Likelihood-free inference from galaxy surveys

Prospects for Euclid

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

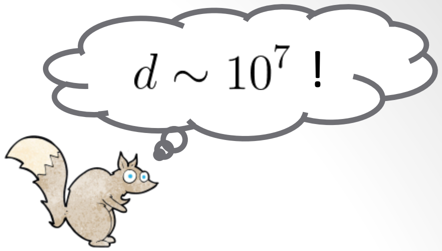
Wolfgang Enzi, Alan Heavens,
Jens Jasche, Guilhem Lavaux,

and the Aquila Consortium
www.aquila-consortium.org

June 27th, 2019

ICIC
Imperial Centre
for Inference & Cosmology

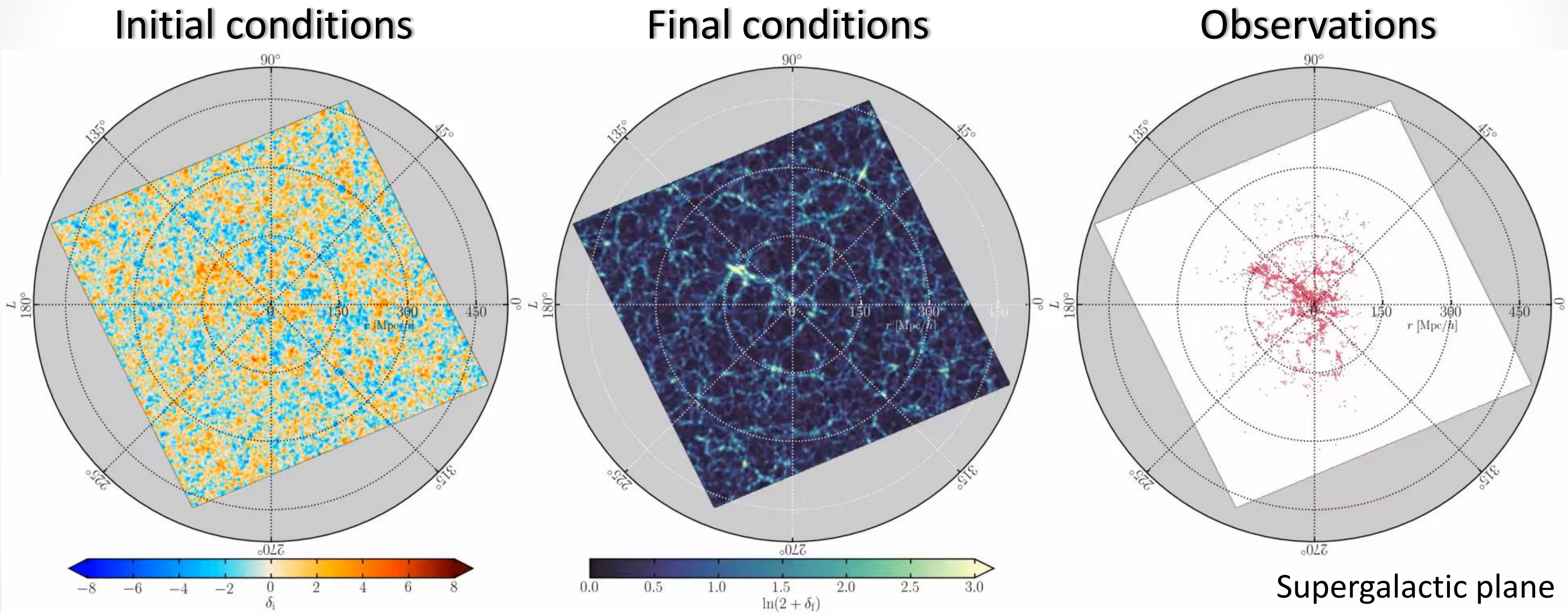
Imperial College
London



Vocabulary considerations I:

What is the likelihood?

In cosmology, the (true?) likelihood should live at the level of the map of the CMB or LSS. e.g. Wiener filtering for the CMB, BORG for the LSS (a 256^3-dimensional Poisson likelihood):

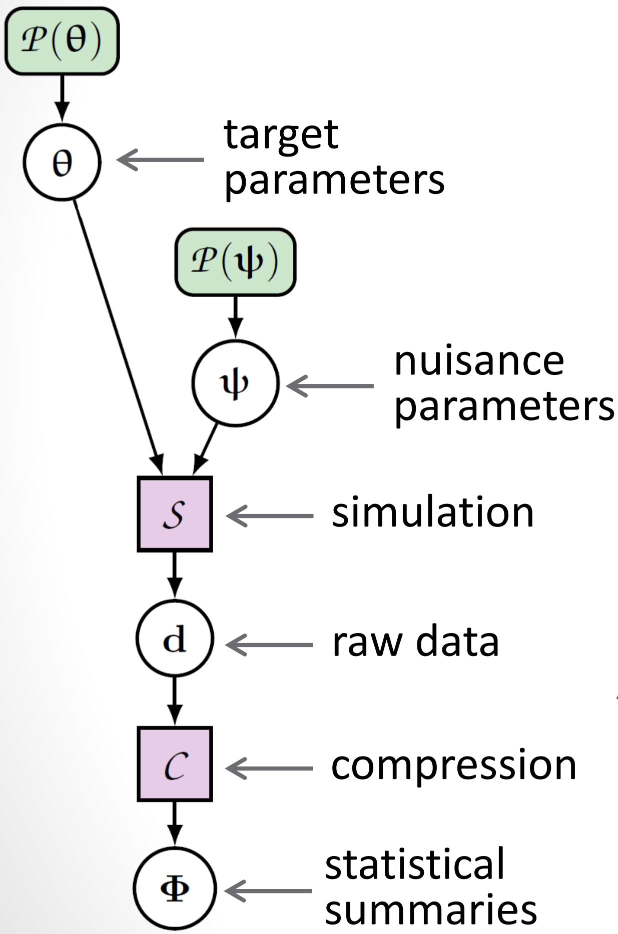


Jasche & Lavaux 2019, 1806.11117 – FL, Lavaux & Jasche, in prep.

Expert knowledge of the likelihood is needed to beat the curse of dimensionality: conditionals/gradients of the likelihood are required by the samplers (Gibbs/Hamiltonian).

Vocabulary considerations II:

You may already be an LFI specialist!



- Likelihood-free inference (LFI) techniques bypass the need for a map-level likelihood, by relying instead only on a "black-box".
- The likelihood is replaced by a measure of the distance/discrepancy Δ between simulated and observed statistical summaries of the data.

e.g. $\mathbf{d} =$ full galaxy survey data
$\boldsymbol{\Phi} = \{\widehat{P}(k)\}$ estimated power spectrum
$\Delta =$ Mahalanobis distance with covariance matrix Σ

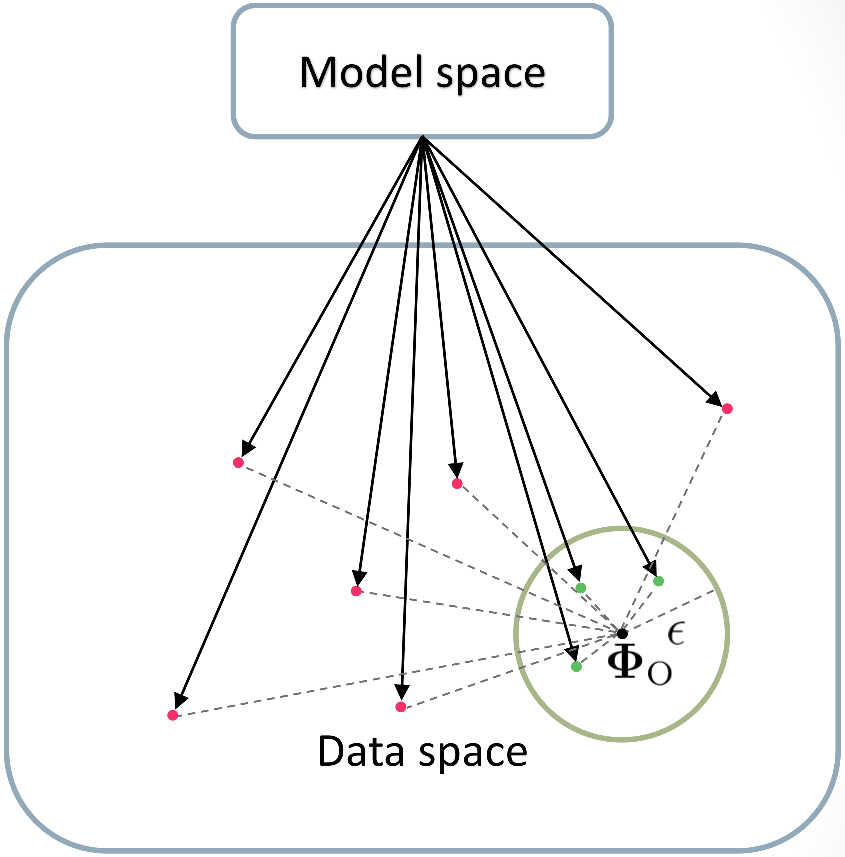
$$\Delta(\boldsymbol{\Phi}_\theta, \boldsymbol{\Phi}_\mathrm{O}) = \sqrt{\sum_{k,k'} \left[\widehat{P}_\theta(k) - \widehat{P}_\mathrm{O}(k)\right]^\intercal \Sigma^{-1}_{k,k'} \left[\widehat{P}_\theta(k') - \widehat{P}_\mathrm{O}(k')\right]}$$

Note that this is what many people would call…
(square root of -2 times) the log-likelihood!

- What is "primordial" depends mostly on your ambition…

Likelihood-free rejection sampling (LFRS)

- Iterate many times:
  - Sample θ from a proposal distribution $q(\theta)$
  - Simulate Φ_θ using the black-box
  - Compute the distance $\Delta(\Phi_\theta, \Phi_\mathrm{O})$ between simulated and observed data
  - Retain θ if $\Delta(\Phi_\theta, \Phi_\mathrm{O}) \leq \epsilon$, otherwise reject



ϵ can be adaptively reduced (Population Monte Carlo)

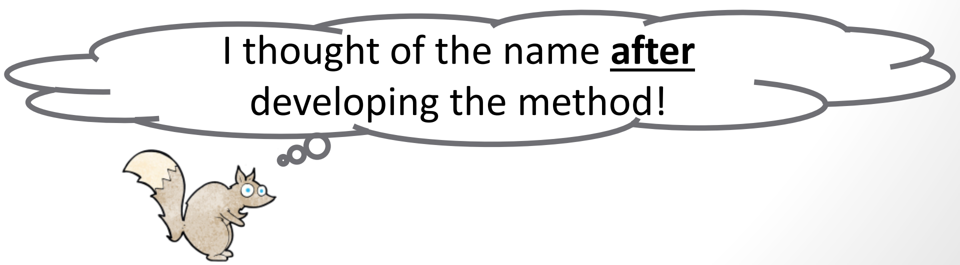
Beyond LFRS: two scenarios

The "number of simulations" route:

- Specific cosmological models ($d \lesssim 10$), general exploration of parameter space
- Density Estimation for Likelihood-Free Inference (DELFI)
  Papamakarios & Murray 2016, 1605.06376
  Alsing, Feeney & Wandelt 2018, 1801.01497
  Alsing, Charnock, Feeney & Wandelt 2019, 1903.00007
- Bayesian Optimisation for Likelihood-Free Inference (BOLFI)
  Gutmann & Corander 2016, 1501.03291
  FL 2018, 1805.07152

The "number of parameters" route:

- Model-independent theoretical parametrisation ($d \gtrsim 100$), strong existing constraints in parameter space
- Simulator Expansion for Likelihood-Free Inference (SELFI)
  FL, Enzi, Jasche & Heavens 2019, 1902.10149

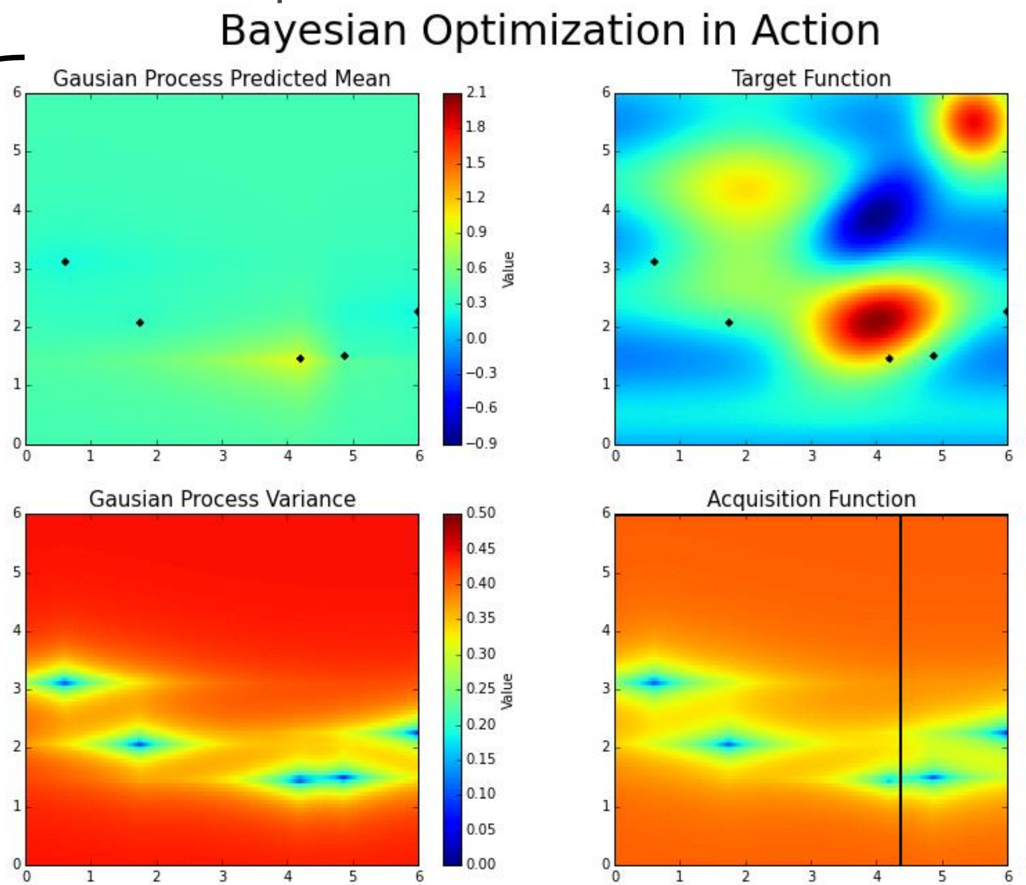


The “number of simulations” route: BOLFI

Bayesian Optimisation for Likelihood-Free Inference

BOLFI: Data acquisition

Regression of the distance between observed and simulated data



F. Nogueira, https://github.com/fmfn/BayesianOptimization

Expected utility of the next simulation in parameter space

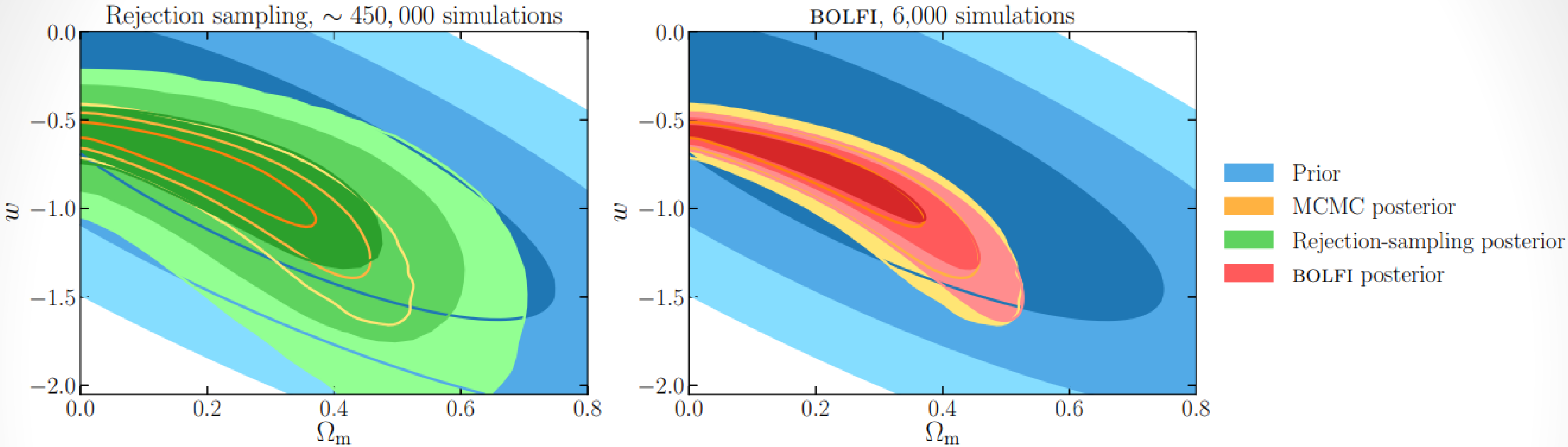
The optimal acquisition function for ABC can be written down:
the **Expected Integrated Variance** (ExpIntVar).

Järvenpää et al. 2017, 1704.00520 (expression of ExpIntVar in the non-parametric approach)

FL 2018, 1805.07152 (expression of ExpIntVar in the parametric approach)

BOLFI: Re-analysis of the JLA supernova sample

Betoule et al. 2014, 1401.4064



- The **number of required simulations is reduced** by:
  - 2 orders of magnitude with respect to likelihood-free rejection sampling (for a much better approximation of the posterior)
  - 3 orders of magnitude with respect to exact Markov Chain Monte Carlo sampling

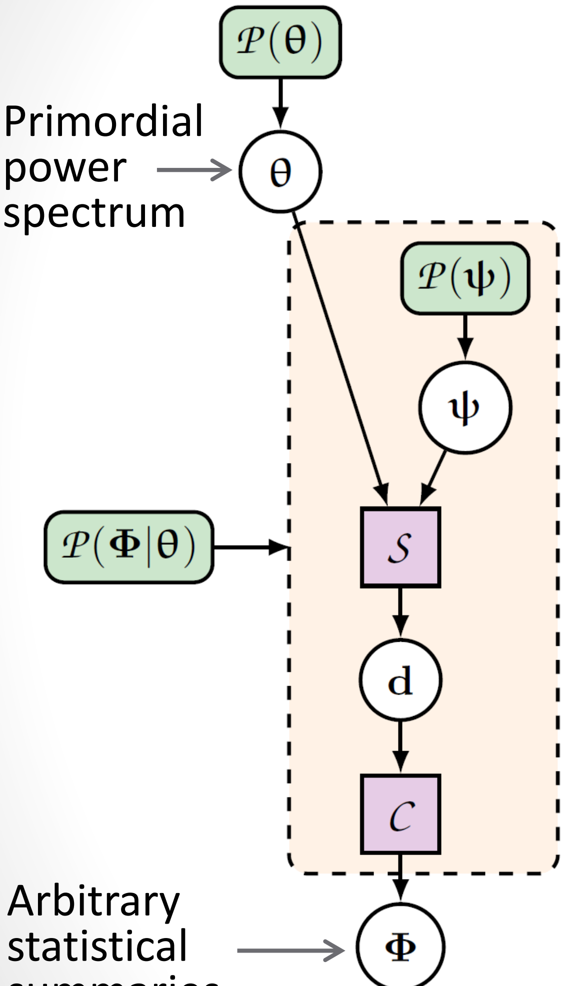
  FL 2018, 1805.07152

- Cheap numerical **data models** (to be evaluated O(10^3) times) will be required for a Euclid-like survey.

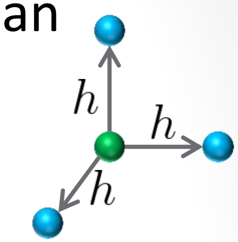
The “number of parameters” route: SELFI

Simulator Expansion for Likelihood-Free Inference

SELFI: Method



- Gaussian prior + Gaussian effective likelihood
- Linearisation of the black-box around an expansion point + finite differences:



$$\hat{\Phi}_\theta \approx \mathbf{f}_0 + \nabla\mathbf{f}_0 \cdot (\theta - \theta_0)$$

The posterior is Gaussian and analogous to a Wiener filter:

expansion point — observed summaries

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

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

prior covariance

covariance of summaries — gradient of the black-box

$\mathbf{f}_0, \mathbf{C}_0$ and $\nabla\mathbf{f}_0$ can be evaluated through simulations only.
The number of required simulations is fixed *a priori*.

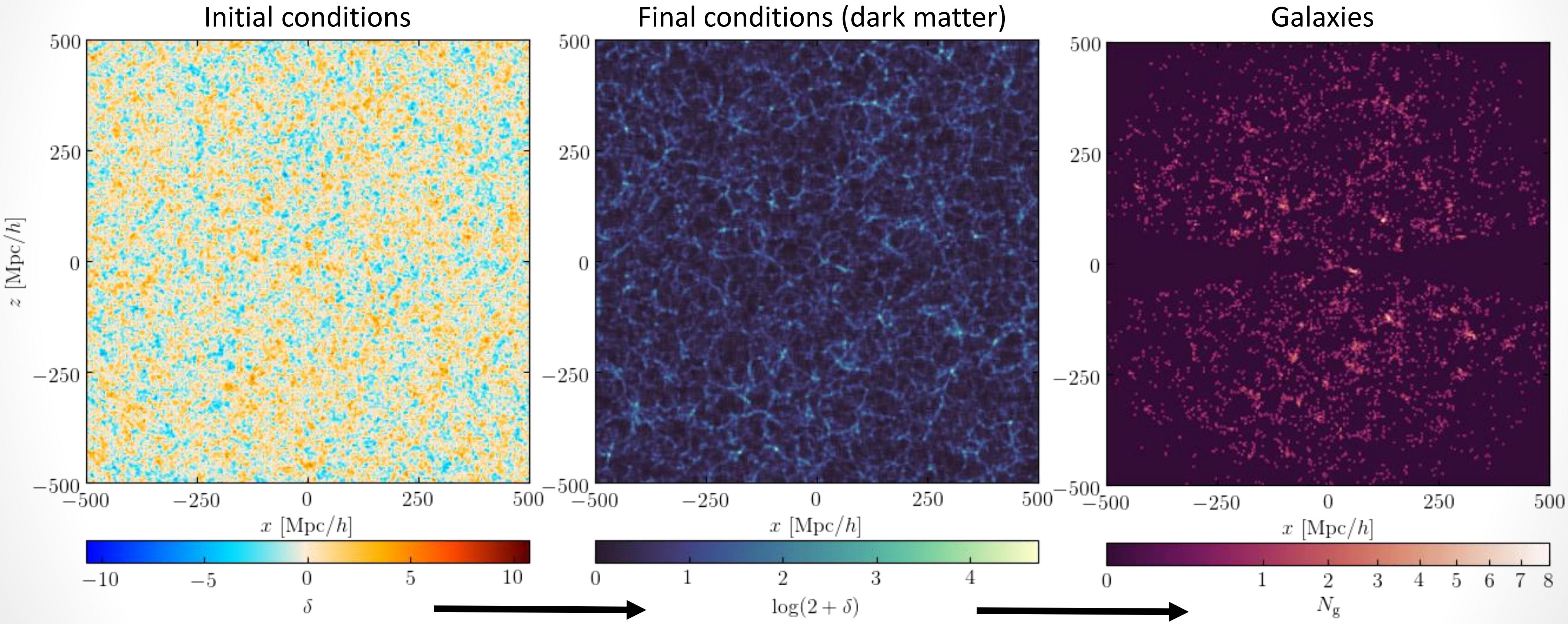
FL, Enzi, Jasche & Heavens 2019, 1902.10149

A black-box: Simbelmynë

I'm happy to explain the name during the coffee break...

Publicly available code:

https://bitbucket.org/florent-leclercq/simbelmyne/

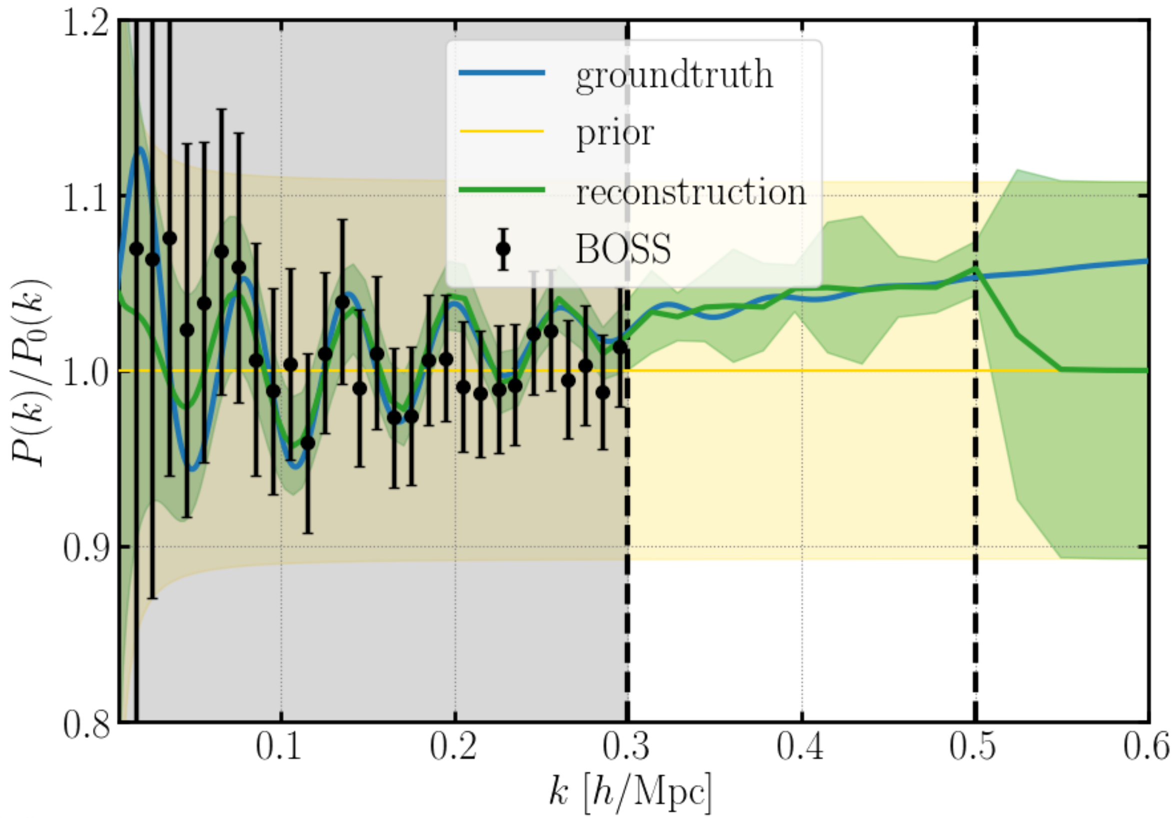


Dark matter simulation with COLA

Tassev, Zaldarriaga & Eisenstein 2013, 1301.0322

Survey simulation: Redshift-space distortions, galaxy bias, selection effects, survey geometry, instrumental noise

SELFI + Simbelmynë: Proof-of-concept



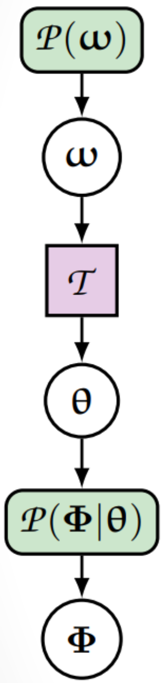
100 parameters are simultaneously inferred from a black-box data model

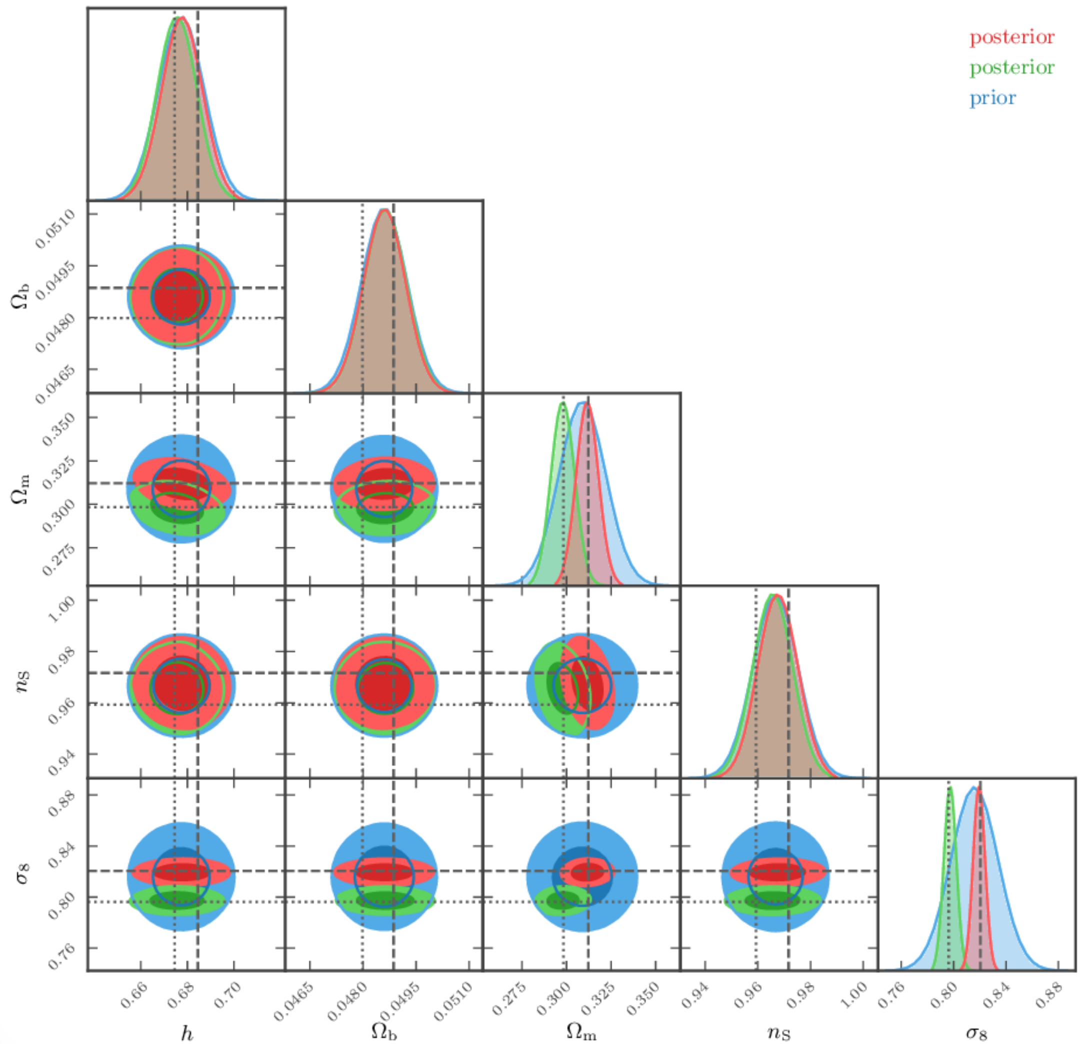
$N_{\text{modes}} \propto k^3$: **5** times more modes are used in the analysis

1 (Gpc/h)^3 only! Much more potential for Euclid data…

FL, Enzi, Jasche & Heavens 2019, 1902.10149

SELFI + Simbelmynë: Proof-of-concept





- Robust inference of cosmological parameters can be easily performed *a posteriori* once the linearised data model is learnt
- pyselfi will be made publicly available soon

FL, Enzi, Jasche & Heavens 2019, 1902.10149

Concluding thoughts

- Goal: developing and using algorithms for targeted questions, allowing the use of simulators including all relevant physical and observational effects.
- Bayesian analyses of galaxy surveys with fully non-linear numerical black-box models is not an impossible task!
- The “number of simulations route” (BOLFI):
  - The optimal acquisition function can be derived: the Expected Integrated Variance.
  - The number of simulations is reduced by several orders of magnitude.
- The “number of parameters route” (SELFI):
  - High-dimensional likelihood-free problems can be addressed.
  - The computational workload is fixed *a priori* and perfectly parallel.