prep.

Final SBI posterior on cosmological parameters

- After likelihood-free rejection sampling (ABC): 1687 selected samples out of 45000 simulations:



Hoellinger & Leclercq, in prep.

Conclusion: a science-ready statistical framework for the galaxy clustering additional probes

- A novel two-step simulation based Bayesian approach, combining SELFI and SBI, to tackle the issue of model misspecification for a large class of BHMs.
- Advantages of the first step (SELFI):
  - Even if the inference is in high dimension, the simulator remains a black-box.
  - The number of simulations is fixed *a priori* by the user.
  - The computational workload is perfectly parallel.
  - The linearised data model is trained once and for all independently of the data vector (amortisation).
- Advantages of the second step (SBI):
  - SELFI quantities provide a score compressor for free.
  - General advantages of SBI with respect to likelihood-based methods are preserved.
  - Inference does not depend on the assumptions made to check for model misspecification.

➢ A computationally efficient and easily applicable framework to perform SBI of BHMs while checking for model misspecification.

pySELFI is publicly available at https://pyselfi.florent-leclercq.eu.

- Thanks for listening!

- References:

FL, Enzi, Jasche & Heavens, 1902.10149

FL, 2209.11057

Hoellinger & Leclercq, in prep.

www.florent-leclercq.eu

www.aquila-consortium.org