Simulation-based inference, Bayesian hierarchical models, and model misspecification

Learning the Universe Meeting

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

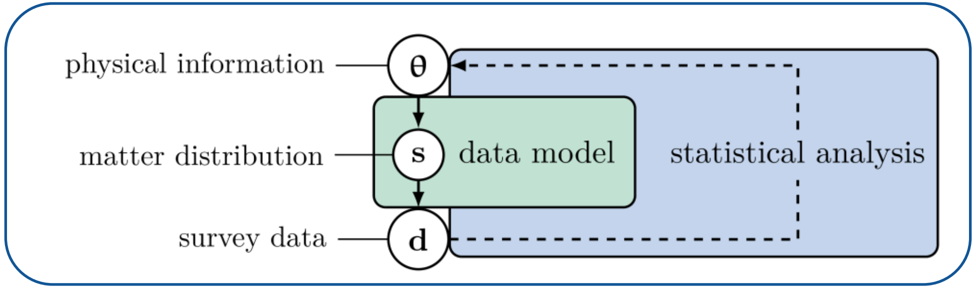
In collaboration with the Aquila Consortium

www.aquila-consortium.org

15 September 2022

Making inferences requires advanced Bayesian techniques

- Complex computer models are incorporated into Bayesian hierarchical models:



- The challenge: using new statistical methods is necessary. Two approaches are possible:

<u>Data assimilation:</u>

exact statistical analysis

approximate data model

<u>Simulation-based inference:</u>

approximate statistical analysis

arbitrary data model

Correlation functions versus field-level inference

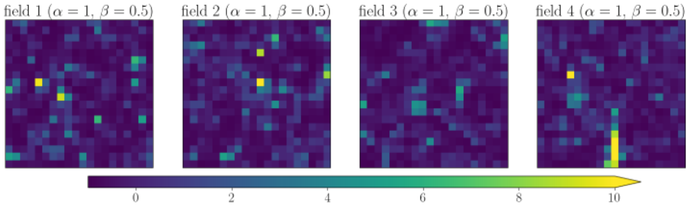
- We checked accuracy and precision of different methods for a log-normal model:

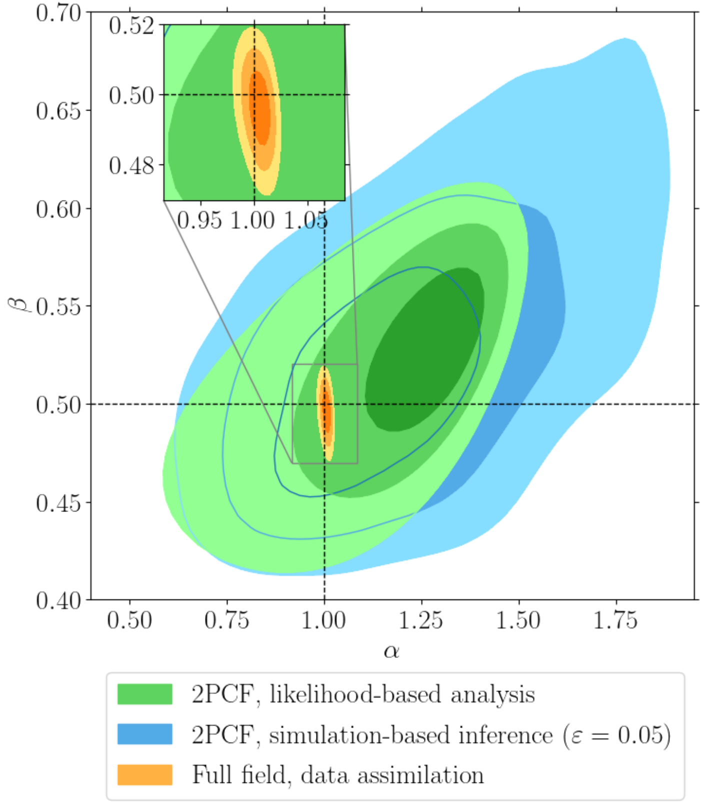
$$f = \frac{1}{\alpha}\left[\exp\left(\alpha g - \frac{1}{2}\alpha^2\right) - 1\right]$$

Log-normal field

Gaussian field with 2PCF:

$$\xi_g(r) = \exp\left(-\frac{1}{4}\frac{r^2}{\beta^2}\right)$$





- 2PCF likelihood-based analysis is *imprecise* and *inaccurate*
- 2PCF simulation-based inference is *imprecise* but *accurate*
- Full-field data assimilation is *precise* and *accurate*

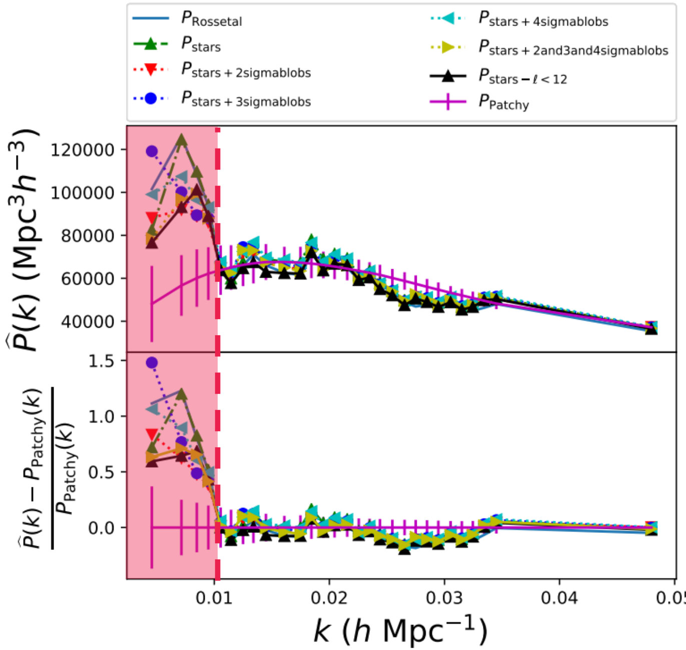
Leclercq & Heavens, 2103.04158

Companion repository:
https://github.com/florent-leclercq/correlations_vs_field

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.
- An example in cosmology: the galaxy *power spectrum.*

Due to observational systematics, we are unable to formulate *any* model that fits the data at large scales.

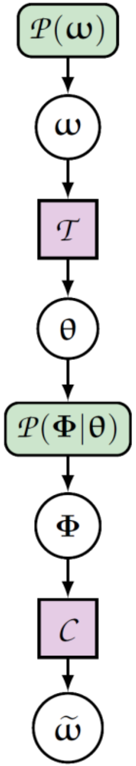


Kalus *et al.* 1806.02789

- Model misspecification: a major challenge particularly for approaches that marginalise over latent variables, such as simulation-based inference (SBI).
- Some recent work: diagnosing such issues and performing conservative belief updates in the presence of model misspecification, via e.g. tempering of the explicit likelihood (when it exists) or of a loss function.

Frazier, Robert & Rousseau 1708.01974, Thomas & Corander 1912.05810, Thomas *et al.* 2002.09377

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



Target parameters → ω

Cheap (deterministic) simulator → $\mathcal{T}$

Latent function → θ

Complex probabilistic observational process → $\mathcal{P}(\Phi|\theta)$

Raw data → Φ

Compressor → $\mathcal{C}$

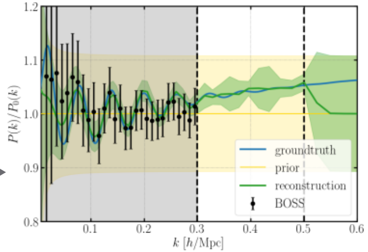
Compressed summary statistics → $\tilde{\omega}$

Example from cosmology:

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

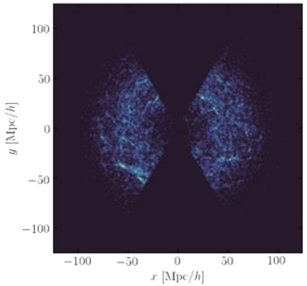
Solver of the Boltzmann equation (CAMB, CLASS)

Matter power spectrum →



N-body simulation, galaxy formation, survey modelling…

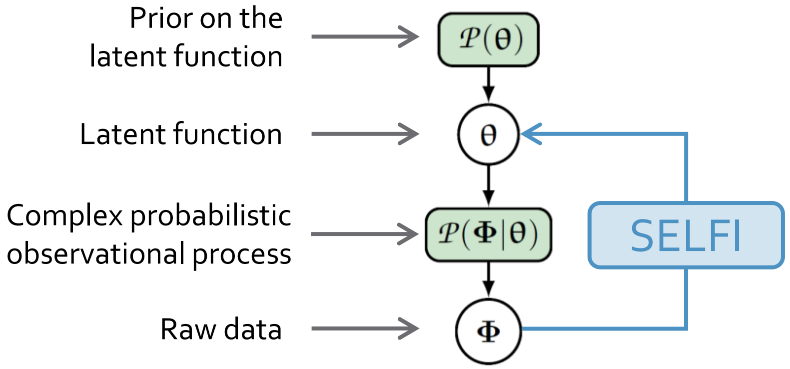
Galaxy survey data →



Score compression

Quasi maximum-likelihood estimators → $\widetilde{\Omega_\mathrm{m}}, \widetilde{\Omega_\mathrm{b}}, \widetilde{H_0}, \widetilde{n_\mathrm{S}}, \widetilde{\sigma_8}, \widetilde{\tau}$

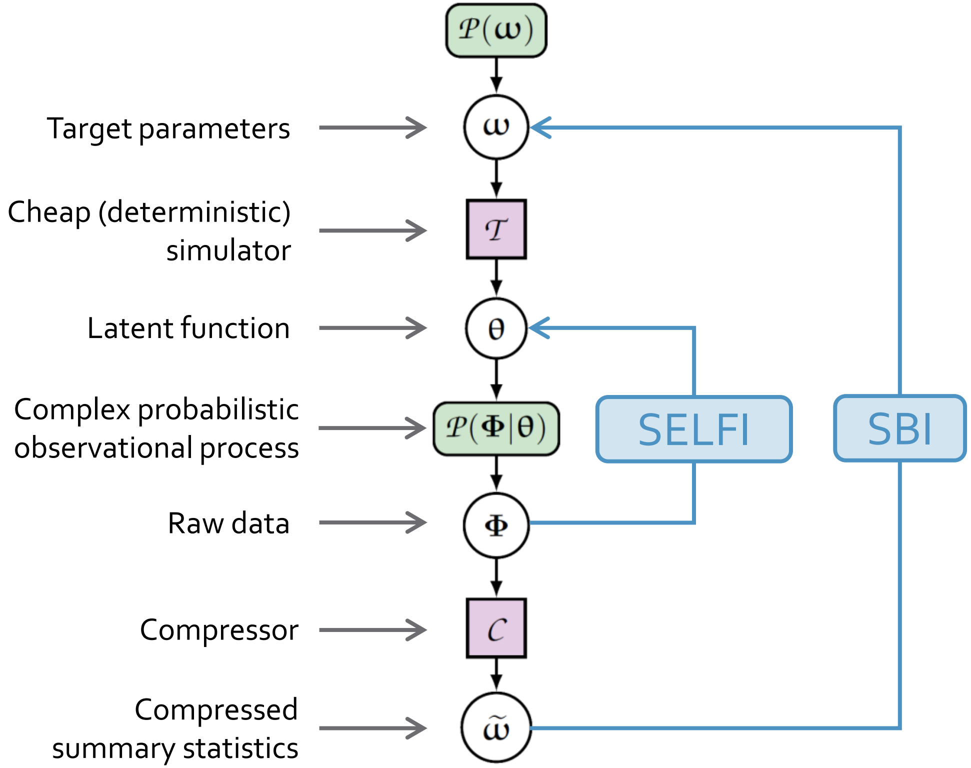
Key idea: a two-step SBI process that recycles simulations



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

Leclercq, MaxEnt 2022 proceedings (on arXiv next week)

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



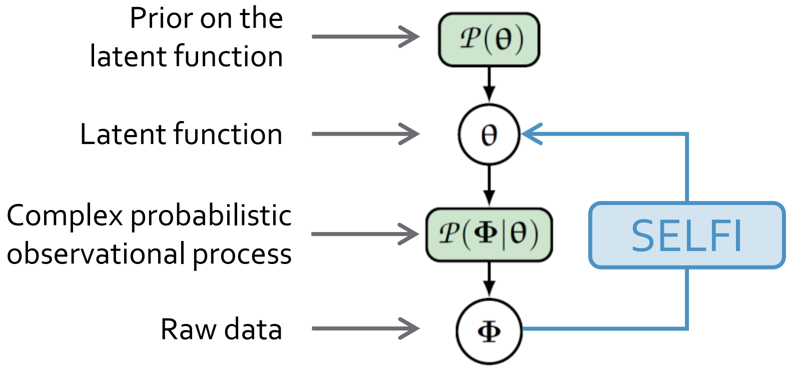
1. Inference of the latent function θ, to check for model misspecification:
   - SELFI algorithm

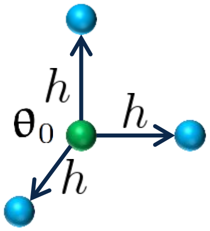
2. Simulation-based inference of ω:
   - Approximate Bayesian Computation (ABC), Likelihood-Free Rejection Sampling
   - Density/ratio estimation (DELFI / NRE)
   - Bayesian optimisation (BOLFI)
   - others…

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

Leclercq, MaxEnt 2022 proceedings (on arXiv next week)

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





Leclercq *et al.* 1902.10149

- We aim at inferring the latent function $\boldsymbol{\theta}$, which usually contains most/all of the information on $\boldsymbol{\omega}$.

  (initial power spectrum in cosmology, prey/predator population functions in ecology)

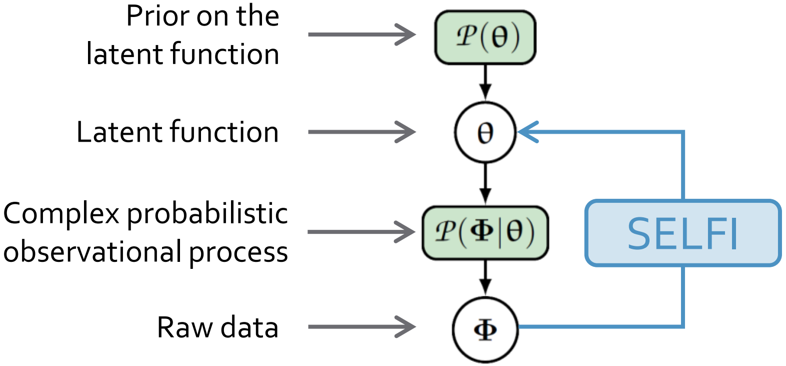
- This requires doing SBI in $d = \mathcal{O}(100) - \mathcal{O}(1,000)$
- If we trust the results of earlier experiments, we can Taylor-expand the black-box around an expansion point $\boldsymbol{\theta}_0$:

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

SELFI-2 (second order): coming soon!

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

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



- Linearisation of the black-box:

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

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

Leclercq *et al.* 1902.10149

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

expansion point — observed summaries

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

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

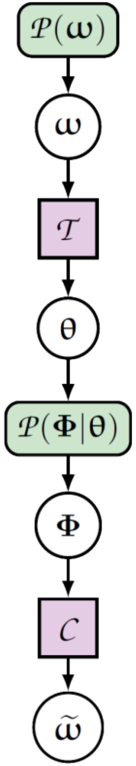
covariance of summaries — gradient of the black-box — prior covariance

- f_0, C_0 and $\nabla\mathrm{f}_0$ can be evaluated through simulations only.
- The number of required simulations is fixed *a priori* (contrary to MCMC).
- The workload is perfectly parallel.

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 $\mathcal{P}(\theta)$:

$$d_\mathrm{M}(\gamma, \theta_0|\mathbf{S}) \equiv \sqrt{(\gamma - \theta_0)^\intercal \, \mathbf{S}^{-1} (\gamma - \theta_0)}$$



- The score function $\nabla_\omega \hat{\ell}_{\omega 0}$ is the gradient of the log-likelihood at fiducial point ω_0 in parameter space.
- A quasi maximum-likelihood estimator for the parameters is

$$\mathcal{C}(\Phi) = \tilde{\omega} \equiv \omega_0 + \mathbf{F}_0^{-1} \left[(\nabla_\omega \mathbf{f}_0)^\intercal \boxed{\mathbf{C}_0^{-1}} (\Phi - \boxed{\mathbf{f}_0}) \right]$$

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

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

Already computed for SELFI

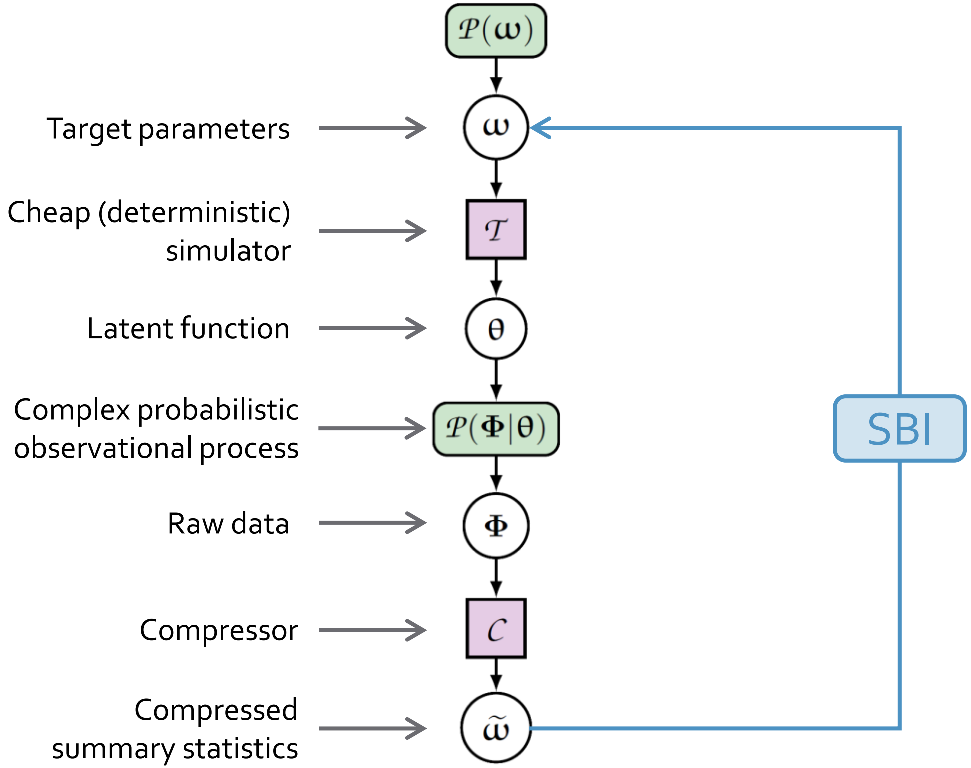
Cheap via finite differences

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

Alsing & Wandelt, 1712.00012

Leclercq, MaxEnt 2022 proceedings (on arXiv next week)

Simulation-based inference of top-level target parameters

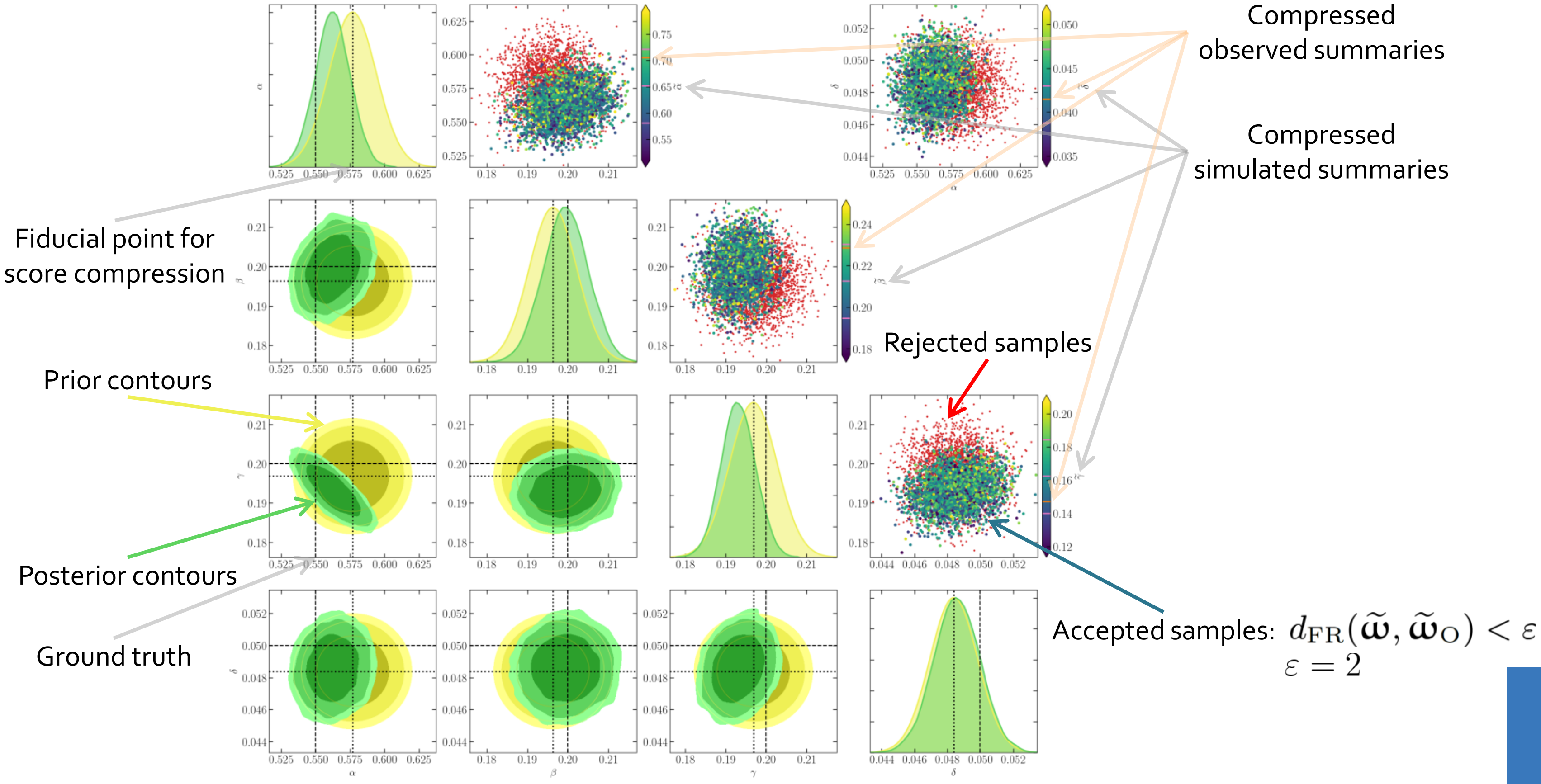


- Any SBI algorithm can be used to obtain the posterior $\mathcal{P}(\omega|\tilde{\omega}_\mathrm{O})$.
- Final inference:
  - does not depend on the assumptions made to check for model misspecification,
  - is unbiased (only more conservative) in case data compression is lossy.
- Non-parametric approaches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

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

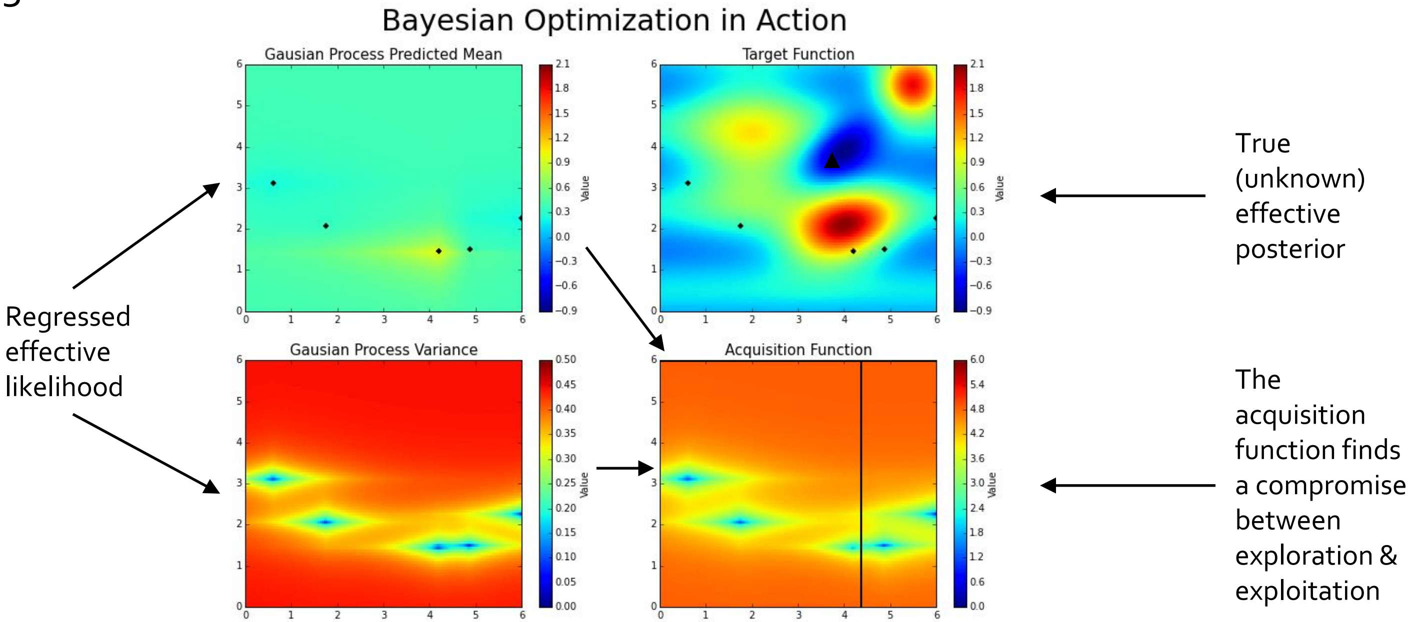
Leclercq, MaxEnt 2022 proceedings (on arXiv next week)

Prey-predator model: inference of target population parameters using Likelihood-Free Rejection Sampling



Simulation-efficient inference of target parameters: BOLFI: Data acquisition

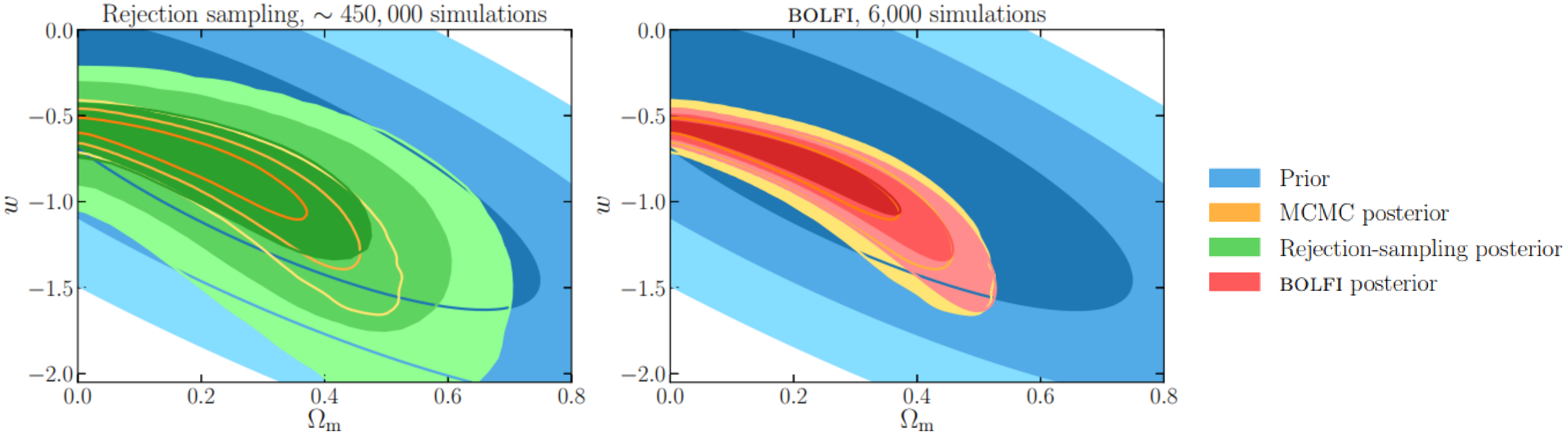
- Simulations are obtained from sampling an adaptively-constructed proposal distribution, using the regressed effective likelihood



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

Simulation-efficient inference of target parameters: 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
- Bayesian optimisation can also be applied to the "true" likelihood (if known) or to build iteratively an emulator of the data model

Leclercq 2018, 1805.07152

Conclusions

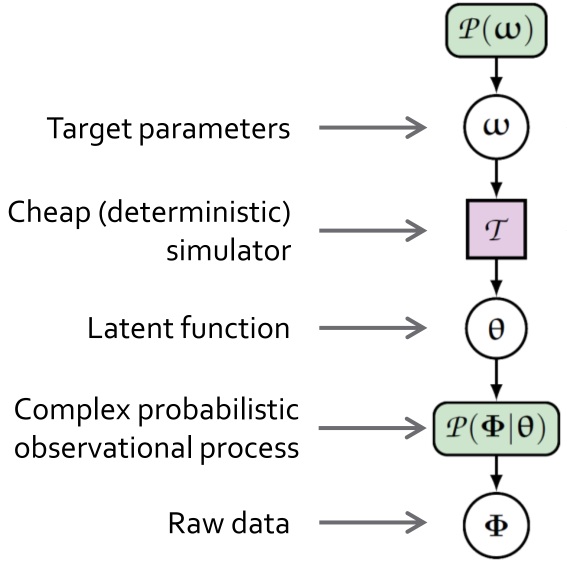
- A novel two-step simulation based Bayesian approach, combining SELFI and SBI, to tackle the issue of model misspecification for a large class of BHMs.
- Advantages of the first step (SELFI):
  - Even if the inference is in high dimension, the simulator remains a black-box.
  - The number of simulations 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/BOLFI):
  - SELFI quantities provide a score compressor for free.
  - Inference does not depend on the assumptions made to check for model misspecification.
  - General advantages of SBI with respect to likelihood-based methods are preserved.
  - BOLFI is particularly suitable for very expensive simulators.

➢ 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.

Backup slides

A prey-predator model with observational effects



$(\alpha, \beta, \gamma, \delta)$

Intrinsic reproduction rate of prey

Mortality rate of prey due to predators encountered

Reproduction rate of predators due to preys eaten

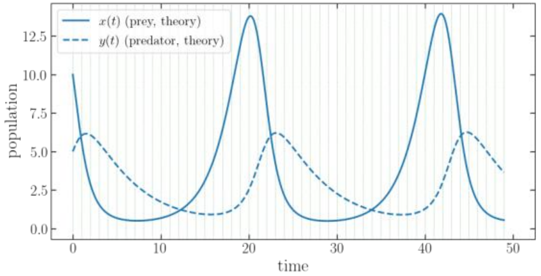
Intrinsic mortality rate of predators

Numerical integration of the Lotka-Volterra equations:

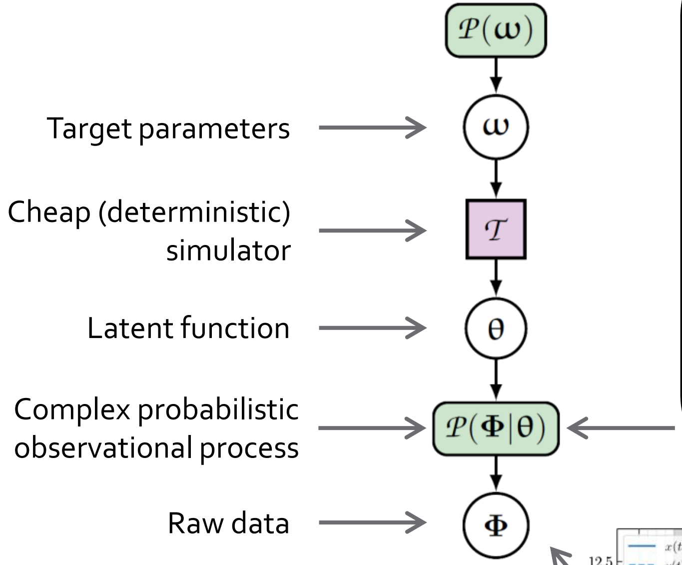
Prey: $\dfrac{\mathrm{d}x}{\mathrm{d}t} = \boxed{\alpha} x - \boxed{\beta} xy$

Predators: $\dfrac{\mathrm{d}y}{\mathrm{d}t} = \boxed{\delta} xy - \boxed{\gamma} y$

(explicit Euler method)



A prey-predator model with observational effects



- Model A (correct):
  - *Signal*: a retarded a non-linear observation of the true functions

$$\begin{cases} s_x(t_{i+1}) &= \boxed{e_x(t_i)} \; [x(t_i) - \boxed{p}x(t_i)y(t_i) + \boxed{q}x(t_i)^2] \\ s_y(t_{i+1}) &= \; y(t_i) + \boxed{p}x(t_i)y(t_i) - \boxed{q}y(t_i)^2 \end{cases}$$

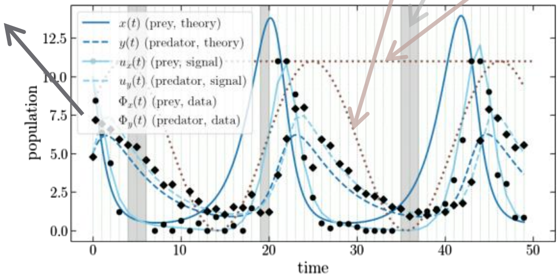
  - *Noise*: additive Gaussian noise with zero mean

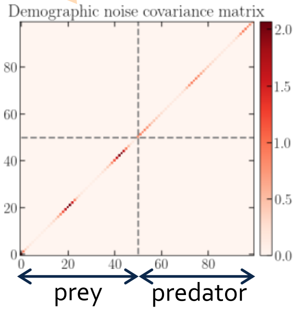
$$\begin{cases} u_x(t) = s_x(t) + n_x^{\mathrm{D}}(t) + n_x^{\mathrm{O}}(t) \\ u_y(t) = s_y(t) + n_y^{\mathrm{D}}(t) + n_y^{\mathrm{O}}(t) \end{cases}$$

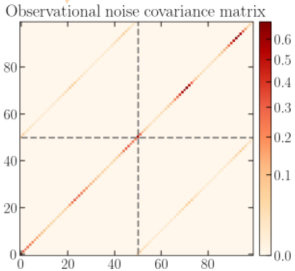
$$\begin{matrix} n_x^{\mathrm{D}}(t) \curvearrowleft \mathcal{G}\left[0, \boxed{r}x(t)\right] \\ n_y^{\mathrm{D}}(t) \curvearrowleft \mathcal{G}\left[0, \boxed{r}y(t)\right] \end{matrix} \quad \begin{pmatrix} n_x^{\mathrm{O}}(t) \\ n_y^{\mathrm{O}}(t) \end{pmatrix} \curvearrowleft \mathcal{G}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boxed{s}\begin{pmatrix} y(t) & \boxed{t}\sqrt{x(t)y(t)} \\ \boxed{t}\sqrt{x(t)y(t)} & x(t) \end{pmatrix}\right]$$

  - *Censoring*: a mask and a threshold

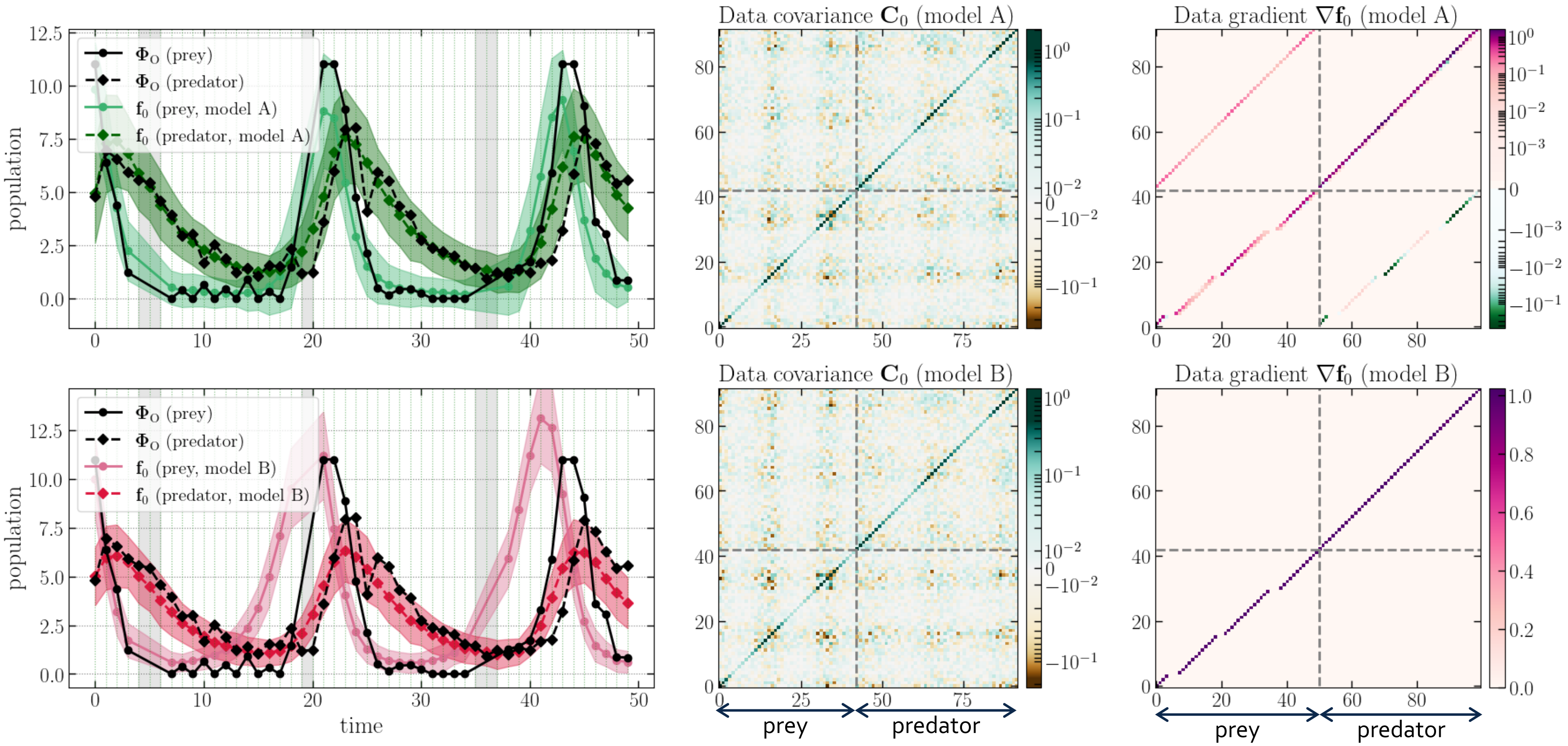
- Model B (misspecified): assumes the true functions are directly observed, does not account for observational noise nor threshold.



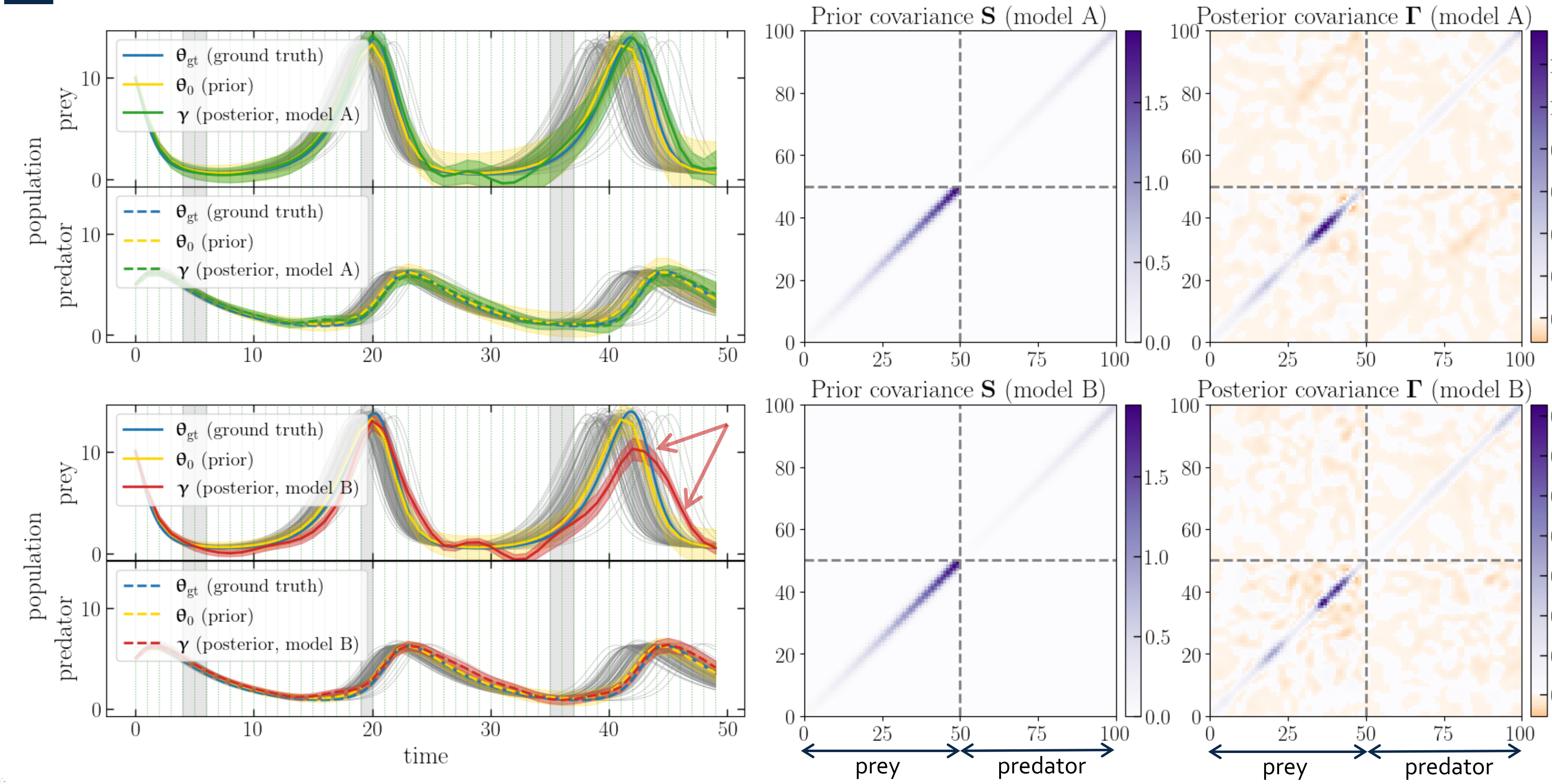




Prey-predator model: diagnostics of the linearised black-box



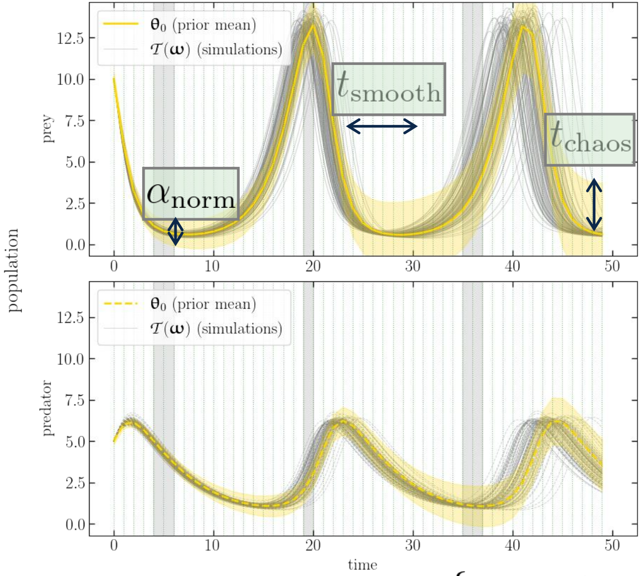
Prey-predator model: inference of the latent population functions



A family of priors for population functions in prey-predator systems

Assumptions:

1. The population functions θ are Gaussian-distributed.
2. They are strongly constrained to live close to $\theta_0 = \mathcal{T}(\omega_0)$.
3. $x(t)$ and $y(t)$ are smooth functions of time.
4. The uncertainty on $x(t)$ and $y(t)$ grows with time.



- Gaussian prior:

  mean: θ_0

  covariance: $\mathbf{S} \equiv \boxed{\alpha_{\mathrm{norm}}^2} \mathbf{K} \circ \mathbf{V}$

  Overall prior uncertainty

- The 3 free hyperparameters $\{\alpha_{\mathrm{norm}}, t_{\mathrm{smooth}}, t_{\mathrm{chaos}}\}$ can be optimised using simulations.

Smoothness of the population function

Chaotic behaviour of the system

$$(\mathbf{K}_z)_{ij} \equiv \left[-\frac{1}{2} \left(\frac{t_i - t_j}{\boxed{t_{\mathrm{smooth}}}} \right)^2 \right] \qquad \mathbf{K} \equiv \begin{pmatrix} \mathbf{K}_x & 0 \\ 0 & \mathbf{K}_y \end{pmatrix} \qquad \mathbf{V} \equiv \begin{pmatrix} x_0\, \mathbf{u}\mathbf{u}^\intercal & 0 \\ 0 & y_0\, \mathbf{u}\mathbf{u}^\intercal \end{pmatrix} \qquad (\mathbf{u})_i \equiv 1 + \frac{t_i}{\boxed{t_{\mathrm{chaos}}}}$$