

Sample-efficient inference for agent-based cognitive models and other computationally intensive simulators

Alexander Aushev



Sample-efficient inference for agent-based cognitive models and other computationally intensive simulators

Alexander Aushev

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the Computer Science Building on 21 December 2023 at 12:00.

Aalto University
School of Science
Computer Science
Probabilistic Machine Learning

Supervising professor

Samuel Kaski, Aalto University, Finland and The University of Manchester, United Kingdom

Preliminary examiners

Christopher Drovandi, Queensland University of Technology, Australia

Jakob Macke, University of Tübingen, Germany

Opponent

Gilles Louppe, University of Liège, Belgium

Aalto University publication series

DOCTORAL THESES 207/2023

© 2023 Alexander Aushev

ISBN 978-952-64-1555-0 (printed)

ISBN 978-952-64-1556-7 (pdf)

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

<http://urn.fi/URN:ISBN:978-952-64-1556-7>

Unigrafia Oy

Helsinki 2023

Finland



Author

Alexander Aushev

Name of the doctoral thesis

Sample-efficient inference for agent-based cognitive models and other computationally intensive simulators

Publisher School of Science**Unit** Department of Computer Science**Series** Aalto University publication series DOCTORAL THESES 207/2023**Field of research** Probabilistic Machine Learning**Manuscript submitted** 3 November 2023**Date of the defence** 21 December 2023**Permission for public defence granted (date)** 21 September 2023**Language** English **Monograph** **Article thesis** **Essay thesis****Abstract**

In recent years, simulator models have become increasingly popular in many scientific domains, such as epidemiology, cosmology, and behavioural sciences. Since simulators often do not have tractable likelihoods, which are either too costly to evaluate or not available, the field needs to resort to likelihood-free inference (LFI), which uses forward simulations instead. With the development of more complex simulators, traditional LFI methods become unfeasible as the cost of simulations significantly increases. This thesis deals with three challenges that arise in the context of computationally heavy simulators and for which the existing LFI methods, such as approximate Bayesian computation, synthetic likelihood, or neural density estimation approaches, are inadequate since they require a large number of simulations.

The first challenge is modelling complex simulator noise, which influences the accuracy of LFI methods and becomes problematic when simulations are computationally costly. The existing methods either oversimplify the noise (e.g., by assuming it to be Gaussian) or require an infeasible number of simulations to accurately model it. We show how to handle multimodal, non-stationary, and heteroscedastic noise distributions in LFI while also assuming a small simulation budget. For this, we adopt deep Gaussian process surrogates in Bayesian Optimisation (BO), along with novel quantile-based multimodal-capable modifications for the acquisition function and posterior extraction procedures.

Another challenge for modern LFI approaches occurs when they are applied to time-series settings, as these methods either need an accurate model of transition dynamics available or always assume it to be linear. We propose a way of estimating the unknown transition dynamics for state predictions in simulator-based dynamical systems, which greatly reduces the required simulation budget and also enables time-series prediction. Our proposed approach uses a multi-objective surrogate for LFI and a semi-parametric model for the transition dynamics.

Finally, we significantly reduce the time required to select agent-based cognitive models with limited experimental designs. The previous methods have primarily focused on either model selection or parameter estimation, while we achieve both in a fraction of the time. This is accomplished through a novel simulator-based utility objective for selecting designs in BO and a LFI approximation of model marginal likelihood for model selection. This new method is needed for developing and verifying computational cognitive theories, which often lack tractable likelihoods.

Keywords Likelihood-free inference, simulator-based inference, Bayesian optimisation**ISBN (printed)** 978-952-64-1555-0**ISBN (pdf)** 978-952-64-1556-7**ISSN (printed)** 1799-4934**ISSN (pdf)** 1799-4942**Location of publisher** Helsinki**Location of printing** Helsinki **Year** 2023**Pages** 155**urn** <http://urn.fi/URN:ISBN:978-952-64-1556-7>

Preface

But Yossarian knew he was right, because, as he explained to Clevinger, to the best of his knowledge he had never been wrong.

Joseph Heller, *Catch-22*

This thesis is the result of my work in the Probabilistic Machine Learning (PML) group between 2018 and 2023. Throughout this period, I was funded by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI; grants Profi3 292334, and B-REAL, Bridging the Reality Gap in Autonomous Learning, 328400), and was provided computational resources by the Aalto Science-IT Project, for which I am grateful.

I would like to thank my supervisor, Prof. Samuel Kaski, and my numerous co-authors, from whom I have learned a lot during this period. I would also like to express my gratitude to Fang Wang, our PML group coordinator, and my colleagues from the PML group, with whom I had many fruitful discussions.

Finally, I thank my family and friends, who have morally supported me throughout this journey.

Helsinki, November 3, 2023,

Alexander Aushev

Contents

Preface	1
Contents	3
List of Publications	5
Author's Contribution	7
1. Introduction	9
2. Likelihood-free inference	13
2.1 Approximate Bayesian Computation	13
2.2 Synthetic likelihood approaches	15
2.3 Surrogate modelling approaches	16
3. Inference with simulators with complex noise distributions	19
3.1 Bayesian optimization for likelihood-free inference	20
3.2 Gaussian processes	21
3.3 Deep Gaussian processes	22
3.4 Multimodality-capable acquisition function and posterior approximation	24
4. State prediction in simulator-based dynamical systems	27
4.1 Prediction of discrete time series	28
4.2 Multi-output surrogate models	29
4.3 Sample-efficient proposals with state transition dynamics surrogates	30
5. Model selection for simulator-based cognitive models	33
5.1 Computational cognitive models	34
5.2 Bayesian experimental design	36
5.3 Likelihood-free model selection	37

5.4	Bayesian optimization for simulator-based model selection	39
6.	Discussion	43
	References	47
	Errata	61
	Publications	63

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

- I** Alexander Aushev, Henri Pesonen, Markus Heinonen, Jukka Corander, and Samuel Kaski. Likelihood-free inference with deep Gaussian processes. *Computational Statistics and Data Analysis*, Volume 174, 107529, May 2022.
- II** Alexander Aushev, Thong Tran, Henri Pesonen, Andrew Howes, and Samuel Kaski. Likelihood-free inference in state-space models with unknown dynamics. *Statistics and Computing*, 10.1007/s11222-023-10339-8, October 2023.
- III** Alexander Aushev, Aini Putkonen, Gregoire Clarte, Suyog Chandramouli, Luigi Acerbi, Samuel Kaski and Andrew Howes. Online simulator-based experimental design for cognitive model selection. *Computational Brain and Behavior*, 10.1007/s42113-023-00180-7, July 2023.

Author's Contribution

Publication I: “Likelihood-free inference with deep Gaussian processes”

AA obtained the primary conclusions, carried out the numerical experiments, and prepared the article (80%), taking into account suggestions from the other authors. HP, MH and SK helped AA with the design of experiments, refinement of the contributions and together with JC gave suggestions for the paper draft.

Publication II: “Likelihood-free inference in state-space models with unknown dynamics”

AA and SK proposed the idea for the paper. AA deduced key findings, conducted all experiments, prepared the first draft of the paper (90%), and assisted TT in implementing the proposed method. HP and SK helped AA with formulating methodological contributions and providing feedback for the paper drafts. AA, AH and SK designed the case studies.

Publication III: “Online simulator-based experimental design for cognitive model selection”

AH and SK proposed the initial concept of the paper. AA co-designed the main experimental pipeline (40%), implemented all methods, obtained experimental results, implemented demonstrative example and memory retention tasks, co-implemented signal detection and risky choice tasks (about 30%), and co-wrote the initial draft of the paper (all sections, about 70%). AP implemented risky-choice models (70%); wrote sections 4.3, 4.4. and 4.5 of the manuscript. GC co-designed experiments, co-wrote

Author's Contribution

the methods and introduction sections and derived theoretical appendices (50%). SC co-wrote the initial introduction (30%). AH co-designed the main experimental pipeline (40%) and along with SK and LA extensively commented on the paper drafts.

1. Introduction

The focus of this thesis is Bayesian inference for computationally intensive simulator-based models. Simulators have been successfully applied in a wide range of scientific and industrial fields, such as financial markets [Barthelmé and Chopin, 2014; Ong et al., 2018b; Peters et al., 2012; Shafi et al., 2018] and cosmology models [Alsing et al., 2018; Jeffrey et al., 2021; Schafer and Freeman, 2012]. Their primary purpose is to explain real-world data using computational models, which entails estimating properties of natural phenomena in the form of simulator parameters. Knowing the simulator parameters aids in understanding the target phenomena and allows one to reproduce synthetic data similar to data that was observed in the real world.

Challenges and motivation. Despite the advantages of incorporating expert knowledge into a Bayesian inference pipeline through simulators, their usage can pose additional challenges that need to be thoroughly addressed for robust and reliable inference. The fundamental issue with conducting inference for such simulator-based models is that their likelihood is intractable, i.e., difficult or too computationally expensive to evaluate. As a result, classic inference techniques like Markov chain Monte Carlo (MCMC) [Brooks et al., 2011; Gilks et al., 1995; Hastings, 1970], variational inference (VI) [Bishop and Nasrabadi, 2006; Blei et al., 2017] or maximum likelihood estimation (MLE) [Edgeworth, 1908; Pfanzagl, 2011] cannot be applied in such a setting. This problem has been addressed by the large and growing family of likelihood-free inference (LFI) approaches, which use simulations instead of the likelihood in the general Bayesian inference framework.

Bayesian inference and LFI. Bayesian inference [Hacking, 1967; O’Hagan et al., 2004; Robert et al., 2007], and particularly LFI, are employed in the context of simulators because of their capacity to include prior information over simulator parameters, allowing for fewer simulations, and also to quantify the uncertainty associated with estimated parameters. Specifically, consider the simulator parameters as $\theta \in \Theta$, and the observed data

as $\mathbf{x}_{\text{obs}} \in \mathcal{X}$, where Θ and \mathcal{X} are finite-dimensional spaces of simulator parameters and observations, respectively. Then, the Bayes rule can be applied to infer θ :

$$p(\theta \mid \mathbf{x}_{\text{obs}}) \propto p(\mathbf{x}_{\text{obs}} \mid \theta) \cdot p(\theta). \quad (1.1)$$

Here, $p(\theta)$ is the prior provided by a human expert, and $p(\mathbf{x}_{\text{obs}} \mid \theta)$ is the likelihood of the specific observed dataset \mathbf{x}_{obs} being produced by the simulator parameters. The goal of Bayesian inference is to compute the posterior $p(\theta \mid \mathbf{x}_{\text{obs}})$ (notice that typically, the normalising constant is not needed for sampling), which captures information about θ based on the dataset \mathbf{x}_{obs} . Throughout the thesis, I refer to computational models that allow forward simulations but do not have tractable likelihoods as simulator-based models.

Research questions and objectives. The growing complexity of simulators has created a formidable inference problem, one that necessitates inferring simulator parameters with as few simulations as possible. For instance, even simulators that generate synthetic data in a matter of minutes need dozens of hours of computation via standard LFI approaches, which is often infeasible in practise. The need for *sample-efficient* inference techniques (i.e., methods that minimise the number of simulations) arose particularly with the development of complex computational systems, such as traffic simulations [Barceló et al., 2010; Fritzsche and Ag, 1994] and human behaviour modelling [Chen et al., 2021; Georgiou and Demiris, 2017; Gimenez et al., 2007]. Not only do these simulators require sample-efficient inference, but they also frequently display irregularly behaving noise distributions, operate in a time-series setting where their transition dynamics are not accounted for, or need to be first selected from a set of multiple competitor models in a series of experiments. The following research questions aim to address these specific issues as well as the computational challenges associated with high simulator complexity:

1. How can one develop a versatile and sample-efficient LFI approach that: (a) can accurately represent simulators with complex and irregularly behaving noise models, such as multimodal, non-stationary, and heteroskedastic; and (b) maintains a comparable number of simulations required for inference as the current state-of-the-art sample-efficient LFI methods?
2. How can one create a sample-efficient LFI approach for simulators operating in a time-series setting that accounts for their transition dynamics while: (a) enhancing sample-efficiency compared to the state-of-the-art LFI techniques in stationary settings; and (b) enabling accurate predictions of future states?

3. How can one design a simulator-based model-selection method that accurately identifies the correct model and its parameters in a series of controlled experiments while requiring significantly less time compared to alternative methods?

Contributions. This thesis contains a detailed list of references on the methodology of surrogate modelling approaches to LFI in relation to complex noise models, time-series settings, and design optimisation for model selection. The main goal of this thesis is to broaden the applicability of simulator-based inference methods by developing more efficient and theoretically sound approaches to the challenges above. This thesis consists of three publications with the following specific contributions corresponding to the listed research questions above:

- **Publication I** develops a new versatile and sample-efficient LFI approach, enhancing Bayesian optimisation (BO) for LFI to accommodate simulators with complex and irregularly behaving noise models, such as multimodal, non-stationary, and heteroskedastic. This is achieved by employing a deep Gaussian process (GP) surrogate model capable of handling these intricate target distributions. The proposed approach effectively addresses inference failures stemming from the inability of traditional LFI methods to accurately represent complex simulators while requiring orders of magnitude fewer simulations ($O(10) - O(10^2)$) than alternatives.
- **Publication II** presents a probabilistic model for a dynamical system in which the simulator serves as an observation model and an observed time-series is provided. In this temporal setting, we utilise a multi-objective surrogate within a state-space model (SSM) to improve the sample-efficiency of LFI and a transition dynamics model to efficiently propose new parameter points for simulations. Estimating transition dynamics offers the added advantage of enabling accurate state predictions. Remarkably, our method achieves comparable prediction accuracy to existing SSM approaches that require likelihoods, while only necessitating $O(1) - O(10)$ additional simulations per new time-step.
- **Publication III** introduces a new approach for accelerating simulator-based model-selection in controlled experimental settings, addressing the challenges faced by LFI methods in sequential experimental design. This innovative method integrates BO for LFI and Bayesian experimental design (BED), incorporating a new simulator-based utility objective and an LFI approximation of the model marginal likelihood, enabling efficient model selection. We showcase its effectiveness on a variety of different computational cognitive models, where it accurately identifies the correct

model and its parameters, requiring significantly less time (up to two orders of magnitude less compared to alternative LFI approaches).

Thesis structure. The rest of the thesis is organised as follows. Chapter 2 presents an overview of recent LFI methods, focusing on sample-efficient surrogate modelling approaches. Chapter 3 delves into the topic of simulator noise models, providing the context for the deep learning surrogates that are capable of handling such noise distributions. Chapter 4 then discusses LFI methods in time-series settings, along with the methodologies required for adapting the surrogate modelling techniques in such contexts. Chapter 5 discusses LFI experimental design and model selection, as well as its applications in cognitive science research. Finally, Chapter 6 concludes with a summary of the publications and discussions on the benefits and drawbacks of the suggested techniques, as well as future work.

2. Likelihood-free inference

The general goal underlying most LFI algorithms is to find parameters $\theta \in \Theta$ that can reproduce data similar to a given observed dataset x_{obs} . To infer simulator parameters θ that likely created the observed data x_{obs} , LFI approaches generate synthetic observations x from the simulator-based model $x \sim p(x | \theta)$. The quality of inference is determined by how well the simulator imitates the true data-generation process that created the observed dataset x_{obs} , the number of simulations used for inference, and the accuracy of the assumptions made by a specific LFI technique. Please refer to Figure 2.1 for a visual breakdown of the core LFI stages discussed in this chapter.

This chapter aims to provide an understanding of the various LFI methodologies and examine their broad application in the context of computationally expensive simulators. I first introduce approximate Bayesian computation (ABC) in Section 2.1, which is perhaps the most prominent family of LFI techniques and lays the foundation for more sophisticated approaches. Next, I discuss synthetic likelihood methods in Section 2.2, which demonstrate how the problem of LFI can be reformulated as likelihood approximation through simulations. Finally, I delve into LFI surrogate modelling methodologies in Section 2.3, which are further explored later in the thesis.

2.1 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is a well-established family of methods for performing Bayesian inference with simulator models [Beaumont et al., 2002; Csilléry et al., 2010; Sunnåker et al., 2013]. The comparison of observed and synthesised datasets is critical in ABC. The simplest form of ABC, ABC with rejection sampling [Pritchard et al., 1999; Tavaré et al., 1997], involves repeatedly sampling parameters θ from the prior $p(\theta)$ and using them in simulations to build synthetic datasets. The rejection-acceptance mechanism compares the discrepancy (e.g., Euclidean distance,

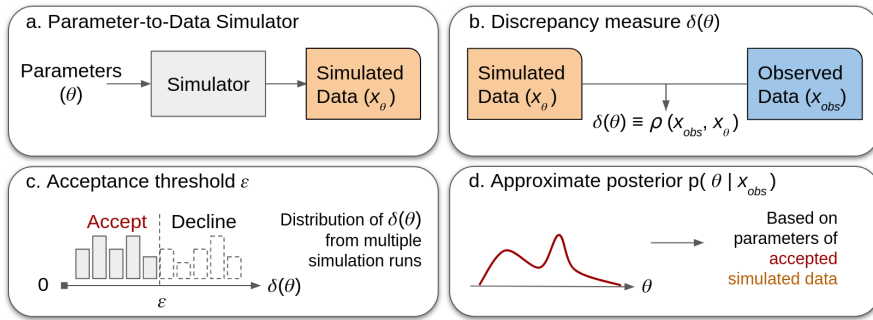


Figure 2.1. Fundamental concepts of LFI. The panels illustrate the steps involved in LFI: (a) parameter-to-data simulation, where chosen parameters θ are used as inputs for the simulator to produce simulated data x_θ ; (b) measuring differences between simulated x_θ and observed data x_{obs} through some distance metric $\rho(x_{\text{obs}}, x_\theta)$; (c) setting acceptance criteria ϵ , where only those simulations with a distance metric below a certain threshold ϵ are accepted; (d) constructing the approximate posterior $p(\theta | x_{\text{obs}})$ based on parameters of accepted simulated data. Note that x_θ notation is used to highlight the parameters θ that gave rise to the simulated data. Summarising statistics have been omitted for clarity.

Kullback-Leibler divergence, or Wasserstein distance) between the observed and synthetic datasets to a user-defined threshold ϵ . Parameters θ with a discrepancy value less than ϵ are accepted as plausible parameters that could produce the observed dataset; otherwise, they are rejected. Lowering the ϵ threshold generally results in more accurate estimations but comes at a significantly higher computational cost, necessitating a tradeoff.

Sampling-based approaches to ABC, such as ABC-MCMC [Marjoram et al., 2003; Sisson and Fan, 2011] and ABC with sequential Monte Carlo (SMC) [Drovandi and Pettitt, 2011; Robert et al., 2008; Sisson et al., 2007], are much more sample-efficient than ABC with rejection sampling. Adaptive algorithms for ABC-SMC [Bonassi and West, 2015; Del Moral et al., 2012] further improved sample-efficiency by replacing simulations from the prior with simulations from a sequentially improved proposal density. However, even these sampling-based approaches require hundreds of thousands or more simulations, even in simple cases, making them unsuitable for computationally costly simulators. For a deeper understanding of ABC and its advancements, one might refer to Lintusaari et al. [2017]. Sisson et al. [2018] provided an extensive review of ABC methods.

Summarising functions in ABC. One principle introduced in ABC that is relevant to this thesis is the use of summarising functions. Calculating discrepancies between high-dimensional datasets can result in significant information loss. To avoid the curse of dimensionality in ABC, the summarising functions $s(\cdot)$ are frequently employed to substitute observed data with low-dimensional summaries (e.g., mean, variance). This way, the

posterior from Equation 1.1 can be approximated with

$$p(\boldsymbol{\theta} \mid \mathbf{x}_{\text{obs}}) \approx p(\boldsymbol{\theta} \mid s(\mathbf{x}_{\text{obs}})) \propto p(s(\mathbf{x}_{\text{obs}}) \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}). \quad (2.1)$$

Of course, the validity of this approximation relies on the choice of summary statistics. If chosen poorly, the resultant posterior can be misleading. Some key works in this area include those by [Bharti et al., 2022; Bi et al., 2022; Fearnhead and Prangle, 2012; Izbicki et al., 2019], as well as explorations into optimal selection strategies [Nunes and Balding, 2010] and adjustments to the ABC distance function [Prangle, 2017].

2.2 Synthetic likelihood approaches

Synthetic likelihood (SL) approaches [Price et al., 2018; Wood, 2010] propose a normal density estimate for the summary statistics. Typically, SL approximates the likelihood function $p(s(\mathbf{x}_{\text{obs}}) \mid \boldsymbol{\theta})$ from Equation (2.1) with a Gaussian distribution:

$$p(s(\mathbf{x}_{\text{obs}}) \mid \boldsymbol{\theta}) \approx \mathcal{N}(s(\mathbf{x}_{\text{obs}}) \mid \mu_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}), \quad (2.2)$$

where $\mu_{\boldsymbol{\theta}}$ and $\Sigma_{\boldsymbol{\theta}}$ are the mean and the covariance of the Gaussian distribution, obtained with MCMC [Wood, 2010] or VI [Ong et al., 2018a]. However, it is important to highlight that the computational performance of this SL approximation might be affected by the number of simulations used for parameter estimation in Equation 2.2. Nevertheless, its influence on the approximation itself remains limited, as examined in Price et al. [2018] and a recent in-depth study on Bayesian SL by Frazier et al. [2022].

Advantages and limitations of SL approaches. The main advantage of the SL approaches over ABC is that they require fewer simulations, and they avoid choosing the tolerance threshold ϵ in Equation (2.1) since they do not need to calculate the discrepancy. Nevertheless, these approaches do introduce an additional approximation of the likelihood by a multivariate Gaussian. This assumption makes the choice of the summary statistics even more important, since if the Gaussian assumption of the likelihood holds exactly, the SL approach can be viewed as using a standard MCMC algorithm with the proposal distribution shaped by factors $p(s(\mathbf{x}_{\text{obs}}) \mid \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$, which has been shown to be more efficient than ABC [Price et al., 2018]. In settings where traditional likelihood methods are challenging, such as epidemic modelling, McKinley et al. [2009] provides insights into making inferences without relying on likelihoods.

Recent works have addressed the limitations of the Gaussian assumption. An et al. [2020] proposed a semi-parametric approach that relaxed the normality assumption of summary statistics, while Frazier and Drovandi [2021] suggested an approach capable of detecting model misspecification

by augmenting the mean or variance in Equation 2.2 with additional free parameters. Moreover, efforts have been made to improve the sample-efficiency of SL methods. For instance, Järvenpää et al. [2021] showed that the computational cost of SL approaches can be significantly reduced, and Priddle et al. [2022] used approximate whitening transformations to break correlation in summary statistics at each algorithm iteration, reducing the number of required model simulations by more than an order of magnitude. Further insights into the challenges posed by high-dimensional LFI (specifically, ABC) are highlighted by Nott et al. [2014]. Despite these advances, the required simulation budget for SL approaches remains large.

2.3 Surrogate modelling approaches

Surrogate modelling methods in LFI [Jabot et al., 2014; Meeds and Welling, 2014; Papamakarios and Murray, 2016] aim to replace time-consuming simulations with predictions from a much faster statistical model. This surrogate can be fitted with a few points from the original model describing a random quantity of interest and then used in the inference procedure. The most prominent examples of LFI surrogates include GPs [Holden et al., 2018; Wilkinson, 2014] and neural density estimators (NDEs; Papamakarios and Murray, 2016; Papamakarios et al., 2019).

Gaussian process surrogates. GPs were initially proposed in the context of ABC, where the GP regression was used as a model for individual summary statistics [Jabot et al., 2014; Meeds and Welling, 2014] or an unknown log-likelihood function [Wilkinson, 2014]. The idea of using GP surrogates was further developed by Gutmann and Corander [2016], who used the GP predictive mean and variance in Bayesian optimisation (BO) to significantly reduce the number of simulations for inference by suggesting locations for simulations. In the following chapter, I examine BO for LFI and GP surrogates in greater detail.

Neural density estimator surrogates. NDEs [Fan et al., 2013; Paige and Wood, 2016; Papamakarios and Murray, 2016] are a broad family of LFI techniques that employ neural networks [Rosenblatt, 1958; Yegnanarayana, 2009] as surrogates for random quantities of interest. Some methods focus on using them as models for the likelihood [Fan et al., 2013] or likelihood ratio [Hermans et al., 2020], which is useful when the prior needs to be adjusted (e.g., for prior sensitivity analysis), while others provide solely the posterior [Papamakarios and Murray, 2016]. Unlike GP-based methods, NDE surrogate variants are more diverse, ranging from masked autoregressive flows [Papamakarios et al., 2017] to invertible neural networks [Radev et al., 2020].

In the context of sample-efficiency, one appealing feature of NDEs is their ability to use intermediate results of fitting the neural network to offer candidates for subsequent simulations [Durkan et al., 2018; Lueckmann et al., 2019; Papamakarios et al., 2019]. While NDEs often require an approximate number, roughly on the order of a thousand simulations, for training, it is important to note that this estimation can be influenced by the dimensionality of the problem at hand.

Amortisation in NDE research. A considerable amount of NDE research focuses on amortisation [Paige and Wood, 2016; Radev et al., 2020], which refers to the model’s ability to condition its outputs using the observed dataset as a random variable. This way, all pre-training can be done before any data is collected, and it is particularly useful when the inference task requires posteriors for multiple independent datasets. Unlike ‘local’ approaches, such as ABC, SL, and GP-based surrogates, amortised approaches are trained on a range of data and do not need to be retrained. This property allowed us to use them as a comparison method in a time-series setting in Publication II. Throughout the rest of the thesis, I repeatedly compare our developed methods with NDEs.

3. Inference with simulators with complex noise distributions

Simulators generate synthetic data x from the data-generation distribution $x \sim p(x | \theta)$, which depends on the parameters θ . In Publication I, we tackle multimodal, non-stationary, and heteroskedastic noise distributions $p(x | \theta)$. These distributions occur in simulators when outputs for identical parameter settings exhibit multiple modes, and when the intrinsic variability and data-generation properties of the simulator change significantly for different parameter values. Existing LFI methods struggle to effectively handle these distributions in a sample-efficient manner because they typically assume the Gaussianity of the noise (or output) distribution. This assumption is prevalent in state-of-the-art sample-efficient LFI approaches, such as GP-ABC [Wilkinson, 2014] and BO for LFI [Gutmann and Corander, 2016]. To address these complex noise distributions, we propose using highly adaptable surrogates in BO and accounting for any abnormalities they may encounter when fitting the objective function. An illustration of the influence the surrogate may have on modelling discrepancy noise distributions is demonstrated in Figure 3.1.

This chapter discusses how simulator noise distributions affect inference. Section 3.1 introduces the general framework of BO for LFI, which makes explicit assumptions about the noise model and is used throughout this thesis. Section 3.2 provides background on GP surrogates and discusses the process of inference failure using BO for LFI as an example. In Section 3.3, I introduce deep GPs, explaining their increased versatility compared to standard GPs and how they can help represent arbitrary noise distributions. Finally, Section 3.4 details the adjustments required in BO for LFI to accommodate deep GP surrogates and multimodal noise distributions, the most challenging subvariant of noise distributions. Publication I contains further information on implementation and inference procedures.

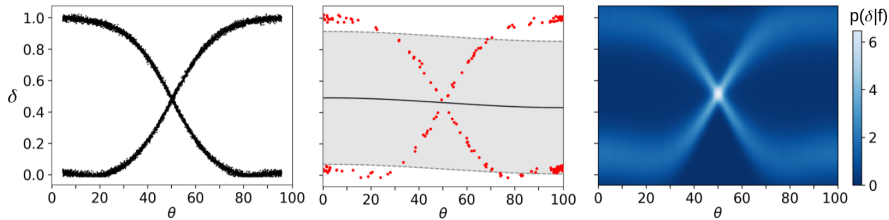


Figure 3.1. Demonstration of GPs modelling bimodal discrepancies. (a) For every parameter value θ , the true discrepancy objective $\delta(\theta)$ displays a clear bimodal structure. (b) When a traditional GP is used as a surrogate for the discrepancy function, it struggles to capture the modes. Here, red points mark observed data, the black line represents the GP mean, and the shaded region conveys the level of uncertainty. (c) By breaking the correlation of the outputs for the same inputs, certain deep GP variants (see Section 3.3) are able to accurately capture and represent the multimodal discrepancy. Here, the shades of blue represent the density of the deep GP outputs. Note: this figure is adapted from Publication I.

3.1 Bayesian optimization for likelihood-free inference

BO for LFI [Gutmann and Corander, 2016] is one of the GP-based surrogate approaches. Similarly to ABC, BO for LFI actively uses the discrepancy $\delta(\theta) \equiv \rho(\mathbf{x}_{\text{obs}}, \mathbf{x}_\theta)$ as the objective function. The predictive distribution for a new discrepancy value obtained from the GP model is used to approximate the posterior through:

$$p(\theta \mid \mathbf{x}_{\text{obs}}) \approx q(\theta \mid \mathbf{x}_{\text{obs}}) \propto p(\theta) \cdot \mathcal{L}(\mathbf{x}_{\text{obs}} \mid \theta), \quad (3.1)$$

where the likelihood function can be approximated through $\mathcal{L}(\mathbf{x}_{\text{obs}} \mid \theta) \approx \mathbb{E}[\kappa_\epsilon(\delta(\theta))]$ with $\kappa_\epsilon(\cdot)$ being the kernel with the maximum at zero, and whose bandwidth ϵ acts as an acceptance-rejection threshold.

Acquisition functions and posterior approximation. In BO for LFI, the Gaussian cdf $\mathcal{N}_{\text{CDF}}\left(\frac{\epsilon - \mu(\theta)}{\sqrt{\nu(\theta) + \sigma^2}}\right)$ with mean 0 and variance 1 can be used as $\mathbb{E}[\kappa(\cdot)]$. Here, ϵ is the tolerance threshold from Section 2.1, σ is the standard deviation of the discrepancy objective, and $\mu(\cdot)$ and $\nu(\cdot)$ are the GP mean and variance functions, respectively. Applying this kernel to obtain the posterior $q(\theta \mid \mathbf{x}_{\text{obs}})$ was shown to be a reasonable approximation of $p(\theta \mid \mathbf{x}_{\text{obs}})$ [Gutmann and Corander, 2016].

The role of BO in LFI is to drastically reduce the number of simulations for the GP surrogate by proposing simulation locations at which to evaluate the discrepancy objective. Since these objective evaluations are noisy due to the stochasticity of the simulator, some locations need to be evaluated multiple times to get a good approximation of the data-generation noise. At each iteration of the BO algorithm, the next evaluation location θ_t is chosen by optimising an acquisition function $u(\cdot)$, such as probability of improvement (PI) [Kushner, 1964], expected improvement (EI) [Jones et al., 1998; Moćkus, 1975] or lower confidence bound criterion (LCBC)

[Srinivas et al., 2009]:

$$\theta'_t = \arg \max_{\theta_t} u(\theta_t). \quad (3.2)$$

The role of the acquisition function $u(\cdot)$ is to measure the potential improvement in the knowledge of θ'_t . Depending on the utility objective, the optimisation procedure can be formulated as a maximisation or minimization problem.

Optimal acquisition rules and recent developments. Lately, there has been little research into the optimal acquisition rules for LFI. Most notably, Järvenpää et al. [2019] proposed BO strategies specifically for estimating the posterior distribution. Nonetheless, BO is more often explored outside of LFI [Daulton et al., 2022; Hvarfner et al., 2022; Verma et al., 2022], while it can still be applied in specific LFI cases, such as when the likelihood function is tractable but costly to evaluate. Yet, these recent trends in BO are beyond the scope of this thesis.

3.2 Gaussian processes

Gaussian processes (GPs) [Matheron, 1963; Williams and Rasmussen, 2006] are Bayesian non-parametric models commonly used in BO, which can be interpreted as taking priors on functions [Barber, 2012]. Assuming a GP prior results in a variety of useful properties for the BO objectives, such as output smoothness and the capability to model the inherent variability in the data. These properties significantly reduce the number of simulations required for LFI in GP-based surrogate modelling methodologies described in the preceding chapter.

GPs in BO for LFI describe the prior mean $\mu(\cdot)$ (e.g. linear, constant) and covariance function $\kappa(\cdot, \cdot)$ (e.g., squared exponential, Matern) [Seeger, 2004] of the discrepancy objective:

$$\mathbb{E}[\delta] = \mu(\theta), \quad (3.3)$$

$$\text{cov}[\delta, \delta'] = \kappa(\theta, \theta'), \quad (3.4)$$

where δ is used as a shorthand for $\delta(\theta)$. By changing the mean and covariance functions, one may affect the properties of a GP surrogate and, as a result, how accurately the surrogate models the true objective function.

Gaussian noise model and simulator noise. Typically, GP surrogates assume the Gaussian noise model, which translates to the noise model of a simulator in LFI,

$$p(\delta \mid \mathbf{f}) = \prod_{n=1}^N \mathcal{N}(\delta_n \mid f(\theta_n), \sigma_n^2), \quad (3.5)$$

where σ_n is the standard deviation of the noise and $f(\cdot)$ are the outputs of the latent GP function. When there is a small simulation budget, the Gaussian noise model prevents a GP from overfitting while also quantifying the uncertainty of GP predictions. The GP noise models are an essential feature of the GPs that I will look at in detail in the next section.

Advanced GP variants and GP limitations. Since GPs were first introduced, a lot of work has been done on developing more advanced GP variants to address the low data scalability and flexibility of vanilla GP inference. The shortcoming in data scalability is caused by the high computational cost of fully analytical or *exact* GP inference. The computational cost of GP inference is relevant to sample-efficient LFI because when the cost of GP inference exceeds the cost of simulations, it is more advantageous to spend the computational budget on performing more simulations rather than on spending time on a costly inference mechanism. Sparse GPs [Leibfried et al., 2020; Titsias, 2009] were suggested to alleviate this problem through approximation of the true GP posterior with a set of pseudo-training examples or *inducing points*. A recent alternative to sparse GP approximation is accessing the kernel matrix only through matrix multiplication [Wang et al., 2019], which was shown to be effective in training a GP with over a million data points.

Another significant shortcoming of vanilla GPs is their lack of flexibility. Although GPs are generally quite flexible, there are certain types of functions, such as non-stationary or heteroskedastic, that are particularly difficult for them. Several approaches have been developed to address these issues, including using non-stationary kernels [Gibbs, 1998; Heinonen et al., 2016; Paciorek and Schervish, 2003], inventing deep kernel learning techniques [Ober et al., 2021; Wilson et al., 2016], and, most notably, adopting deep GPs [Damianou and Lawrence, 2013; Havasi et al., 2018; Salimbeni and Deisenroth, 2017]. These advancements aim to improve the flexibility of GPs, allowing them to better model complex functions and enhance their applicability in various problem domains, including LFI.

3.3 Deep Gaussian processes

BO for LFI traditionally assumes Gaussian noise in the objective function. Consequently, if this assumption is violated, the LFI approximation becomes poor. A natural solution to this problem is to analyse the simulator noise and then employ the stochastic process that assumes the appropriate noise model (e.g., beta, student-t). However, this solution necessitates extensive knowledge of the simulator data-generation mechanism, which may be particularly challenging for computationally intensive simulators. Careful analysis would necessitate many more simulations than what surrogate modelling techniques for LFI typically require. Instead, in Pub-

lication I, we advocate using a considerably more flexible prior, a deep GP, as a solution.

Modeling complex noise distributions. Deep GPs overcome the limitations of GPs by composing multiple GPs together, yielding more flexible and powerful function representations [Damianou and Lawrence, 2013; Dunlop et al., 2018]. For example, a deep GP containing two GPs with latent functions $f(\cdot)$ and $g(\cdot)$ can be defined as:

$$p(\boldsymbol{\delta}|\mathbf{f}, \mathbf{g}) = \mathcal{N}(\boldsymbol{\delta}|f(g(\boldsymbol{\theta})), \sigma^2). \quad (3.6)$$

Here, f and g are the latent outputs for $f(\cdot)$ and $g(\cdot)$ respectively. Despite being heavily inspired by deep neural networks [Damianou and Lawrence, 2013; Pleiss and Cunningham, 2021], deep GPs require much less data to train due to the Bayesian treatment of their components, which allows them to cope with complexity. However, deep GPs are intractable, necessitating the use of approximate inference techniques for training, such as variational [Salimbeni and Deisenroth, 2017] or Monte Carlo [Havasi et al., 2018]. The main motivation for using deep GPs in Publication I was to have a flexible model capable of representing non-Gaussian and multimodal noise distributions with minimal data. However, not every deep GP variant exhibits these properties, and multimodality, in particular, is extremely challenging to model using traditional inference approaches.

Addressing multimodal noise distributions. In order to approximate multimodal target noise distributions, deep GP inputs should yield uncorrelated latent outputs [Salimbeni et al., 2019], which correspond to multiple modes. One way to achieve this is by maintaining the distribution over deep GP parameters through Monte Carlo simulations and later sampling these parameters for deep GP predictions [Havasi et al., 2018]. If the Monte Carlo simulations are successful, the parameter distribution retains information about the modes, and two separate samples related to specific modes result in uncorrelated latent outputs. The competitor importance-weighted variational inference approach by Salimbeni et al. [2019], which we use in Publication I, introduces a latent variable w that augments the deep GP input vector $\boldsymbol{\theta}$ with Gaussian noise $\mathcal{N}(0, 1)$. This method allows the inference mechanism to treat an extra dimension in the input vector as information about the mode, breaking the correlation of deep GP outputs for various values of w . The deep GP representation for the two-GP layer architecture mentioned earlier would be:

$$p(\boldsymbol{\delta}|\mathbf{f}, \mathbf{g}, w) = \mathcal{N}(\boldsymbol{\delta}|f(g([\boldsymbol{\theta}, w])), \sigma^2), \quad (3.7)$$

$$p(\boldsymbol{\delta}|\mathbf{f}, \mathbf{g}) = \mathbb{E}_{p(w)}\mathcal{N}(\boldsymbol{\delta}|f(g([\boldsymbol{\theta}, w])), \sigma^2), \quad (3.8)$$

where the latent variable (LV) augmentation is applied to the inputs of the first stacked GP with the latent function $g(\cdot)$, and $p(w)$ is the distribution

of the latent variable. The resulting LV-GP-GP architecture provides more flexible DGP posterior approximations [Salimbeni et al., 2019], which we apply in LFI.

3.4 Multimodality-capable acquisition function and posterior approximation

A multimodal surrogate model alone is insufficient for effectively handling multimodal simulator noise distributions. The LFI method should also be able to exploit multimodality to propose locations for subsequent simulations and accurately extract the estimated posteriors. Both of these issues have previously remained unresolved for multimodal distributions in BO for LFI, as both the acquisition and posterior extraction procedures were designed exclusively for Gaussian noise models. In Publication I, we suggest using a quantile-based threshold to focus on the important lower-valued regions of the discrepancy surface, which is a key aspect of the proposed solution. The resulting method addresses multimodality while maintaining all the useful properties of BO for LFI, such as its extreme sample-efficiency.

Quantile-based threshold. The quantile-threshold ϵ_q conditions deep GP predictive samples, estimating the lowest values of the discrepancies more accurately. Specifically, it modifies the mean $\mu_q(\boldsymbol{\theta})$ and variance $\nu_q(\boldsymbol{\theta})$ of a deep GP through a quantile function $Q(\cdot)$:

$$\mu_q(\boldsymbol{\theta}) = \mathbb{E}\{\boldsymbol{\delta}^n : \boldsymbol{\delta}^n \leq Q(\epsilon_q)\}_{n=1}^N, \quad (3.9)$$

$$\nu_q(\boldsymbol{\theta}) = \text{var} \{\boldsymbol{\delta}^n : \boldsymbol{\delta}^n \leq Q(\epsilon_q)\}_{n=1}^N, \quad (3.10)$$

where N is the number of predictive samples. This procedure shifts the focus of the noise representation to the important lower-valued regions of the discrepancy surface, filtering out the predictive samples that are lower than the value $Q(\epsilon_q)$ associated with the quantile-threshold ϵ_q .

In other words, when faced with multiple noise distribution modes, the decision on which parameters to simulate next and which to classify as the ones most likely to yield the observed dataset is based only on the mode with the lowest discrepancy values. The role of the quantile-threshold ϵ_q is to regulate these discrepancy values by setting the acceptable signal-to-noise ratio in the simulator distribution, but also to help maintain the sample-efficiency of the LFI approach. For instance, a quantile threshold that is too low would require more predictive samples (and hence more computations), with the risk of catching merely noise in the resulting quantile. Therefore, values between 0.1 and 0.4 are recommended.

Application in acquisition functions and posterior approximation. The modified mean $\mu_q(\boldsymbol{\theta})$ and variance $\nu_q(\boldsymbol{\theta})$ from Equations (3.9) and (3.10) can be

directly used in acquisition functions, such as LCBSC or EI, and the LFI approximation of the posterior in Equation (3.1). The proposed modification maintains the sample-efficiency of BO for LFI with a small computational overhead, which, in the context of computationally intensive simulators, is negligible. It also enables BO to handle multimodal or skewed uncertainties. This was demonstrated on a range of simulators, most notably in an inverse reinforcement learning grid-world planning problem, which was used as an example of multimodal noise in the simulator, where the same behavioural parameters of an agent could lead to it taking different routes on a map. In conclusion, the proposed approach effectively addresses the research question by enabling a sample-efficient LFI approach that can handle complex noise distributions.

Conclusion. This chapter presented a versatile and sample-efficient LFI approach that successfully handles simulators with complex and irregularly behaving noise models, such as multimodal, non-stationary, and heteroskedastic. By introducing a quantile-based threshold, we focus on the important lower-valued regions of the discrepancy surface, effectively addressing multimodality while maintaining the extreme sample-efficiency of BO for LFI. Our method modifies the mean and variance of a deep GP using a quantile function, which can be directly used in acquisition functions and the LFI approximation of the posterior. This modification retains the sample-efficiency of BO for LFI with a small computational overhead, enabling it to handle multimodal or skewed uncertainties.

Having a method capable of handling simulators with complex noise distributions allowed us to consider more challenging settings. We kept looking into how to infer characteristics of actual people in a cognitive task (where the human model serves as a simulator) in response to the inverse reinforcement learning application in Publication I. Specifically, we focused on humans' capacity for situational adaptation and how this can be accounted for in LFI. The use of LFI approaches in a time-series scenario, where the simulator model must continuously adapt to evolving observations, is covered in the next chapter.

4. State prediction in simulator-based dynamical systems

State-space models (SSMs) [Baum and Eagon, 1967; Baum and Petrie, 1966; Koller and Friedman, 2009] provide a principled framework for analysing the simulator-based dynamical systems considered in this chapter. These models involve a sequence of latent variables, *states*, which are governed by an observation model, $g(\boldsymbol{\theta}_t) \sim p(x_t | \boldsymbol{\theta}_t)$, and transition dynamics, $h(\boldsymbol{\theta}_t) \sim p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t)$. Both the observation model and transition dynamics define state evolution over time, with the data x_t being collected at each time-step t :

$$\boldsymbol{\theta}_{t+1} = h(\boldsymbol{\theta}_t) + v_t, \quad (4.1)$$

$$x_{t+1} = g(\boldsymbol{\theta}_{t+1}) + e_t, \quad (4.2)$$

where v_t and e_t are white noise vectors. In this setting, the simulator represents the observation model, and the latent states correspond to simulator parameters. In this chapter, I explore the creation of a sample-efficient LFI approach for simulators operating in a time-series setting, accounting for their transition dynamics. This proposed approach aims to achieve a similar or better performance than current state-of-the-art LFI techniques while requiring fewer simulations and also enabling accurate state predictions for future observations. The setting where the observational model is a simulator and the transition dynamics do not allow simulations and need to be learned non-parameterically can be encountered in brain activity mapping [Bassett and Sporns, 2017; Shenoy et al., 2013] and epidemiological modelling [Grenfell et al., 2001; Tamerius et al., 2011]. Figure 4.1 underscores the key distinctions between conventional SSMs and the specific setting explored in this chapter.

The chapter is dedicated to sample-efficient LFI and state prediction in dynamical systems, covering the fundamentals and alternative approaches to time-series prediction in Section 4.1. In Section 4.2, multi-output surrogates are introduced as a better alternative to GPs in LFI for SSMs. Section 4.3 concludes the chapter by demonstrating how the transition dynamics surrogate can be used to generate informative simulation inputs in LFI. The proposed approach in this chapter is detailed in Publication II.

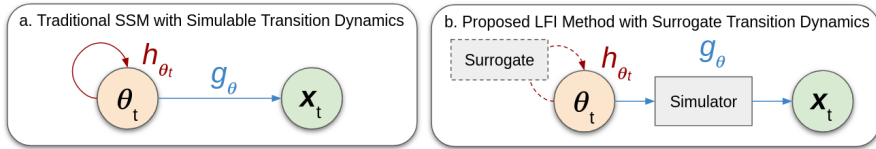


Figure 4.1. Comparison of traditional SSMs and the proposed LFI approach. Panel (a) showcases the standard SSM with its simulable transition dynamics. In contrast, Panel (b) introduces our LFI method, employing simulators for observations while leveraging a surrogate to emulate the elusive transition dynamics. While many Bayesian SSMs typically use a specific form for both transition and observation densities, aiming to deduce the posterior distribution of hyperparameters, this chapter delves into a more unconventional issue: a non-parametric estimation of the states and their transition density.

4.1 Prediction of discrete time series

Time-series prediction methods [Weigend, 2018] encompass a wide range, from simple linear autoregressive [Sims et al., 1990; Staudenmayer and Buonaccorsi, 2005; Udney Yule, 1927] and integrated moving average predictors [Adebiyi et al., 2014; Liu et al., 2016] to more complex nonlinear neural networks [Connor et al., 1994; Frank et al., 2001; Lapedes and Farber, 1987] and support vector machine-based approaches [Sapankevych and Sankar, 2009; Thissen et al., 2003]. The objective of time-series prediction (or forecasting) is to find a continuation for a given sequence (x_1, x_2, \dots, x_n) formed by some stochastic or deterministic dynamic system process. Traditional prediction approaches involve developing an underlying model that generates the observed sequence, with the selection of the appropriate method hinged on the properties of the dynamic system.

Learning and leveraging observational models. In certain scenarios, such as speech [Juang and Rabiner, 1991; Nilsson and Ejnarsson, 2002; Rabiner, 1989] or security research [Liu et al., 2019; Tang and Dong, 2019], modelling the underlying dynamics that generate the time-series through an observational model, as SSMs, proved beneficial. Probabilistic time-series prediction was traditionally accomplished by filtering-based methods [Joo and Kim, 2015; Wu and Wang, 2012]; however, in recent years, powerful deep learning methods have emerged with an encoder-decoder prediction paradigm, such as recurrent neural networks [Hochreiter and Schmidhuber, 1997; Li et al., 2017; Yu et al., 2017] and transformers [Vaswani et al., 2017; Zhou et al., 2021]. Most of these methods learn latent representations of the time-series and capture dynamics in the embedding space.

When states require interpretability, the learned representations cannot be used in their place, at least in the traditional SSM setting that was described earlier in this chapter. Instead, the burden of interpretation rests on the observation model (or the likelihood), which is why a broad range of methods focuses on the transition dynamics. These methods include

Kalman filters [Yang et al., 2020; Zerdali and Barut, 2017], GP SSMs with variational inference [Curi et al., 2020; Doerr et al., 2018; Ialongo et al., 2019] and MCMC [Chopin et al., 2013; Flury and Shephard, 2011; Kattwinkel and Reichert, 2017].

Limitation of LFI methods in SSMs. On the other hand, LFI methods [Calvet and Czellar, 2015; Hasegawa et al., 2016; Jasra et al., 2012] are capable of inferring the states of an SSM without a tractable likelihood; however, they suffer from poor sample-efficiency, as most of the filtering approaches to some extent do. They also make assumptions about the state dynamics that are too simplistic, preventing them from making time-series predictions in the cases we considered in Publication II.

In summary, various time-series prediction methods were explored in this section, with some instances favouring the use of observational models. While certain methods focused on interpretable states and transition dynamics, LFI methods encountered limitations in sample-efficiency and overly simplistic assumptions about state dynamics.

4.2 Multi-output surrogate models

In Publication II, we propose a multi-objective LFI approach that combines information between multiple states in a single surrogate while simultaneously learning the unknown transition dynamics. Our hypothesis was that sharing the parameters of the surrogate between multiple states would enable robust inference with as few simulations as possible. As more simulations are collected and the observed data is revealed in an online fashion, we gradually improve posterior approximations of individual states while training the dynamics model. In this section, I describe the LFI aspect of the problem, while in the next section, I cover state predictions.

Sharing state information. To share information between states, we employ a multi-output model, in which each state represents a separate objective. This model should be flexible enough to fit a discrepancy function, following the surrogate modelling approaches I mentioned in earlier chapters. It uses the same I simulator parameters $\theta^{(1:I)}$ as inputs and the same corresponding synthetic observations $x^{(1:I)}$ to calculate the discrepancy objectives for states, assuming that the observational model remains the same (notice, the discrepancy is deterministic and states differ only in observed datasets). This way, we share synthetic observations across all states without requiring any unnecessary simulations. In Publication II, we use a linear model of coregionalization (LMC) [Fanshawe and Diggle, 2012; Liu et al., 2018] as an example of such a multi-output surrogate, where T outputs $f_i(\cdot)$ are expressed as a linear combination of Q latent

GPs \mathbf{u}_q :

$$f_t(\boldsymbol{\theta}^{(1:T)}) = \sum_{q=1}^Q a_{t,q} \mathbf{u}_q, \quad (4.3)$$

$$\mathbf{u}_q \sim \text{GP}(0, \kappa(\boldsymbol{\theta}, \boldsymbol{\theta}')). \quad (4.4)$$

where $a_{t,q}$ are linear coefficients, and q is the number of GPs shared across states by the surrogate. Once we fit the surrogate, the LFI approximations can be extracted similarly as in the previous chapter with Equation (3.1).

Moving window simplification. Considering all states at the same time would be very computationally costly, especially when the number of observed datasets in a time-series T is large. Therefore, we propose fitting the multi-output surrogate only for the most recent observations in a so-called *moving window*:

$$p(\boldsymbol{\theta}_{t_0-1:T} \mid \mathbf{x}_{t_0:T}) \propto \prod_{t=t_0}^T p(\mathbf{x}_t \mid \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}), \quad (4.5)$$

which is equivalent to filtering approaches used when the moving window's size $t_0 = 0$. This simplification still makes use of the shared information between states, focusing on the general properties of the objective functions without overfitting. The process of learning the transition dynamics and generating simulation proposals is discussed in the next section.

4.3 Sample-efficient proposals with state transition dynamics surrogates

When states can be interpreted as simulator parameters, predicting them is sufficient for time-series forecasting. The predicted states can later be used in an observational model (a simulator) to generate observations. Therefore, in Publication II, we propose a method for learning a state transition dynamics model that can predict the next states and use them as simulation locations for sample-efficient inference in a time-series setting. Such a transition dynamics model \tilde{h}_θ requires the current observation \mathbf{x}_t to produce a predictive posterior of the next state (note that the transition model assumes the Markov property and uses pairs of states for training instead of their whole trajectories):

$$p(\boldsymbol{\theta}_{t+1} \mid \mathbf{x}_t) = \int \tilde{h}_\theta(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t \mid \mathbf{x}_t) d\boldsymbol{\theta}_t. \quad (4.6)$$

We use samples from the LFI posteriors extracted from the multi-output surrogate with Equation (3.1) to train the transition dynamics model. This way, the more accurate our posterior approximations are, the higher the quality of training samples we can provide for the dynamics model.

Approaches to modelling transition dynamics. In Publication II, we explore two approaches to modelling unknown transition dynamics. The first involves approximating the transition dynamics in Equation (4.6) locally with a simple linear model. This results in a series of models that can linearize the transition dynamics of any form, but its downside is that all these models must be stored to generate new state trajectories. The second approach is more general, adapting a highly flexible semi-parametric model that can be constantly improved with new data instead of being retrained from scratch. Ultimately, we favour the latter approach, using a Bayesian neural network [Esposito, 2020; Kononenko, 1989] as a surrogate for the dynamics.

Application in online and interactive settings. Utilising the transition dynamics surrogate for proposing simulation locations through Equation (4.6), the proposed method enables LFI for dynamical systems with minimal simulations required for accurate parameter estimates. This is particularly important in situations where observations are collected gradually and immediate intervention may be necessary, such as medical treatment in the ICU. Importantly, the proposed method can be used in an *online* or *interactive* setting, making use of many already existing simulators in a dynamical setting. It is noteworthy that while certain simulators can generate the whole time-series without a transition dynamics model, we opted for a separate observation model for each state to accommodate gradual data collection and increase the applicability of the method to a broader range of simulators.

Conclusion. This chapter introduced a sample-efficient LFI approach for time-series simulators that accounts for transition dynamics, enhances sample-efficiency, and enables accurate future state predictions. The multi-output surrogate model facilitates information sharing between states, while the highly flexible semi-parametric model, such as a Bayesian neural network, serves as a surrogate for state transition dynamics. Combining these elements into a single approach allows for efficient parameter estimates with minimal simulations, making it suitable for online and interactive settings.

In online settings, where sample-efficient LFI is crucial, the state prediction model does not influence actual future observations. However, people or other rational agents frequently use these state predictions to make decisions that alter the dynamical system [Dayan and Daw, 2008; Loftin and Oliehoek, 2022]. In the following study, I investigate the cognitive diagnostics of humans as an example of such a dynamical system in collaboration with cognitive scientists. The next chapter delves into this application, focusing on static behaviour since applying experimental design for LFI model selection has proven notoriously difficult.

5. Model selection for simulator-based cognitive models

An important problem in computational cognitive science is to determine which computational model $m \in \mathcal{M}$ in a given set of competitor models \mathcal{M} offers the best explanation of human behaviour in a certain task. In the previous chapter, I considered human behaviour as a dynamic system whose observations x_t were produced as a function of human model parameters. In contrast, here the human participant produces observations in a controlled experimental environment, i.e., according to some experimental designs d_t : $x_t \sim p(\cdot \mid m, \theta, d_t)$, where θ remains constant. The controlled experimental environment allows for multiple experimental trials with the purpose of revealing as much information about human behaviour as possible. The procedure of selecting experimental designs so that they maximise the utility of running experiments with a particular design is called *experimental design optimisation*.

In the experimental design optimisation for model selection, we are given several models $m \in \mathcal{M}$, and we need to choose a model m' and its parameters θ' that are the most consistent with human behaviour across all experimental trials. The goal is to accurately identify the correct model and its parameters in a series of controlled experiments while requiring significantly less time compared to alternative methods. Table 5.1 provides a comprehensive comparison of experimental design methods, detailing their capabilities and applicability in various contexts.

The mechanisms for sample-efficient likelihood-free model selection are examined in this chapter. Section 5.1 first presents computational cognitive models and their usage in cognitive science as the main motivation for the methods that are later discussed in this chapter. Second, Section 5.2 describes the background of Bayesian experimental design, which is a common strategy for design optimisation. Third, Section 5.3 discusses alternate experimental design procedures for model selection, emphasising the difficulties encountered by modern methods when dealing with models that do not have tractable likelihoods. Finally, Section 5.4 describes a newly developed sample-efficient experimental design strategy for simulator-based model selection, which greatly improves the computa-

Reference	Method	Attributes						
		PE	MS	SBM	Amor.	LTP	Adapt.	LFI
Cavagnaro et al. [2010]	ADO	✓	✓	×	×	×	✓	×
Kleingesse and Gutmann [2020]	MINEBED	✓	×	✓	✓	×	×	✓
Blau et al. [2022]	RL-BOED	✓	×	✓	✓	✓	✓	×
Moon et al. [2022]	BOLFI	✓	×	×	×	-	-	✓
Pudlo et al. [2016]	RF-ABC	×	✓	×	✓	-	-	✓
Proposed method	BOSMOS	✓	✓	✓	×	×	✓	✓

Table 5.1. Summary of methods for experimental design, distinguishing based on several attributes. *Parameter Estimation* (PE): Estimating the values of parameters. *Model Selection* (MS): Choosing among various models. Some methods, such as MINEBED, recast MS as a PE problem, considering only the model index as a parameter and fixing the rest. This table specifically emphasizes MS for simulator models that also allow parameter changes. *Simulator-Based Models* (SBM): Methods proposing designs without needing tractable model likelihoods. *Amortisation* (Amor.): Enables quick design proposals or inference but necessitates retraining when beliefs update. *Long-Term Planning* (LTP): Essential for efficient design space exploration, especially when the budget is predetermined, contrasting with short-sighted methods. *Adaptive* (Adapt.): Allows for sequential design optimisation. *Likelihood-Free Inference* (LFI): Inference mechanism based on new experimental data. The '×' indicates design selection is not applicable to that method. Adapted from Publication III.

tional cost needed for simulator cognitive models. Publication III contains full implementation details for the suggested technique.

5.1 Computational cognitive models

Cognitive science is a synthesis of numerous disciplines, including psychology, linguistics, anthropology, computer science, neuroscience, and philosophy [Bechtel, 2013; Collins, 1977; Hunt, 1989]. It seeks to create theoretical frameworks for describing human and animal behaviour, for instance, when doing a certain cognitive task.

Early models and recent advances. The earliest cognitive models can be traced back to the early 1970s, when mathematical linguists and artificial intelligence scientists identified semantic similarities in the domains of colour vocabulary [Berlin B, 1969; Rosch, 1973] and spatial description of three-dimensional sceneries [Falk, 1972; Marr and Nishihara, 1976]. Recent advances in cognitive modelling saw improvements in both analytical models [Anderson and Schooler, 1991; Hahn and Oaksford, 2006, 2007; O’Doherty et al., 2003], such as Bayesian models with tractable (or closed-form) solutions, and computational models [Duggins, 2014; Madsen et al., 2018; Orr and Plaut, 2014; Orr et al., 2013; Sun et al., 2006], which use contemporary computing techniques ranging from multi-agent simulations to deep learning. Since inference for the former class of computational cog-

nitive models had been severely constrained, it became one of the focuses of this thesis.

Agent-based paradigm. Computational cognitive models often operate in an agent-based paradigm, where the model is represented by a reinforcement learning [Kaelbling et al., 1996; Sutton and Barto, 2018] agent. In reinforcement learning, the problem is typically formulated as partly observable Markov decision processes (POMDPs) [Astrom, 1965; Littman, 2009; Spaan, 2012], which cognitive scientists use as a framework for studying normative behaviour in experimental settings. POMDPs were proven to be a valid model for dopamine regulation in animals [Day et al., 2007; Schultz, 2002] and optimal exploitative risky choice decisions in humans Daw et al. [2006], where POMDPs were used to estimate belief states. In direct psychological and neurological applications, it is also common to interpret these models through the lenses of Bayesian decision theory [Dayan and Daw, 2008; Kim et al., 2019; Lee, 2018; McCarthy et al., 2021], since the activity of populations of neurons can be represented and maintained through the Bayes rule. These theories of human behaviour led to the development of multiple models whose data-generating mechanisms are intractable for non-trivial cognitive tasks.

Applications and characteristics. Applications of computational cognitive models include belief propagation in social networks [Duggins, 2014; Madsen et al., 2018; Pilditch, 2017], crowd behaviour [Wijermans et al., 2013] and voter turnout [Fieldhouse et al., 2016]. They are commonly used in cognitive science to calibrate models for predictions in complex systems [Grazzini and Richiardi, 2015; Lee et al., 2015], as well as to build novel theories [Madsen et al., 2019], because they can emerge from observing and analysing the simulated behaviour of agent interaction. The key distinction between computational models and their closed-form counterparts is the setting in which they are used. When used to forecast the behaviours of dynamic, adaptable, and heterogeneous agents, models often do not have a tractable solution unless major simplifications are made [Madsen et al., 2019]. This is due to agent interactions that have an unpredictable effect on each other's behaviour and the system's dynamics. When tasks consist of a series of actions, the unpredictability of agent behaviour due to their attention lapses or belief changes, their partial observability, and complex, often unknown, behavioural dynamics (e.g., related to some external factors) make solutions to these computational cognitive models intractable. More details regarding the characteristics of these models can be found in Madsen et al. [2019].

Importance of forward simulations. Finally, in the context of this thesis, the most crucial aspect of these computational cognitive models is their capacity to allow forward simulations, which allows the use of LFI methods. Cognitive scientists can utilise LFI to calibrate and choose computational

models for specific human participants. The intricacies of inference mechanisms, with an emphasis on sample-efficient methods, are discussed in the following sections.

5.2 Bayesian experimental design

When the number of trials is restricted due to resource constraints (e.g., time or money), the experimental designs need to be carefully selected to provide as much information on the models and their parameters as possible. Bayesian experimental design (BED) [Chaloner and Verdinelli, 1995; Ryan et al., 2016] is one such technique that shares many similarities with the BO, which was briefly discussed in Chapter 3. Similarly to BO, the goal of BED is to optimise the utility function while reducing the amount of function evaluation. The distinction between the two lies in the quantity the utility function represents, which in BED quantifies the *reward* (or *regret*) associated with choosing a particular design $d \in R^+$.

Design selection. The BED utility function encourages experimental designs that reduce the resulting posterior uncertainty over the joint space of models and their parameters as much as possible. For example, the expected entropy $H(\cdot)$ of the posterior $p(m, \theta_m | \mathcal{D}_{1:t})$ at the next step $t + 1$ can be used to select the designs:

$$d_t = \arg \min_{d_t} \mathbb{E}_{p(m, \theta_m | \mathcal{D}_{1:t})} [H(m, \theta_m | \mathcal{D}_{1:t-1} \cup (d_t, x_t))] \quad (5.1)$$

$$= \arg \min_{d_t} \mathbb{E}_{p(m, \theta_m | \mathcal{D}_{1:t})} [-\log(p(x_t | d_t, \theta_m, m))] \\ + \mathbb{E}_{x_t | \mathcal{D}_{1:t-1}} \log p(x_t | d_t, \mathcal{D}_{1:t-1}). \quad (5.2)$$

Here, the decreased entropy translates to a narrower, more concentrated posterior distribution (with maximal information about models and parameters), showing the design location where the behavioural data should be collected next.

Static vs. sequential design optimization. Design of experiments can be broadly categorised into static and sequential design optimisation. In static designs, all experimental settings are determined upfront, before any data collection, leading to an inflexible but computationally manageable approach [Atkinson et al., 2007; Dean et al., 2015; Santner et al., 2003]. On the other hand, sequential design optimisation refines the experimental conditions iteratively based on observations from previous trials [Box and Draper, 1987; Chernoff, 1973; Chipman et al., 2012]. This adaptability can lead to more efficient designs, especially in situations with high uncertainty or where the design space is vast. The downside, however, is that they may demand more computational resources due to the need for continuous re-evaluation and adjustment. In the context of BED, the

decision between static and sequential optimisation leans on the trade-offs between experimental flexibility and computational burden. While static designs may suffice for well-understood systems with limited uncertainty [Berger and Wolpert, 1988; Cox, 1958], sequential designs have shown promise in scenarios demanding adaptability, such as in high-dimensional or non-linear systems [Cressie, 1985]. Moreover, modern computational tools and the surge of machine learning techniques have made sequential design increasingly accessible and effective [Gramacy and Lee, 2012; Shen and Huan, 2023]. This chapter primarily focuses on sequential designs.

Progress beyond traditional BED theory. A significant amount of progress in cognitive science has been made beyond traditional BED theory. For instance, Cavagnaro et al. [2010, 2013] offered choosing designs based on an average value of the local utilities over all possible data samples and model parameters, in which the likelihood function and the model and parameter priors were used to weight these local utilities [Myung et al., 2013]. There were other adaptations to this method, like Kim et al. [2014], which brought a hierarchical modelling viewpoint to design optimisation, allowing it to utilise previously gathered data to construct a more informative prior for model parameters. Despite the progress, these approaches remained very computationally intensive and required a tractable likelihood to work.

Reinforcement learning approaches. The recent promising direction that addresses the high computational bottleneck when evaluating utility functions and avoids the need for tractable likelihood is through reinforcement learning. The general idea of these approaches is to pre-train a reinforcement learning agent [Blau et al., 2022] in an amortised fashion that would propose design locations that maximise the utility function as its reward objective. This way, the model needs to be trained only once before the actual experimental trials begin, resulting in fast design proposals. The reinforcement learning approach provides a better exploration of the design space, does not require access to a differentiable probabilistic model, and can handle both continuous and discrete design spaces, unlike previous amortised approaches [Foster et al., 2021; Ivanova et al., 2021]. At the moment, these design optimisation methods were proposed solely for parameter inference in a single model, with the possibility of their extension to full model selection in the near future.

5.3 Likelihood-free model selection

When selecting computational models that do not have tractable likelihoods, LFI solutions are needed. Unfortunately, model selection in simulator-based models is notoriously difficult for a variety of reasons

Sisson et al. [2018], ranging from the difficulty of selecting adequate summary statistics to the significant approximation error introduced by these approaches.

Regression-based model selection. One typical method for selecting LFI models is to train a regression model that predicts the probability of the model given the data $p(m | \mathcal{D}_{1:t})$. Pudlo et al. [2016], for instance, proposed a clever procedure that first trains random forest (RF) classifiers [Ho, 1995, 1998; Parmar et al., 2018] using pairs of model labels and simulated data, where model labels are sampled from the prior. Then, these RF classifiers provide their classification error as a target and simulated data as inputs for the RF regression [Smith et al., 2013], which acts as an error calibration tool while interpreting the classification error as the predictive model posterior.

Model discrimination enhancement. Another type of methods focuses on enhancing model discrimination. For instance, Olofsson et al. [2018] have introduced GP surrogates for the approximation of the predictive distribution to compute the non-analytical model likelihood and perform model discrimination. Ouyang et al. [2018] is another example of these methods, in which they offered a probabilistic programming technique to model discrimination, effectively providing accessible adaptive design optimisation tools for cognitive scientists. Once model likelihoods are obtained, these approaches may be employed in the experimental design processes discussed in the preceding section.

Efficient model selection. Recent research on simulator-based approaches has focused on developing efficient methods for model selection, aiming to reduce the time required compared to alternative methods. For instance, Moon et al. [2022] proposed a way of creating a *generalised* (or *unified*) computational model that combines several models into one and can quickly adapt to multiple behaviours. This model is computationally cheap to evaluate, and the entire technique effectively reformulates the model selection issue as standard LFI; however, it cannot be used to improve existing models owing to its lack of interpretability. A similar effort by Kwon et al. [2020] reformulates the problem through the lenses of inverse control theory, introducing novel inference methods for sequential computational models that have the potential to increase sample-efficiency. Despite these advancements, the number of settings in which these methods can be used at the moment is somewhat restricted; nevertheless, they can still be further developed to enhance the inference of the existing models.

Extending BED and LFI approaches. In an attempt to address the issue of sample-efficient simulator-based model selection, we propose an extension of the current BED and LFI approaches in Publication III. The goal of the method is to approximate the model predictive posterior $p(\theta_m, m | \mathcal{D}_{1:t})$,

which can be later used to conduct model selection by estimating the model and its parameters using, for example, the maximum a posteriori (MAP) rule:

$$m' = \arg \max_{m \in \mathcal{M}} p(m \mid \mathcal{D}_{1:t}), \quad (5.3)$$

$$\theta' = \arg \max_{\theta \in \Theta} p(\theta \mid \mathcal{D}, m'), \quad (5.4)$$

where $p(m \mid \mathcal{D}_{1:t})$ and $p(\theta_m \mid \mathcal{D}_{1:t}, m')$ can be obtained through marginalisation of the distribution $p(\theta_m, m \mid \mathcal{D}_{1:t})$. This Bayesian formulation of the problem needs a highly accurate LFI approximation and some design selection procedure that would take into account both parameter inference and model selection. Such a formulation is compatible with other modern BED LFI techniques such as mutual information neural estimation for BED (MINEBED) [Kleinegesse and Gutmann, 2020; Valentin et al., 2021], which maximises a lower bound on the expected information gain for a particular design through a neural network trained on synthetic data. In the next section, I describe a unified method for selecting simulator-based models with a restricted number of experimental designs.

5.4 Bayesian optimization for simulator-based model selection

In cognitive science, it is essential to develop efficient model-selection methods to reduce the time required for experiments with human participants. Current model selection techniques for computational simulator-based models have limitations, such as: 1) the inability to accurately compare model likelihoods; 2) not accounting for uncertainty in both the model and its parameters (the former is needed for model calibration, i.e., fitting the behaviour); and 3) running numerous simulations, which can increase the wait time between designs to several hours (depending on the simulations), making them less efficient compared to desired methods. In Publication III, we address the first problem by proposing a LFI approximation for the model likelihood and the former two by proposing a simulator-based utility function that uses few simulations and takes into account the uncertainty around the current estimation of the model and its parameters. In this section, I describe our newly proposed method, Bayesian optimisation for simulator-based model selection (BOSMOS), which combines these solutions.

Design-conduct-infer loop. The overall procedure of BOSMOS revolves around the design-conduct-infer loop, drawing inspiration from the principles of sequential or adaptive design [Berry, 2006; Chernoff, 1992; Pallmann et al., 2018; Pukelsheim, 2006]. In sequential designs, experimental conditions are adjusted based on the outcomes of prior experiments, thereby optimising the data collection process. This is particularly useful

when experimentation is costly or time-consuming, ensuring that each new data point adds maximal information [Lindley, 1956]. Following this strategy, BOSMOS first proposes a design location for the experiments, then it collects the data generated with the proposed design, and finally it updates beliefs about all models and their parameters based on the collected data. The process repeats itself until the experimental design budget is exhausted. Notably, sequential designs have been applied in various scientific disciplines, from clinical trials [Rosenberger and Lachin, 2015] to psychology [Myung and Pitt, 2009]. In the context of BOSMOS, the LFI approximation of the marginal likelihood affects the last and most important stage of the loop: belief updates.

Belief updates with marginal likelihood approximation. The general idea of belief updates is to conduct LFI for the models separately, and then use an approximation of the marginal likelihood for model selection. In order to approximate the marginal likelihood of the models without compromising parameter inference, we frame model selection as a separate LFI problem with its own kernel, which satisfies the conditions of being non-negative, non-concave, and having a maximum at 0, as outlined by Gutmann and Corander [2016]. Specifically, we are using the Gaussian kernel:

$$\kappa_\eta(u) = \mathcal{N}(u \mid 0, \eta^2), \quad (5.5)$$

where $\eta > 0$ is the kernel bandwidth. It is straightforward to show that the value of $\kappa(\cdot)$ monotonically increases as the model m produces smaller discrepancy values.

This kernel leads to the following approximation of the likelihood, which uses predictions $\hat{\rho}$ of the GP surrogate for the discrepancy from Chapter 3 to reduce computational overhead:

$$\mathcal{L}(\mathbf{x}_t \mid m, \mathcal{D}_{t-1}) \propto \mathbb{E}_{\mathbf{x}_\theta \sim p(\cdot \mid \boldsymbol{\theta}_m, m) \cdot q(\boldsymbol{\theta}_m \mid m, \mathcal{D}_{t-1})} \kappa_\eta(\hat{\rho}(\mathbf{x}_\theta, \mathbf{x}_t)). \quad (5.6)$$

The resulting model likelihood $\mathcal{L}(\mathbf{x}_t \mid m)$ is then used in the importance-weighted sampling procedure to obtain $q(\boldsymbol{\theta}_m, m \mid \mathcal{D}_t)$:

$$q(\boldsymbol{\theta}_m, m \mid \mathcal{D}_t) \propto \mathcal{L}_{\epsilon_m}(\mathbf{x}_t \mid \boldsymbol{\theta}_m, m) \cdot \mathcal{L}(\mathbf{x}_t \mid m, \mathcal{D}_{t-1}) \cdot q(\boldsymbol{\theta}_m, m \mid \mathcal{D}_{t-1}). \quad (5.7)$$

Simulator-based utility function and design selection. The computational bottleneck of most BED methods is the design selection procedure. A good utility function for model selection should encourage exploration of the design space and facilitate the estimation of the model. Since neither $p(m, \boldsymbol{\theta}_m \mid \mathcal{D}_{1:t})$ nor (5.1) are tractable in the case of simulator-based models, we use a simulator-based approximation of the utility function. It chooses such designs \mathbf{d}_t that would maximise identifiability (maximise the entropy) between N responses \mathbf{x}' from the posterior predictive simulated from the human behaviour model $\pi(\mathbf{d}_t, \boldsymbol{\theta}_m, m)$:

$$\mathbf{d}_t = \arg \min_{\mathbf{d}_t} \mathbb{E}_{q_t(m, \boldsymbol{\theta}_m \mid \mathcal{D}_{1:t-1})} [\hat{H}(\mathbf{x}'_t \mid m, \boldsymbol{\theta}_m)]. \quad (5.8)$$

where q_t is a particle approximation of the posterior at time t , and \hat{H} is a MCMC approximation of H . The θ_m and m are sampled from current beliefs $q_t(\theta_m \mid m, \mathcal{D}_{1:t-1})$ and $q_t(m \mid \mathcal{D}_{1:t-1})$. This simulator-based approximation was shown to perform orders of magnitude faster than the alternatives while also facilitating the convergence of the model selection.

Conclusion. This chapter presented BOSMOS, an innovative simulator-based model-selection method that efficiently and accurately identifies the correct model and its parameters in controlled experiments. BOSMOS combines a marginal likelihood approximation with a simulator-based design selection procedure, which takes into account the uncertainty around the estimation of the model and its parameters. This combination proves crucial for achieving significantly faster performance compared to alternative methods while still maintaining accurate model identification and parameter fitting.

Unlike the settings from the previous chapters, here we had control over the collected observations through the experimental design variables. The ability to select informative design values while also avoiding bias in collected observations (e.g., when the designs misrepresent the ground truth) is a very challenging problem, which is exacerbated by LFI approximations. The proposed method was demonstrated to be significantly more efficient, accurately identifying the correct model and fitting the target behaviour of the participants at least two orders of magnitude faster than alternative methods. The fully Bayesian solution allows quantification of uncertainty in cognitive modelling settings where multiple model configurations may produce the same behaviour.

6. Discussion

The thesis developed sample-efficient LFI techniques relevant to surrogate modelling, dynamical systems, and experimental design for model selection. The use of hundreds of simulations for inference opens up many new possibilities for simulator-based models in the future. The suggested BO-based techniques for LFI can handle computationally expensive simulators while also addressing setting-specific issues. In this chapter, we look at possible ways to build on this thesis's contributions.

Addressing research question 1: developing a versatile and sample-efficient LFI approach. In Publication I, we addressed the challenge of handling complex noise distributions in simulators while maintaining sample-efficiency. The proposed deep GP surrogates proved to be more flexible than regular GPs; however, their inability to support amortisation limits their suitability as a general method for LFI. This limitation is directly related to sample-efficiency, as amortised approaches typically require significantly more simulations. As demonstrated in Publication II, a potential compromise involves reusing synthetic observations, necessitating only a few additional simulations. The reused synthetic observations can be employed to compute the deterministic discrepancy objective for the new observed dataset and serve as initial evidence for BO. Future work may further explore this strategy by adapting BO's acquisition function to provide simulation locations that are informative for all conceivable observed datasets. Moreover, the advent of more scalable and powerful neural network architectures may provide new opportunities for improving deep GP surrogates, allowing for enhanced sample-efficiency.

Addressing research question 2: sample-efficient LFI approach for time-series simulators. As for Publication II, we introduced a new method for inferring adaptive simulator-based models, which greatly reduces the number of required simulations by enabling state predictions. The suggested state transition dynamics model proved beneficial in behavioural cloning studies, where individuals exhibit only short-term behavioural changes. However, many real-world stochastic processes exhibit long-term dependen-

cies, such as people demonstrating strategic flexibility in tasks requiring long-term planning [Madsen et al., 2019]. Extending the concepts from Publication II to more challenging settings would require enhancing the transition dynamics model and incorporating new modelling assumptions.

Learning long-term dependencies is inherently more difficult, and this advancement would likely require far more time-series observations than the initial estimate of one hundred. Additionally, to avoid overfitting the transition dynamics, multiple time-series trajectories (such as those produced by a population of humans) will be needed. Although these concepts have not yet been applied to LFI, they have been effectively employed in reinforcement learning [Sutton and Barto, 2018] and control theory [Baleanu et al., 2021] literature. Implementing architectures inspired by transformers Lin et al. [2022]; Vaswani et al. [2017], which have shown success in modeling long-term dependencies in sequence data, could also be a promising avenue for future research in this area.

Addressing research question 3: designing a simulator-based model-selection method. In Publication III, we presented a novel approach for experimental design for simulator-based model selection that was designed to minimise simulations of computational models during the inference and design selection phases. While BOSMOS effectively addresses the challenges of the setting, it shares some limitations with other LFI approaches for model selection, such as inference failure due to an incorrect choice of the summary statistics [Fearnhead and Prangle, 2012] and cumulative error associated with LFI approximation. Although increasing the number of trials can help reduce LFI approximation error, resolving the issue of selecting summary statistics for likelihood-free model selection remains a non-trivial task and continues to be a focal point in contemporary LFI research.

An intriguing future direction for methods like BOSMOS involves adopting recent advancements in experimental design literature that utilise reinforcement learning techniques to propose non-myopic designs, i.e., designs that influence subsequent trials. This approach necessitates knowing the number of designs beforehand, whereas BOSMOS does not rely on this assumption and opts for myopic designs, which may be suboptimal in various situations. Non-myopic methods, such as those found in [Kantaros et al., 2019; Yue and Kontar, 2020], are now accessible, and some adaptations, including those based on reinforcement learning [Blau et al., 2022], hold potential for implementation in experimental design optimisation for simulator-based models. Exploring hybrid models that combine strengths of both myopic and non-myopic approaches can be a potential future research direction.

Future prospects and challenges in LFI applications. The popularity of the approaches presented in this thesis is heavily reliant on the industrial

application of simulators. Currently, in the industry, simulator-based inference is less known than other popular methods, such as reinforcement learning or control theory approaches. These other techniques often accomplish inference in a less interpretable manner. The recent paradigm of relying on purely data-driven approaches, which has been largely led by the development of deep learning techniques, overshadows the utility of applying LFI in many contexts. However, the clear benefits of data-efficiency and model-based (or in the case of this thesis, simulator-based) approaches highlight the potential importance of LFI in the coming years.

LFI holds promise in enhancing agent-based systems, which serve as platforms for exploring cognitive models in expansive social systems. These computational cognitive models allow for the probing of interventions and hypotheses that real-world experiments would deem unethical, given that the "participants" are simulated rather than real [Madsen et al., 2019]. While a unified theory that connects individual behaviors to broader population dynamics in cognitive modeling is still lacking, LFI could be the pivotal instrument for this integration.

Apart from areas where LFI is already well-established, I believe its most promising future lies in human-centric applications such as diagnostic, analytical, or support tools that bolster human creativity and expertise. However, with the potential of these systems to gather behavioral data for purposes like targeted advertising or manipulation, ensuring user privacy and upholding transparency is essential. For instance, it is crucial to adhere to responsible AI principles and practices [Arrieta et al., 2020; Dignum, 2019]. Furthermore, incorporating methods like differential privacy [Dwork, 2006] can also guarantee that individual data cannot be deciphered from simulation outcomes.

As for the future challenges, high complexity of noise distributions and interactivity are two characteristics that are anticipated to be present in the next generation of simulator-based systems, as discussed in this thesis. I expect that the established methodologies and contributions from this thesis will aid in the adoption of the simulator-based inference paradigm in future industrial applications of simulators and machine learning in general.

References

- Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014.
- Alsing, J., Wandelt, B., and Feeney, S. (2018). Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *Monthly Notices of the Royal Astronomical Society*, 477(3):2874–2885.
- An, Z., Nott, D. J., and Drovandi, C. (2020). Robust bayesian synthetic likelihood via a semi-parametric approach. *Statistics and Computing*, 30(3):543–557.
- Anderson, J. R. and Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, 2(6):396–408.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.
- Astrom, K. J. (1965). Optimal control of markov decision processes with incomplete state estimation. *J. Math. Anal. Applic.*, 10:174–205.
- Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum experimental designs, with SAS*, volume 34. OUP Oxford.
- Baleanu, D., Sajjadi, S. S., Jajarmi, A., and Deftferli, Ö. (2021). On a nonlinear dynamical system with both chaotic and nonchaotic behaviors: a new fractional analysis and control. *Advances in Difference Equations*, 2021(1):1–17.
- Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
- Barceló, J. et al. (2010). *Fundamentals of traffic simulation*, volume 145. Springer.
- Barthelmé, S. and Chopin, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315–333.
- Bassett, D. S. and Sporns, O. (2017). Network neuroscience. *Nature neuroscience*, 20(3):353–364.
- Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363.
- Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- Bechtel, W. (2013). *Philosophy of science: An overview for cognitive science*. Psychology Press.
- Berger, J. O. and Wolpert, R. L. (1988). The likelihood principle. IMS.
- Berlin B, K. P. (1969). Basic color terms: their universality and evolution. *Berkeley & Los Angeles*.
- Berry, D. A. (2006). Bayesian clinical trials. *Nature reviews Drug discovery*, 5(1):27–36.
- Bharti, A., Filstroff, L., and Kaski, S. (2022). Approximate bayesian computation with domain expert in the loop. *arXiv preprint arXiv:2201.12090*.
- Bi, J., Shen, W., and Zhu, W. (2022). Random forest adjustment for approximate bayesian computation. *Journal of Computational and Graphical Statistics*, 31(1):64–73.
- Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.
- Blau, T., Bonilla, E. V., Chades, I., and Dezfouli, A. (2022). Optimizing sequential experimental design with deep reinforcement learning. In *International Conference on Machine Learning*, pages 2107–2128. PMLR.
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.
- Bonassi, F. V. and West, M. (2015). Sequential monte carlo with adaptive weights for approximate bayesian computation. *Bayesian Analysis*, 10(1):171–187.
- Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.
- Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.
- Calvet, L. E. and Czellar, V. (2015). Accurate methods for approximate bayesian computation filtering. *Journal of Financial Econometrics*, 13(4):798–838.
- Cavagnaro, D. R., Myung, J. I., Pitt, M. A., and Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural computation*, 22(4):887–905.
- Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., and Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of risk and uncertainty*, 47(3):255–289.
- Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.
- Chen, X., Acharya, A., Oulasvirta, A., and Howes, A. (2021). An adaptive model of gaze-based selection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11.
- Chernoff, H. (1973). *Approaches in sequential design of experiments*. Stanford University. Department of Statistics.
- Chernoff, H. (1992). *Sequential design of experiments*. Springer.

- Chipman, H., Ranjan, P., and Wang, W. (2012). Sequential design for computer experiments with a flexible bayesian additive model. *Canadian Journal of Statistics*, 40(4):663–678.
- Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). Smc2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426.
- Collins, A. (1977). Why cognitive science. *Cognitive Science*, 1(1):1–2.
- Connor, J. T., Martin, R. D., and Atlas, L. E. (1994). Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254.
- Cox, D. R. (1958). Planning of experiments.
- Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the international Association for mathematical Geology*, 17:563–586.
- Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418.
- Curi, S., Melchior, S., Berkenkamp, F., and Krause, A. (2020). Structured variational inference in partially observable unstable gaussian process state space models. In *Learning for Dynamics and Control*, pages 147–157. PMLR.
- Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.
- Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. (2022). Multi-objective bayesian optimization over high-dimensional search spaces. In *Uncertainty in Artificial Intelligence*, pages 507–517. PMLR.
- Daw, N. D., O’doherly, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879.
- Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature neuroscience*, 10(8):1020–1028.
- Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.
- Dean, A. M., Morris, M., Stufken, J., and Bingham, D. (2015). *Handbook of design and analysis of experiments*, volume 7. CRC Press Boca Raton, FL, USA:.
- Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and computing*, 22(5):1009–1020.
- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.
- Doerr, A., Daniel, C., Schiegg, M., Duy, N.-T., Schaal, S., Toussaint, M., and Sebastian, T. (2018). Probabilistic recurrent state-space models. In *International Conference on Machine Learning*, pages 1280–1289. PMLR.
- Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556.

- Duggins, P. (2014). A psychologically-motivated model of opinion change with applications to american politics. *arXiv preprint arXiv:1406.7770*.
- Dunlop, M. M., Girolami, M. A., Stuart, A. M., and Teckentrup, A. L. (2018). How deep are deep Gaussian processes? *The Journal of Machine Learning Research*, 19(1):2100–2145.
- Durkan, C., Papamakarios, G., and Murray, I. (2018). Sequential neural methods for likelihood-free inference. *arXiv preprint arXiv:1811.08723*.
- Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.
- Edgeworth, F. Y. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2):381–397.
- Esposito, P. (2020). BLiTz - Bayesian layers in Torch zoo (a Bayesian deep learning library for torch). <https://github.com/piEsposito/blitz-bayesian-deep-learning/>.
- Falk, G. (1972). Interpretation of imperfect line data as a three-dimensional scene. *Artificial intelligence*, 3:101–144.
- Fan, Y., Nott, D. J., and Sisson, S. A. (2013). Approximate bayesian computation via regression density estimation. *Stat*, 2(1):34–48.
- Fanshawe, T. R. and Diggle, P. J. (2012). Bivariate geostatistical modelling: a review and an application to spatial variation in radon concentrations. *Environmental and ecological statistics*, 19(2):139–160.
- Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.
- Fieldhouse, E., Lessard-Phillips, L., and Edmonds, B. (2016). Cascade or echo chamber? a complex agent-based simulation of voter turnout. *Party Politics*, 22(2):241–256.
- Flury, T. and Shephard, N. (2011). Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Econometric Theory*, 27(5):933–956.
- Foster, A., Ivanova, D. R., Malik, I., and Rainforth, T. (2021). Deep adaptive design: Amortizing sequential bayesian experimental design. In *International Conference on Machine Learning*, pages 3384–3395. PMLR.
- Frank, R. J., Davey, N., and Hunt, S. P. (2001). Time series prediction and neural networks. *Journal of intelligent and robotic systems*, 31(1):91–103.
- Frazier, D. T. and Drovandi, C. (2021). Robust approximate bayesian inference with synthetic likelihood. *Journal of Computational and Graphical Statistics*, 30(4):958–976.
- Frazier, D. T., Nott, D. J., Drovandi, C., and Kohn, R. (2022). Bayesian inference using synthetic likelihood: asymptotics and adjustments. *Journal of the American Statistical Association*, (just-accepted):1–28.
- Fritzsche, H.-T. and Ag, D.-b. (1994). A model for traffic simulation. *Traffic Engineering+ Control*, 35(5):317–21.
- Georgiou, T. and Demiris, Y. (2017). Adaptive user modelling in car racing games using behavioural and physiological data. *User Modeling and User-Adapted Interaction*, 27(2):267–311.

- Gibbs, M. N. (1998). *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.
- Gimenez, O., Rossi, V., Choquet, R., Dehais, C., Doris, B., Varella, H., Vila, J.-P., and Pradel, R. (2007). State-space modelling of data on marked individuals. *ecological modelling*, 206(3-4):431–438.
- Gramacy, R. B. and Lee, H. K. (2012). Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22:713–722.
- Grazzini, J. and Richiardi, M. (2015). Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 51:148–165.
- Grenfell, B. T., Bjørnstad, O. N., and Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723.
- Gutmann, M. U. and Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*.
- Hacking, I. (1967). Slightly more realistic personal probability. *Philosophy of Science*, 34(4):311–325.
- Hahn, U. and Oaksford, M. (2006). A normative theory of argument strength. *Informal Logic*, 26(1):1–24.
- Hahn, U. and Oaksford, M. (2007). The rationality of informal argumentation: a bayesian approach to reasoning fallacies. *Psychological review*, 114(3):704.
- Hasegawa, T., Niida, A., Mori, T., Shimamura, T., Yamaguchi, R., Miyano, S., Akutsu, T., and Imoto, S. (2016). A likelihood-free filtering method via approximate bayesian computation in evaluating biological simulation models. *Computational Statistics & Data Analysis*, 94:63–74.
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.
- Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 7506–7516.
- Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pages 732–740. PMLR.
- Hermans, J., Begy, V., and Louppe, G. (2020). Likelihood-free MCMC with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–4248. PMLR.
- Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

- Holden, P. B., Edwards, N. R., Hensman, J., and Wilkinson, R. D. (2018). Abc for climate: dealing with expensive simulators. In *Handbook of approximate Bayesian computation*, pages 569–595. Chapman and Hall/CRC.
- Hunt, E. (1989). Cognitive science: Definition, status, and questions. *Annual Review of psychology*, 40(1):603–629.
- Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F., and Nardi, L. (2022). $\backslash\pi$ bo: Augmenting acquisition functions with user beliefs for bayesian optimization. *arXiv preprint arXiv:2204.11051*.
- Ialongo, A. D., Van Der Wilk, M., Hensman, J., and Rasmussen, C. E. (2019). Overcoming mean-field approximations in recurrent gaussian process models. In *International Conference on Machine Learning*, pages 2931–2940. PMLR.
- Ivanova, D. R., Foster, A., Kleinegesse, S., Gutmann, M. U., and Rainforth, T. (2021). Implicit deep adaptive design: policy-based experimental design without likelihoods. *Advances in Neural Information Processing Systems*, 34:25785–25798.
- Izbicki, R., Lee, A. B., and Pospisil, T. (2019). Abc-cde: Toward approximate bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, 28(3):481–492.
- Jabot, F., Lagarrigues, G., Courbaud, B., and Dumoulin, N. (2014). A comparison of emulation methods for approximate bayesian computation. *arXiv preprint arXiv:1412.7560*.
- Järvenpää, M., Gutmann, M. U., Pleska, A., Vehtari, A., and Marttinen, P. (2019). Efficient acquisition rules for model-based approximate bayesian computation. *Bayesian Analysis*, 14(2):595–622.
- Järvenpää, M., Gutmann, M. U., Vehtari, A., and Marttinen, P. (2021). Parallel gaussian process surrogate bayesian inference with noisy likelihood evaluations. *Bayesian Analysis*, 16(1):147–178.
- Jasra, A., Singh, S. S., Martin, J. S., and McCoy, E. (2012). Filtering via approximate bayesian computation. *Statistics and Computing*, 22(6):1223–1237.
- Jeffrey, N., Alsing, J., and Lanusse, F. (2021). Likelihood-free inference with neural compression of des sv weak lensing map statistics. *Monthly Notices of the Royal Astronomical Society*, 501(1):954–969.
- Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.
- Joo, T. W. and Kim, S. B. (2015). Time series forecasting based on wavelet filtering. *Expert Systems with Applications*, 42(8):3868–3874.
- Juang, B. H. and Rabiner, L. R. (1991). Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272.
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.
- Kantaros, Y., Schlotfeldt, B., Atanasov, N., and Pappas, G. J. (2019). Asymptotically optimal planning for non-myopic multi-robot information gathering. In *Robotics: Science and Systems*, pages 22–26.
- Kattwinkel, M. and Reichert, P. (2017). Bayesian parameter inference for individual-based models using a particle markov chain monte carlo method. *Environmental modelling & software*, 87:110–119.

- Kim, W., Pitt, M. A., Lu, Z.-L., Steyvers, M., and Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural computation*, 26(11):2465–2492.
- Kim, Y.-S., Walls, L. A., Krafft, P., and Hullman, J. (2019). A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14.
- Kleinegesse, S. and Gutmann, M. U. (2020). Bayesian experimental design for implicit models by mutual information neural estimation. In *International Conference on Machine Learning*, pages 5316–5326. PMLR.
- Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
- Kononenko, I. (1989). Bayesian neural networks. *Biological Cybernetics*, 61(5):361–370.
- Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise.
- Kwon, M., Daptardar, S., Schrater, P. R., and Pitkow, X. (2020). Inverse rational control with partially observable continuous nonlinear dynamics. *Advances in neural information processing systems*, 33:7898–7909.
- Lapedes, A. and Farber, R. (1987). Nonlinear signal processing using neural networks: Prediction and system modelling. Technical report.
- Lee, J.-S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., Voinov, A., Polhill, J. G., Sun, Z., and Parker, D. C. (2015). The complexities of agent-based modeling output analysis. *Journal of Artificial Societies and Social Simulation*, 18(4).
- Lee, M. D. (2018). Bayesian methods in cognitive modeling. *The Stevens' handbook of experimental psychology and cognitive neuroscience*, 5:37–84.
- Leibfried, F., Dutoir, V., John, S., and Durrande, N. (2020). A tutorial on sparse gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*.
- Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.
- Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*.
- Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.
- Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1):e66–e82.
- Littman, M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125.
- Liu, C., Hoi, S. C., Zhao, P., and Sun, J. (2016). Online arima algorithms for time series prediction. In *Thirtieth AAAI conference on artificial intelligence*.
- Liu, H., Cai, J., and Ong, Y.-S. (2018). Remarks on multi-output gaussian process regression. *Knowledge-Based Systems*, 144:102–121.
- Liu, X., Zhuo, Z., Du, X., Zhang, X., Zhu, Q., and Guizani, M. (2019). Adversarial attacks against profile hmm website fingerprinting detection model. *Cognitive Systems Research*, 54:83–89.

- Loftin, R. and Oliehoek, F. A. (2022). On the impossibility of learning to cooperate with adaptive partner strategies in repeated games. In *International Conference on Machine Learning*, pages 14197–14209. PMLR.
- Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. (2019). Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53. PMLR.
- Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30.
- Madsen, J. K., Bailey, R., Carrella, E., and Koralus, P. (2019). Analytic versus computational cognitive models: Agent-based modeling as a tool in cognitive sciences. *Current Directions in Psychological Science*, 28(3):299–305.
- Madsen, J. K., Bailey, R. M., and Pilditch, T. D. (2018). Large networks of rational agents form persistent echo chambers. *Scientific reports*, 8(1):1–8.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.
- Marr, D. and Nishihara, H. K. (1976). Representation and recognition of the spatial organization of three dimensional shapes. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.
- Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.
- McCarthy, D. M., McCarty, K. N., Hatz, L. E., Prestigiacomo, C. J., Park, S., and Davis-Stober, C. P. (2021). Applying bayesian cognitive models to decisions to drive after drinking. *Addiction*, 116(6):1424–1430.
- McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1).
- Meeds, E. and Welling, M. (2014). Gps-abc: Gaussian process surrogate approximate bayesian computation. *arXiv preprint arXiv:1401.2838*.
- Moćkus, J. (1975). On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer.
- Moon, H.-S., Do, S., Kim, W., Seo, J., Chang, M., and Lee, B. (2022). Speeding up inference with user simulators through policy modulation. In *CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Myung, J. I., Cavagnaro, D. R., and Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3-4):53–67.
- Myung, J. I. and Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological review*, 116(3):499.
- Nilsson, M. and Ejnarsson, M. (2002). Speech recognition using hidden markov model.
- Nott, D. J., Fan, Y., Marshall, L., and Sisson, S. (2014). Approximate bayesian computation and bayes’ linear analysis: toward high-dimensional abc. *Journal of Computational and Graphical Statistics*, 23(1):65–86.
- Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical applications in genetics and molecular biology*, 9(1).

- Ober, S. W., Rasmussen, C. E., and van der Wilk, M. (2021). The promises and pitfalls of deep kernel learning. In *Uncertainty in Artificial Intelligence*, pages 1206–1216. PMLR.
- O’Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.
- O’Hagan, A., Kendall, M. G., and Forster, J. (2004). *Kendall’s Advanced Theory of Statistics: Bayesian Statistics. Vol. 2B*. Arnold.
- Olofsson, S., Deisenroth, M., and Misener, R. (2018). Design of experiments for model discrimination hybridising analytical and data-driven approaches. In *International Conference on Machine Learning*, pages 3908–3917. PMLR.
- Ong, V. M., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018a). Variational bayes with synthetic likelihood. *Statistics and Computing*, 28(4):971–988.
- Ong, V. M.-H., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018b). Likelihood-free inference in high dimensions with synthetic likelihood. *Computational Statistics & Data Analysis*, 128:271–291.
- Orr, M. G. and Plaut, D. C. (2014). Complex systems and health behavior change: insights from cognitive science. *American journal of health behavior*, 38(3):404–413.
- Orr, M. G., Thrush, R., and Plaut, D. C. (2013). The theory of reasoned action as parallel constraint satisfaction: Towards a dynamic computational model of health behavior. *PloS one*, 8(5):e62490.
- Ouyang, L., Tessler, M. H., Ly, D., and Goodman, N. D. (2018). webppl-oed: A practical optimal experiment design system. In *CogSci*.
- Paciorek, C. and Schervish, M. (2003). Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems*, 16.
- Paige, B. and Wood, F. (2016). Inference networks for sequential monte carlo in graphical models. In *International Conference on Machine Learning*, pages 3040–3049. PMLR.
- Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odoni, L., Sydes, M. R., et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1):1–15.
- Papamakarios, G. and Murray, I. (2016). Fast ε -free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29.
- Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30.
- Papamakarios, G., Sterratt, D., and Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR.
- Parmar, A., Katariya, R., and Patel, V. (2018). A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things*, pages 758–763. Springer.

- Peters, G. W., Sisson, S. A., and Fan, Y. (2012). Likelihood-free bayesian inference for α -stable models. *Computational Statistics & Data Analysis*, 56(11):3743–3756.
- Pfanzagl, J. (2011). *Parametric statistical theory*. Walter de Gruyter.
- Pilditch, T. D. (2017). Opinion cascades and echo-chambers in online networks: A proof of concept agent-based model. Cognitive Science Society.
- Pleiss, G. and Cunningham, J. P. (2021). The limitations of large width in neural networks: A deep gaussian process perspective. *Advances in Neural Information Processing Systems*, 34:3349–3363.
- Prangle, D. (2017). Adapting the abc distance function.
- Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11.
- Priddle, J. W., Sisson, S. A., Frazier, D. T., Turner, I., and Drovandi, C. (2022). Efficient bayesian synthetic likelihood with whitening transformations. *Journal of Computational and Graphical Statistics*, 31(1):50–63.
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798.
- Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). Reliable abc model choice via random forests. *Bioinformatics*, 32(6):859–866.
- Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.
- Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*.
- Robert, C. P., Beaumont, M. A., Marin, J.-M., and Cornuet, J.-M. (2008). Adaptivity for abc algorithms: the abc-pmc scheme. *arXiv preprint arXiv:0805.2256*.
- Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.
- Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, 4(3):328–350.
- Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154.
- Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599.
- Salimbeni, H., Dutordoir, V., Hensman, J., and Deisenroth, M. (2019). Deep gaussian processes with importance-weighted variational inference. In *International Conference on Machine Learning*, pages 5589–5598. PMLR.

- Santner, T. J., Williams, B. J., Notz, W. I., and Williams, B. J. (2003). *The design and analysis of computer experiments*, volume 1. Springer.
- Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE computational intelligence magazine*, 4(2):24–38.
- Schafer, C. M. and Freeman, P. E. (2012). Likelihood-free inference in cosmology: Potential for the estimation of luminosity functions. In *Statistical Challenges in Modern Astronomy V*, pages 3–19. Springer.
- Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2):241–263.
- Seeger, M. (2004). Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106.
- Shafi, K., Latif, N., Shad, S. A., Idrees, Z., and Gulzar, S. (2018). Estimating option greeks under the stochastic volatility using simulation. *Physica A: Statistical Mechanics and its Applications*, 503:1288–1296.
- Shen, W. and Huan, X. (2023). Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 416:116304.
- Shenoy, K. V., Sahani, M., and Churchland, M. M. (2013). Cortical control of arm movements: a dynamical systems perspective. *Annual review of neuroscience*, 36:337–359.
- Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society*, pages 113–144.
- Sisson, S. A. and Fan, Y. (2011). Likelihood-free mcmc. *Handbook of Markov Chain Monte Carlo*, pages 313–335.
- Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press.
- Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.
- Smith, P. F., Ganesh, S., and Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods*, 220(1):85–91.
- Spaan, M. T. (2012). Partially observable markov decision processes. In *Reinforcement Learning*, pages 387–414. Springer.
- Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.
- Staudenmayer, J. and Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association*, 100(471):841–852.
- Sun, R. et al. (2006). *Cognition and multi-agent interaction: From cognitive modeling to social simulation*. Cambridge University Press.
- Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Comput Biol*, 9(1):e1002803.

- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Tamerius, J., Nelson, M. I., Zhou, S. Z., Viboud, C., Miller, M. A., and Alonso, W. J. (2011). Global influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental health perspectives*, 119(4):439–445.
- Tang, H. and Dong, C. (2019). Detection of malicious domain names based on an improved hidden markov model. *International Journal of Wireless and Mobile Computing*, 16(1):58–65.
- Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518.
- Thissen, U., Van Brakel, R., De Weijer, A., Melssen, W., and Buydens, L. (2003). Using support vector machines for time series prediction. *Chemometrics and intelligent laboratory systems*, 69(1-2):35–49.
- Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR.
- Udny Yule, G. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer’s sunspot numbers. *Philosophical Transactions of the Royal Society of London Series A*, 226:267–298.
- Valentin, S., Kleinegesse, S., Bramley, N. R., Gutmann, M. U., and Lucas, C. G. (2021). Bayesian optimal experimental design for simulator models of cognition. *arXiv preprint arXiv:2110.15632*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Verma, A., Dai, Z., and Low, B. K. H. (2022). Bayesian optimization under stochastic delayed feedback. In *International Conference on Machine Learning*, pages 22145–22167. PMLR.
- Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., and Wilson, A. G. (2019). Exact gaussian processes on a million data points. *Advances in Neural Information Processing Systems*, 32.
- Weigend, A. S. (2018). *Time series prediction: forecasting the future and understanding the past*. Routledge.
- Wijermans, N., Jorna, R., Jager, W., van Vliet, T., and Adang, O. (2013). Cross: modelling crowd behaviour with social-cognitive agents. *Journal of Artificial Societies and Social Simulation*, 16(4):1.
- Wilkinson, R. (2014). Accelerating abc methods using gaussian processes. In *Artificial Intelligence and Statistics*, pages 1015–1023. PMLR.
- Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.
- Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR.
- Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.
- Wu, X. and Wang, Y. (2012). Extended and unscented kalman filtering based feedforward neural networks for time series prediction. *Applied Mathematical Modelling*, 36(3):1123–1131.

- Yang, C., Gao, Z., Liu, F., and Ma, R. (2020). Extended kalman filters for nonlinear fractional-order systems perturbed by colored noises. *ISA transactions*, 102:68–80.
- Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.
- Yu, F., Koltun, V., and Funkhouser, T. (2017). Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480.
- Yue, X. and Kontar, R. A. (2020). Why non-myopic bayesian optimization is promising and how far should we look-ahead? a study via rollout. In *International Conference on Artificial Intelligence and Statistics*, pages 2808–2818. PMLR.
- Zerdali, E. and Barut, M. (2017). The comparisons of optimized extended kalman filters for speed-sensorless control of induction motors. *IEEE Transactions on industrial electronics*, 64(6):4340–4351.
- Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115.

Errata

Publication II

Table 1 provides approximate orders of magnitude for the number of simulations needed by each method to achieve similar accuracy levels. These figures aim to highlight differences in sample-efficiency across methods rather than offer a qualitative analysis. The exact numbers will vary based on specific task properties, such as the dimensionality of simulator parameters or outputs. Furthermore, reference [40] should correctly be cited as Lueckmann et al. [2017].



ISBN 978-952-64-1555-0 (printed)

ISBN 978-952-64-1556-7 (pdf)

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

Aalto University
School of Science
Computer Science
www.aalto.fi

**BUSINESS +
ECONOMY**

**ART +
DESIGN +
ARCHITECTURE**

**SCIENCE +
TECHNOLOGY**

CROSSOVER

**DOCTORAL
THESES**