physical sciences forum

MDPI

Proceeding Paper

Simulation-Based Inference of Bayesian Hierarchical Models While Checking for Model Misspecification †

Florent Leclercq

CNRS & Sorbonne Université, UMR 7095, Institut d'Astrophysique de Paris, 98 bis boulevard Arago, F-75014 Paris, France; florent.leclercq@iap.fr

† Presented at the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris, France, 18–22 July 2022.

Abstract: This paper presents recent methodological advances for performing simulation-based inference (SBI) of a general class of Bayesian hierarchical models (BHMs) while checking for model misspecification. Our approach is based on a two-step framework. First, the latent function that appears as a second layer of the BHM is inferred and used to diagnose possible model misspecification. Second, target parameters of the trusted model are inferred via SBI. Simulations used in the first step are recycled for score compression, which is necessary for the second step. As a proof of concept, we apply our framework to a prey–predator model built upon the Lotka–Volterra equations and involving complex observational processes.

Keywords: Bayesian inference; Bayesian hierarchical models; simulation-based inference

Citation: Leclercq, F. Simulation-Based Inference of Bayesian Hierarchical Models While Checking for Model Misspecification. *Phys. Sci. Forum* **2022**, *5*, 4. https://doi.org/10.3390/psf2022005004

Academic Editors: Frédéric Barbaresco, Ali Mohammad-Djafari, Frank Nielsen and Martino Trassinelli

Published: 2 November 2022

1. Introduction

Model misspecification is a long-standing problem for Bayesian inference: when the model differs from the actual data-generating process, posteriors tend to be biased and/or overly concentrated. In this paper, we are interested the problem of model misspecification for a particular, but common, class of Bayesian hierarchical models (BHMs): those that involve a latent function, such as the primordial power spectrum in cosmology (e.g., [1]) or the population model in genetics (e.g., [2]).

Simulation-based inference (SBI) only provides the posterior of top-level target parameters and marginalizes over all other latent variables of the BHM. Alone, it is therefore unable to diagnose whether the model is misspecified. Key insights regarding the issue of model misspecification can usually be obtained from the posterior distribution of the latent function, as there often exists an independent theoretical understanding of its values. An approximate posterior for the latent function (a much higher-dimensional quantity than the target vector of parameters) can be obtained using SELFI (simulator expansion for likelihood-free inference, [1]), an approach based on the likelihood of an alternative parametric model, constructed by linearizing model predictions around an expansion point.

This paper presents a framework that combines SELFI and SBI while recycling the necessary simulations. The simulator is first linearized to obtain the SELFI posterior of the latent function. Next, the same simulations are used for data compression to the score function (the gradient of the log-likelihood with respect to the parameters), and the final SBI posterior of target parameters is obtained.

2. Method

2.1. Bayesian Hierarchical Models with a Latent Function

In this paper, we assume a given BHM consisting of the following variables: $\boldsymbol{\omega} \in \mathbb{R}^N$ (vector of N target parameters), $\boldsymbol{\theta} \in \mathbb{R}^S$ (vector containing the values of the latent function θ at S support points), $\boldsymbol{\Phi} \in \mathbb{R}^P$ (data vector of P components), and $\widetilde{\boldsymbol{\omega}} \in \mathbb{R}^N$

(compressed data vector of size N). We typically expect $N \sim \mathcal{O}(5-10)$ target parameters, $S \sim \mathcal{O}(10^2 - 10^3)$ support points; P can be any number and as large as $\mathcal{O}(10^7)$ for complex data models. We further assume that $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ are linked by a deterministic function $\mathcal{T}$, usually theoretically well-understood and numerically cheap. Therefore, the expensive and potentially misspecified part of the BHM is the probabilistic simulator linking the latent function $\boldsymbol{\theta}$ to the data $\boldsymbol{\Phi}$, $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$. The deterministic compression step $\mathcal{C}$ linking $\boldsymbol{\Phi}$ to $\widetilde{\boldsymbol{\omega}}$ is discussed in Section 2.4.

2.2. Latent Function Inference with SELFI

The first part of the framework proposed in this paper is to infer the latent function $\boldsymbol{\theta}$ conditional on observed data $\boldsymbol{\Phi}_\mathrm{O}$. This is an inference problem in high dimension (S, the number of support points for the latent function θ), which means that usual SBI frameworks, allowing a general exploration of parameter space, will fail and that stronger assumptions are required. SELFI [1] relies upon the simplification of the inference problem around an expansion point $\boldsymbol{\theta}_0$.

The first assumption is a Taylor expansion (linearization) of the mean data model around $\boldsymbol{\theta}_0$. Namely, if $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}} \equiv \mathrm{E}[\boldsymbol{\Phi}_{\boldsymbol{\theta}}]$ is the expectation value of $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$, where $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ are simulations of $\boldsymbol{\Phi}$ given $\boldsymbol{\theta}$ (i.e., $\boldsymbol{\Phi}_{\boldsymbol{\theta}} \curvearrowleft \mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$), we assume that

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}} \approx \mathbf{f}_0 + \nabla \mathbf{f}_0 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \equiv \mathbf{f}(\boldsymbol{\theta}), \tag{1}$$

where $\mathbf{f}_0 \equiv \hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}_0}$ is the mean data model at the expansion point $\boldsymbol{\theta}_0$, and $\nabla \mathbf{f}_0$ is the gradient of $\mathbf{f}_0$ at the expansion point (for simplification, we note $\nabla \mathbf{f}_0 = \nabla_{\boldsymbol{\theta}} \mathbf{f}_0$, where the gradient is taken with respect to $\boldsymbol{\theta}$). The second assumption is that the (true) implicit likelihood of the problem is replaced by a Gaussian effective likelihood: $\mathcal{P}(\boldsymbol{\Phi}_\mathrm{O}|\boldsymbol{\theta}) \equiv \exp\left[\hat{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right]$ with

$$-2\hat{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \approx \log|2\pi\mathbf{C}_0| + [\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}(\boldsymbol{\theta})]^\intercal \mathbf{C}_0^{-1} [\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}(\boldsymbol{\theta})], \tag{2}$$

where $\mathbf{C}_0$ is the data covariance matrix at the expansion point $\boldsymbol{\theta}_0$.

The SELFI framework is fully characterized by $\mathbf{f}_0$, $\mathbf{C}_0$, and $\nabla \mathbf{f}_0$, which, if unknown, can be evaluated through forward simulations only. The numerical computation requires N_0 simulations at the expansion point (to evaluate the empirical mean $\mathbf{f}_0$ and empirical covariance matrix $\mathbf{C}_0$), and N_s simulations in each direction of parameter space (to evaluate the empirical gradient $\nabla \mathbf{f}_0$ via first-order forward finite differences). The total is $N_0 + N_s \times S$ simulations; N_0 and N_s should be of the order of the dimensionality of the data space P, giving a total cost of $\mathcal{O}(\gtrsim P(S+1))$ model evaluations.

To fully characterize the Bayesian problem, one requires a prior on $\boldsymbol{\theta}$, $\mathcal{P}(\boldsymbol{\theta})$. Any prior can be used if one is ready to use numerical techniques to explore the posterior (such as standard Markov Chain Monte Carlo), using the linearized data model and Gaussian effective likelihood. However, a remarkable analytic result with SELFI is that, if the prior is Gaussian with a mean equal to the expansion point $\boldsymbol{\theta}_0$, i.e.,

$$-2\log \mathcal{P}(\boldsymbol{\theta}) \equiv \log|2\pi\mathbf{S}| + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \mathbf{S}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{3}$$

then the effective posterior is also Gaussian:

$$-2\log \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi}_\mathrm{O}) \approx \log|2\pi\boldsymbol{\Gamma}| + (\boldsymbol{\theta} - \boldsymbol{\gamma})^\intercal \boldsymbol{\Gamma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\gamma}). \tag{4}$$

The posterior mean and covariance matrix are given by

$$\boldsymbol{\gamma} \equiv \boldsymbol{\theta}_0 + \boldsymbol{\Gamma} (\nabla \mathbf{f}_0)^\intercal \mathbf{C}_0^{-1} (\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0), \tag{5}$$

$$\boldsymbol{\Gamma} \equiv \left[(\nabla \mathbf{f}_0)^\intercal \mathbf{C}_0^{-1} \nabla \mathbf{f}_0 + \mathbf{S}^{-1}\right]^{-1} \tag{6}$$

(see [1] Appendix B, for a derivation). They are fully characterized by the expansion variables $\boldsymbol{\theta}_0$, $\mathbf{f}_0$, $\mathbf{C}_0$, and $\nabla \mathbf{f}_0$, as well as the prior covariance matrix $\mathbf{S}$.

2.3. Check for Model Misspecification

The SELFI posterior can be used as a check for model misspecification. Visually checking the reconstructed $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ can yield interesting insights, especially if the latent function has some properties (such as an expected shape, periodicity, etc.) to which the data model may be sensitive if misspecified (see Section 4.2).

If a quantitative check for model misspecification is desired, we propose using the Mahalanobis distance between the reconstruction $\boldsymbol{\gamma}$ and the prior distribution $\mathcal{P}(\boldsymbol{\theta})$, defined formally by

$$d_{\mathrm{M}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0|\mathbf{S}) \equiv \sqrt{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\intercal} \mathbf{S}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}. \tag{7}$$

The value of $d_{\mathrm{M}}(\boldsymbol{\gamma}, \boldsymbol{\theta}_0|\mathbf{S})$ for the SELFI posterior mean $\boldsymbol{\gamma}$ can be compared to an ensemble of values of $d_{\mathrm{M}}(\boldsymbol{\theta}_{\boldsymbol{\omega}}, \boldsymbol{\theta}_0|\mathbf{S})$ for simulated latent functions $\boldsymbol{\theta}_{\boldsymbol{\omega}} = \mathcal{T}(\boldsymbol{\omega})$, where samples $\boldsymbol{\omega}$ are drawn from the prior $\mathcal{P}(\boldsymbol{\omega})$.

2.4. Score Compression and Simulation-Based Inference

Having checked the BHM for model misspecification, we now address the second part of the framework, aimed at inferring top-level parameters $\boldsymbol{\omega}$ given observations. SBI is known to be difficult when the dimensionality of the data space P is high. For this reason, data compression is usually necessary. Data compression can be thought of as an additional layer at the bottom of the BHM, made of a deterministic function $\mathcal{C}$ acting on $\boldsymbol{\Phi}$. In practical scenarios, data compression shall preserve as much information about $\boldsymbol{\omega}$ as possible, meaning that compressed summaries $\mathcal{C}(\boldsymbol{\Phi})$ shall be as close as possible to sufficient summary statistics of $\boldsymbol{\Phi}$, i.e., $\mathcal{P}(\boldsymbol{\omega}|\mathcal{C}(\boldsymbol{\Phi})) = \mathcal{P}(\boldsymbol{\omega}|\boldsymbol{\Phi})$.

Here, we propose to use score compression [3]. We make the assumption (for compression only, not for later inference) that $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\omega})$ is Gaussian distributed: $\mathcal{P}(\boldsymbol{\Phi}_{\mathrm{O}}|\boldsymbol{\omega}) \equiv \exp\left[\hat{\ell}_{\boldsymbol{\omega}}(\boldsymbol{\omega})\right]$ where $\hat{\ell}_{\boldsymbol{\omega}}(\boldsymbol{\omega}) = \hat{\ell}_{\boldsymbol{\theta}}(\mathcal{T}(\boldsymbol{\omega}))$ (see Equation (2)). The score function $\nabla_{\boldsymbol{\omega}} \hat{\ell}_{\boldsymbol{\omega}0}$ is the gradient of this log-likelihood with respect to the parameters $\boldsymbol{\omega}$ at a fiducial point $\boldsymbol{\omega}_0$ in parameter space. Using as fiducial point the values that generate the SELFI expansion point (i.e., $\boldsymbol{\omega}_0$ such that $\boldsymbol{\theta}_0 = \mathcal{T}(\boldsymbol{\omega}_0)$), a quasi maximum-likelihood estimator for the parameters is $\widetilde{\boldsymbol{\omega}}_{\mathrm{O}} \equiv \boldsymbol{\omega}_0 + \mathbf{F}_0^{-1} \nabla_{\boldsymbol{\omega}} \hat{\ell}_{\boldsymbol{\omega}0}$, where the Fisher matrix $\mathbf{F}_0$ and the gradient of the log-likelihood are evaluated at $\boldsymbol{\omega}_0$. Compression of $\boldsymbol{\Phi}_{\mathrm{O}}$ to $\widetilde{\boldsymbol{\omega}}_{\mathrm{O}}$ yields N compressed statistics that are optimal in the sense that they preserve the Fisher information content of the data [3].

In our case, the covariance matrix $\mathbf{C}_0$ is assumed not to depend on parameters ($\nabla_{\boldsymbol{\omega}} \mathbf{C}_0 = 0$), and the expression for $\mathcal{C}(\boldsymbol{\Phi})$ is therefore

$$\mathcal{C}(\boldsymbol{\Phi}) = \widetilde{\boldsymbol{\omega}} \equiv \boldsymbol{\omega}_0 + \mathbf{F}_0^{-1} \left[(\nabla_{\boldsymbol{\omega}} \mathbf{f}_0)^{\intercal} \mathbf{C}_0^{-1} (\boldsymbol{\Phi} - \mathbf{f}_0) \right]. \tag{8}$$

The Fisher matrix of the problem further takes a simple form:

$$\mathbf{F}_0 \equiv -\mathrm{E}\left[\nabla_{\boldsymbol{\omega}} \nabla_{\boldsymbol{\omega}} \hat{\ell}_{\boldsymbol{\omega}0}(\boldsymbol{\omega}) \right] = (\nabla_{\boldsymbol{\omega}} \mathbf{f}_0)^{\intercal} \mathbf{C}_0^{-1} \nabla_{\boldsymbol{\omega}} \mathbf{f}_0. \tag{9}$$

We therefore need to evaluate

$$\nabla_{\boldsymbol{\omega}} \mathbf{f}_0 = \nabla \mathbf{f}_0 \cdot \left. \frac{\partial \mathcal{T}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \right|_{\boldsymbol{\omega} = \boldsymbol{\omega}_0}. \tag{10}$$

Importantly, in Equations (8)–(10), $\mathbf{C}_0$ and $\nabla \mathbf{f}_0$ have already been computed for latent function inference with SELFI. The only missing quantity is the second matrix in the right-hand side of Equation (10), that is, $\nabla_{\boldsymbol{\omega}} \mathcal{T}_0$, the gradient of $\mathcal{T}$ evaluated at $\boldsymbol{\omega}_0$. If unknown, its computation (e.g., via finite differences) does not require any more simulation of $\boldsymbol{\Phi}$. It is usually easy, as there are only N directions in parameter space and $\mathcal{T}$ is the numerically cheap part of the BHM. We note that, because we have to calculate $\mathbf{F}_0$, we can

easily get the Fisher–Rao distance between any simulated summaries $\widetilde{\boldsymbol{\omega}}$ and the observed summaries $\widetilde{\boldsymbol{\omega}}_{\mathrm{O}}$,

$$d_{\mathrm{FR}}(\widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\omega}}_{\mathrm{O}}) \equiv \sqrt{(\widetilde{\boldsymbol{\omega}} - \widetilde{\boldsymbol{\omega}}_{\mathrm{O}})^{\mathsf{T}} \mathbf{F}_0 (\widetilde{\boldsymbol{\omega}} - \widetilde{\boldsymbol{\omega}}_{\mathrm{O}})}, \tag{11}$$

which can be used by any non-parametric SBI method.

We specify a prior $\mathcal{P}(\boldsymbol{\omega})$ (typically peaking at or centered on $\boldsymbol{\omega}_0$, for consistency with the assumptions made for data compression). Having defined $\mathcal{C}$, we now have a full BHM that maps $\boldsymbol{\omega}$ (of dimension N) to compressed summaries $\widetilde{\boldsymbol{\omega}}$ (of size N) and has been checked for model misspecification for the part linking $\boldsymbol{\theta}$ to $\boldsymbol{\Phi}$. We can then proceed with SBI via usual techniques. These can include likelihood-free rejection sampling, but also more sophisticated techniques such as DELFI (e.g., [4,5]) or BOLFI (e.g., [6–8]).

3. Lotka–Volterra BHM

3.1. Lotka–Volterra Solver

The Lotka–Volterra equations describe the dynamics of an ecological system in which two species interact, as a pair of first-order non-linear differential equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha x - \beta xy, \tag{12}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \delta xy - \gamma y. \tag{13}$$

where $x(t)$ is the number of prey at time t, and $y(t)$ is the number of predators at time t. The model is characterized by $\boldsymbol{\omega} = (\alpha, \beta, \gamma, \delta)$, a vector of four real parameters describing the interaction of the two species.

The initial conditions of the problem $\{x(0), y(0)\} = \{x_0, y_0\}$ are assumed to be exactly known. Throughout the paper, timestepping and number of timesteps are fixed: $t_i = i\Delta t$ for $i \in [\![0, S/2]\!]$.

The expression $\mathcal{T}$ is an algorithm that numerically solves the ordinary differential equations. For simplicity, we choose an explicit Euler method: for all $i \in [\![0, S/2 - 1]\!]$,

$$x(t_{i+1}) = x(t_i) \times [1 + \alpha - \beta y(t_i)] \times \Delta_t, \tag{14}$$

$$y(t_{i+1}) = y(t_i) \times [1 + \delta x(t_i) - \gamma] \times \Delta_t. \tag{15}$$

The latent function $\theta(t)$ is a concatenation of $x(t)$ and $y(t)$ evaluated at the timesteps of the problem. The corresponding vector is $\boldsymbol{\theta} \equiv \left\{ \{x(t_i)\}_{0 \le i < S/2}, \{y(t_i)\}_{0 \le i < S/2} \right\}$ of size S.

3.2. Lotka–Volterra Observer

3.2.1. Full Data Model

To go from $\boldsymbol{\theta}$ to $\boldsymbol{\Phi}$, we assume a complex, probabilistic observational process of prey and predator populations, later referred to as "model A" and defined as follows.

Signal. The (unobserved) signal s_z is a delayed and non-linearly perturbed observation of the true population function for species $z \in \{x, y\}$, modulated by some seasonal efficiency $e_z(t)$. Formally, $s_x(0) = x_0$, $s_y(0) = y_0$, and for $i \in [\![0, S/2 - 1]\!]$,

$$s_x(t_{i+1}) = e_x(t_i) \Big[x(t_i) - p x(t_i) y(t_i) + q x(t_i)^2 \Big], \tag{16}$$

$$s_y(t_{i+1}) = e_y(t_i) \Big[y(t_i) + p x(t_i) y(t_i) - q y(t_i)^2 \Big]. \tag{17}$$

These equations involve two parameters: p accounts for hunts between t_i and t_{i+1} (temporarily making prey more likely to hide and predators more likely to be visible), and q accounts for the gregariousness of prey and independence of predators. The free functions $e_x(t)$ and $e_y(t)$, valued in $[0, 1]$, describe how prey and predators are likely to be detectable at any time, accounting, for example, for seasonal variation (hibernation, migration).

Noise. The signal s_z is subject to additive noise, giving a noisy signal $u_z(t) = s_z(t) + n_z^{\mathrm{D}}(t) + n_z^{\mathrm{O}}(t)$, where the noise has two components:

- Demographic Gaussian noise with zero mean and variance proportional to the true underlying population, i.e., $n_x^{\mathrm{D}}(t) \curvearrowleft \mathcal{G}[0, rx(t)]$ and $n_y^{\mathrm{D}}(t) \curvearrowleft \mathcal{G}[0, ry(t)]$. The parameter r gives the strength of demographic noise.
- Observational Gaussian noise that accounts for observer efficiency, coupling prey and predators such that

$$\begin{pmatrix} n_x^{\mathrm{O}}(t) \\ n_y^{\mathrm{O}}(t) \end{pmatrix} \curvearrowleft \mathcal{G}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, s\begin{pmatrix} y(t) & t\sqrt{x(t)y(t)} \\ t\sqrt{x(t)y(t)} & x(t) \end{pmatrix}\right]. \tag{18}$$

  The parameter s gives the overall amplitude of observational noise, and the parameter t controls the strength of the non-diagonal component (it should be chosen such that the covariance matrix appearing in Equation (18) is positive semi-definite).

Censoring. Finally, observed data are a censored and thresholded version of the noisy signal: for each timestep t_i, $\Phi_z(t_i) = m_z(t_i) \times \min[u_z(t_i), M_z]$, where M_z is the maximum number of prey or predators that can be detected by the observer, and m_z is a mask (taking either the value 0 or 1). Masked data points are discarded. The data vector is $\boldsymbol{\Phi} = \left\{\{\Phi_x(t_i)\}, \{\Phi_y(t_i)\}\right\}$. It contains $P \leq S$ elements depending on the number of masked timesteps for each species z (formally, $P = \sum_{i=0}^{S/2-1}\left(\delta_{\mathrm{K}}^{m_x(t_i),1} + \delta_{\mathrm{K}}^{m_y(t_i),1}\right)$, where δ_{K} is a Kronecker delta symbol).

All of the free parameters (p, q, r, s, t, M_x, M_y) and free functions ($e_x(t)$, $e_y(t)$, $m_x(t)$, $m_y(t)$) appearing in the Lotka–Volterra observer data model described in this section are assumed known and fixed throughout the paper. Parameters used are $x_0 = 10$, $y_0 = 5$, $p = 0.05$, $q = 0.01$, $r = 0.15$, $s = 0.05$, $t = 0.2$.

3.2.2. Simplified Data Model

In this section, we introduce "model B", a simplified (misspecified) data model linking $\boldsymbol{\theta}$ to $\boldsymbol{\Phi}$. Model B assumes that underlying functions are directly observed, i.e., $s_z(t) = z(t)$. It omits observational noise, such that $u_z(t) = s_z(t) + n_z^{\mathrm{D}}(t)$. In model B, parameters p, q, s, and t are not involved, and the value of r (strength of demographic noise) can be incorrect (we used $r = 0.105$). Finally, model B fails to account for the thresholds: $\Phi_z(t) = m_z(t)u_z(t)$.

4. Results

In this section, we apply the two-step inference method described in Section 2 to the Lotka–Volterra BHM introduced in Section 3. We generate mock data $\boldsymbol{\Phi}_{\mathrm{O}}$ from model A, using ground truth parameters $\boldsymbol{\omega}_{\mathrm{gt}} = (\alpha_{\mathrm{gt}}, \beta_{\mathrm{gt}}, \gamma_{\mathrm{gt}}, \delta_{\mathrm{gt}}) = (0.55, 0.2, 0.2, 0.05)$. We assume that ground truth parameters are known a priori with a precision of approximately 3%. Consistently, we choose a Gaussian prior $\mathcal{P}(\boldsymbol{\omega})$ with mean $\boldsymbol{\omega}_0 = (0.5768, 0.1963, 0.1968, 0.0484)$ and diagonal covariance matrix $\mathrm{diag}(0.0173^2, 0.0059^2, 0.0059^2, 0.0015^2)$.

4.1. Inference of Population Functions with SELFI

We first seek to reconstruct the latent population functions $x(t)$ and $y(t)$, conditional on the data $\boldsymbol{\Phi}_{\mathrm{O}}$, using SELFI. We choose as an expansion point the population functions simulated from the mean of the prior on $\boldsymbol{\omega}$, i.e., $\boldsymbol{\theta}_0 = \mathcal{T}(\boldsymbol{\omega}_0)$. We use $N_0 = 150$ and $N_s = 100$; the computational workload is therefore a fixed number of 10,150 simulations for each model. It is known a priori and perfectly parallel.

We adopt a Gaussian prior $\mathcal{P}(\boldsymbol{\theta})$ and combine it with the effective likelihood to obtain the SELFI effective posterior $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}})$. Figure 1 (left panels) shows the inferred population functions $\boldsymbol{\gamma}$ in comparison with the prior mean and expansion point $\boldsymbol{\theta}_0$ and the ground truth $\boldsymbol{\theta}_{\mathrm{gt}}$. The figure shows 2σ credible regions for the prior and the posterior (i.e., $2\sqrt{\mathrm{diag}(\mathbf{S})}$

and $2\sqrt{\mathrm{diag}(\boldsymbol{\Gamma})}$, respectively). The full posterior covariance matrix $\boldsymbol{\Gamma}$ for each model is shown in the rightmost column of Figure 1.

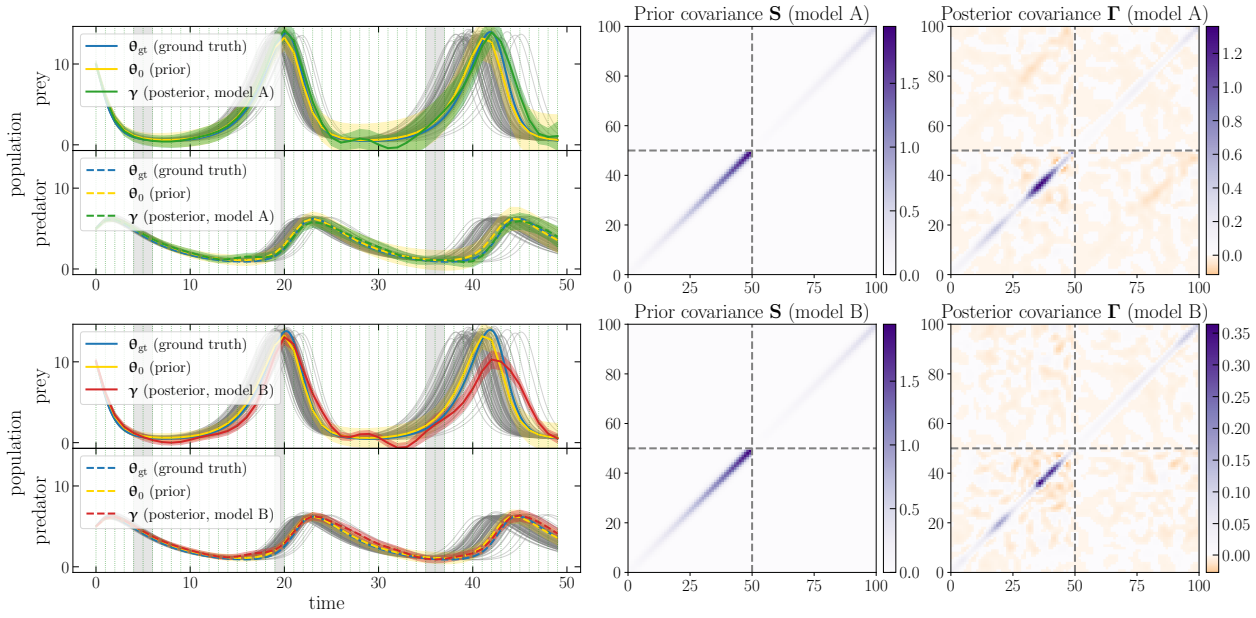

Figure 1. SELFI inference of the population function θ given the observed data $\boldsymbol{\Phi}_\mathrm{O}$, used as a check for model misspecification. **Left panels**: the prior mean and expansion point θ_0 and the effective posterior mean $\boldsymbol{\gamma}$ are represented as yellow and green/red lines, respectively, with their 2σ credible intervals. For comparison, simulations $\mathcal{T}(\boldsymbol{\omega})$ with $\boldsymbol{\omega} \curvearrowleft \mathcal{P}(\boldsymbol{\omega})$, and the ground truth θ_gt are shown in grey and blue, respectively. **Middle and right panels**: the prior covariance matrix $\mathbf{S}$ and the posterior covariance matrix $\boldsymbol{\Gamma}$, respectively. The first row corresponds to model A (see Section 3.2.1) and the second row to model B (see Section 3.2.2).

4.2. Check for Model Misspecification

The inferred population functions allow us to check for model misspecification. From Figure 1, it is clear that model B fails to produce a plausible reconstruction of population functions: model B breaks the (pseudo-)periodicity of the predator population function $y(t)$, which is a property required by the model. In the bottom left-hand panels, the red lines differ in shape from fiducial functions $\mathcal{T}(\boldsymbol{\omega})$ (grey lines), and the credible intervals exclude the expansion point. On the contrary, with model A, the reconstructed population functions are consistent with the expansion point. The inference is unbiased, as the ground truth typically lies within the 2σ credible region of the reconstruction.

As a quantitative check, we compute the Mahalanobis distance between $\boldsymbol{\gamma}$ and $\mathcal{P}(\theta)$ (Equation (7)) for each model. We find that $d_\mathrm{M}(\boldsymbol{\gamma}, \theta_0|\mathbf{S})$ is much smaller for model A than for model B (5.35 versus 12.54). The numbers can be compared to the empirical mean among our set of fiducial populations functions, $\langle d_\mathrm{M}(\mathcal{T}(\boldsymbol{\omega}_n), \theta_0|\mathbf{S})\rangle = 9.43$.

At this stage, we therefore consider that model B is excluded, and we proceed further with model A.

4.3. Score Compression

As $\mathcal{T}$ is numerically cheap, we get $\nabla_{\boldsymbol{\omega}}\mathcal{T}_0$ via sixth-order central finite differences around $\boldsymbol{\omega}_0$, then obtain $\nabla_{\boldsymbol{\omega}}\mathbf{f}_0$ using Equation (10). This does not require any further evaluation of the data model $\mathcal{P}(\boldsymbol{\Phi}|\theta)$, as $\nabla\mathbf{f}_0$ has already been computed.

Using Equations (8) and (9), we compress $\boldsymbol{\Phi}_\mathrm{O}$ and obtain $\tilde{\boldsymbol{\omega}}_\mathrm{O} = (0.7050, 0.2287, 0.1471, 0.0415)$.

4.4. Inference of Parameters Using Likelihood-Free Rejection Sampling

As a last step, we infer top-level parameters $\boldsymbol{\omega}$ given compressed summaries $\tilde{\boldsymbol{\omega}}_{\mathrm{O}}$. As the problem studied in this paper is sufficiently simple, we rely on the simplest solution for SBI, namely likelihood-free rejection sampling (sometimes also known as approximate Bayesian computation, e.g., [9]). To do so, we use the Fisher–Rao distance between simulated $\tilde{\boldsymbol{\omega}}$ and observed $\tilde{\boldsymbol{\omega}}_{\mathrm{O}}$, which comes naturally from score compression (see Equation (11)), and we set a threshold $\varepsilon = 2$. We draw samples from the prior $\mathcal{P}(\boldsymbol{\omega})$, simulate $\tilde{\boldsymbol{\omega}}$, then accept $\boldsymbol{\omega}$ as a sample of $\mathcal{P}(\boldsymbol{\omega}|\tilde{\boldsymbol{\omega}}_{\mathrm{O}})$ if $d_{\mathrm{FR}}(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}_{\mathrm{O}}) < \varepsilon$, and reject it otherwise.

In Figure 2, we find that the inference of top-level parameters is unbiased, with the ground truth $\boldsymbol{\omega}_{\mathrm{gt}}$ (dashed lines) lying within the 2σ credible region of the posterior. We observe that the data correctly drive some features that are not built into the prior, for instance, the degeneracy between α and γ, respectively, the reproduction rate of prey and the mortality rate of predators.

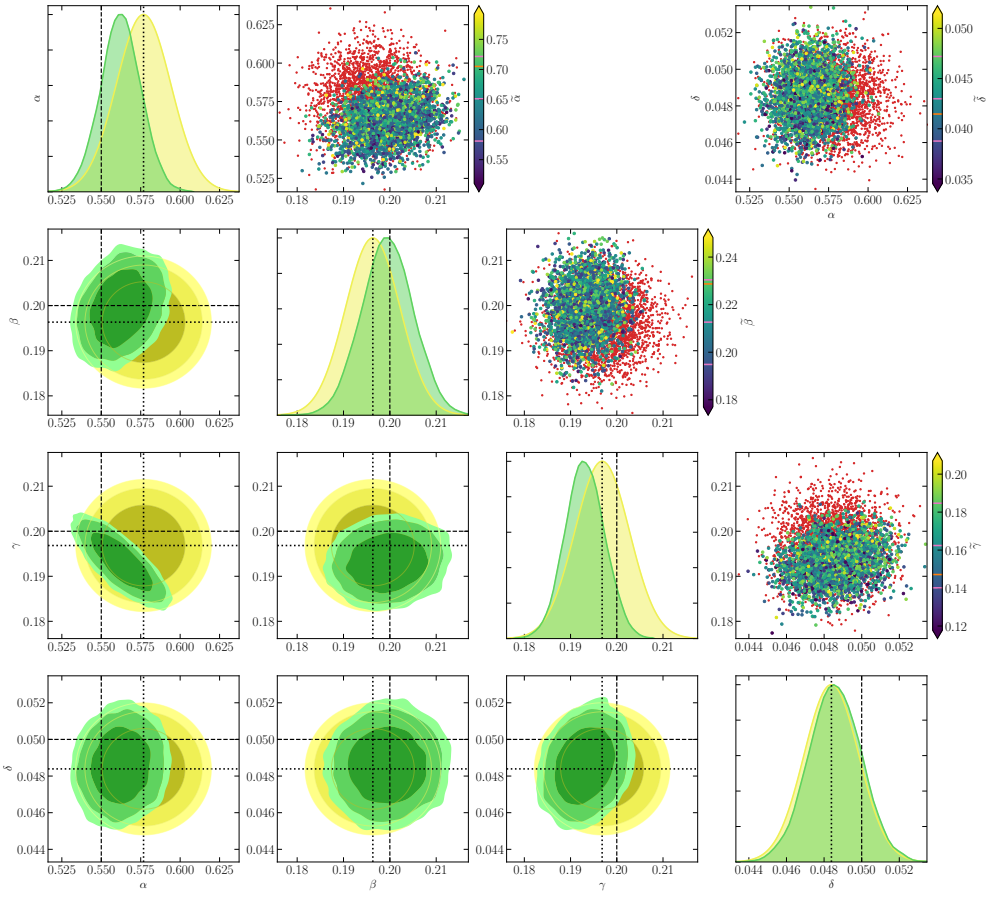

Figure 2. Simulation-based inference of the Lotka–Volterra parameters $\boldsymbol{\omega} = (\alpha, \beta, \gamma, \delta)$ given the compressed observed data $\tilde{\boldsymbol{\omega}}_{\mathrm{O}}$. Plots in the lower corner show two-dimensional marginals of the prior $\mathcal{P}(\boldsymbol{\omega})$ (yellow contours) and of the SBI posterior $\mathcal{P}(\boldsymbol{\omega}|\tilde{\boldsymbol{\omega}}_{\mathrm{O}})$ (green contours), using a threshold $\varepsilon = 2$ on the Fisher–Rao distance between simulated $\tilde{\boldsymbol{\omega}}$ and observed $\tilde{\boldsymbol{\omega}}_{\mathrm{O}}$, $d_{\mathrm{FR}}(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}_{\mathrm{O}})$. Contours show 1, 2, and 3σ credible regions. Plots on the diagonal show one-dimensional marginal distributions of the parameters, using the same color scheme. Dotted and dashed lines denote the position of the fiducial point for score compression $\boldsymbol{\omega}_0$ and of the ground truth parameters $\boldsymbol{\omega}_{\mathrm{gt}}$, respectively. The scatter plots in the upper corner illustrate score compression for pairs of parameters. There, red dots represent some simulated samples. Larger dots show some accepted samples (i.e., for which $d_{\mathrm{FR}}(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}_{\mathrm{O}}) < \varepsilon$), with a color map corresponding to the value of one component of $\tilde{\boldsymbol{\omega}}$. In the color bars, pink lines denote the mean and 1σ scatter among accepted samples of the component of $\tilde{\boldsymbol{\omega}}$, and the orange line denotes its value in $\tilde{\boldsymbol{\omega}}_{\mathrm{O}}$.

5. Conclusions

One of the biggest challenges in statistical data analysis is checking data models for misspecification, so as to obtain meaningful parameter inferences. In this work, we described a novel two-step simulation-based Bayesian approach, combining SELFI and SBI, which can be used to tackle this issue for a large class of models. BHMs to which the approach can be applied involve a latent function depending on parameters and observed through a complex probabilistic process. They are ubiquitous, e.g., in astrophysics and ecology.

In this paper, we introduced a prey–predator model, consisting of a numerical solver of the Lotka–Volterra system of equations and of a complex observational process of population functions. As a proof of concept, we applied our technique to this model and to a simplified (misspecified) version of it. We demonstrated successful identification of the misspecified model and unbiased inference of the parameters of the correct model.

In conclusion, the method developed constitutes a computationally efficient and easily applicable framework to perform SBI of BHMs while checking for model misspecification. It allows one to infer the latent function as an intermediate product, then to perform score compression at no additional simulation cost. This study opens up a new avenue to increase the robustness and reliability of Bayesian data analysis using fully non-linear, simulator-based models.

Funding: This work was done within the Aquila Consortium (https://aquila-consortium.org, accessed on 31 October 2022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code and data underlying this paper, as well as additional plots, have been made publicly available as part of the pySELFI code at https://pyselfi.florent-leclercq.eu (accessed on 31 October 2022).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Leclercq, F.; Enzi, W.; Jasche, J.; Heavens, A. Primordial power spectrum and cosmology from black-box galaxy surveys. *Mon. Not. R. Astron. Soc.* **2019**, *490*, 4237–4253. [CrossRef]
2. Rousset, F. Inferences from Spatial Population Genetics. In *Handbook of Statistical Genetics*; John Wiley & Sons, Ltd.: London, UK, 2007; Chapter 28, pp. 945–979. [CrossRef]
3. Alsing, J.; Wandelt, B. Generalized massive optimal data compression. *Mon. Not. R. Astron. Soc. Lett.* **2018**, *476*, L60–L64. [CrossRef]
4. Papamakarios, G.; Murray, I. Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation. In *Advances in Neural Information Processing Systems 29: Proceedings of the 30th International Conference on Neural Information Processing Systems, 5–10 December 2016, Barcelona, Spain*; Curran Associates Inc.: Red Hook, NY, USA, 2016. pp. 1036–1044.
5. Alsing, J.; Wandelt, B.; Feeney, S. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *Mon. Not. R. Astron. Soc.* **2018**, *477*, 2874–2885. [CrossRef]
6. Gutmann, M.U.; Corander, J. Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *J. Mach. Learn. Res.* **2016**, *17*, 1–47.
7. Leclercq, F. Bayesian optimization for likelihood-free cosmological inference. *Phys. Rev. D* **2018**, *98*, 063511. [CrossRef]
8. Thomas, O.; Pesonen, H.; Sá-Leão, R.; de Lencastre, H.; Kaski, S.; Corander, J. Split-BOLFI for misspecification-robust likelihood free inference in high dimensions. *arXiv*, **2020**, arXiv:2002.09377v1.
9. Beaumont, M.A. Approximate Bayesian Computation. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 379–403. [CrossRef]