Minerva Schools at KGI

Global Surrogate Likelihood-Free Inference

Braden Scherting

March 24, 2020

Abstract

Probabilistic inference on models lacking tractable likelihoods is important in a range of disciplines. However, such models evade standard posterior inference, leading researchers to employ likelihood-free inference methods, which rely on simulations from the likelihood-free models. Likelihood-free inference methods have undergone substantial development but struggle to solve problems with multiple observations, especially when the generative process defined by the model is computationally expensive to simulate. In this Capstone, we develop the theory of inverse probabilistic modeling and introduce a novel workflow designed specifically to addresses multiple observations and expensive simulators. The workflow harnesses neural density estimation to learn a global, surrogate approximation to the likelihood, which is used in conjunction with an existing sample-efficient likelihood-free inference method. The workflow is shown to reduce the number of simulations required by 5-10 times on well-known problems, which serves as a proof-of-concept.

Contents

Author's Note

This document is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science at Minerva Schools at KGI. An earlier version of the workflow presented below was first proposed by Henri Pesonen. Both he and Professor Samuel Kaski were instrumental in supporting and directing the early stages of this work. I am immensely grateful for their contributions. Refinements, extensions, implementation, and validation of the workflow as well as all other material presented here are solely the work of the author.

The paper is organized such that each section builds on the previous. Those familiar with Bayesian inference as it applies to inverse modeling may prefer to proceed to section 2.3 directly. The Python file implementing real NVP and Jupyter notebooks for performing experiments are included in the secondary materials submission.

I would additionally like to thank Professors Ribeiro and Scheffler for supporting my interest in Bayesian inference and Professor Wilkins for his support throughout the development of this deliverable.

1. Introduction

Bayesian statistics is inherently model-based; data is incorporated into a prior belief according to a model or likelihood function (Gelman et al., 2013). It is therefore intrinsically related to the practice (Shen et al., 2019) and philosophy of science (Gelman and Shalizi, 2012). Classical Bayesian statistics methods require access to a tractable, often statistical likelihood function (MacKay, 2003). However, expressive models of naturally occurring complex or chaotic phenomena are commonly noisy, highly non-linear, and altogether evasive of standard, classical statistical treatment (Wood, 2010). Such expressive models may be straightforward to implement as mechanistic or rule-based simulations (Lueckmann et al., 2017) or systems of differential equations (Owen et al., 2015); forward simulation is easy, but computing the likelihood is impossible. Despite the intractability of the likelihoods, performing statistical inference on models of this nature is of great interest.

Methods for performing statistical inference on likelihood-free models have been developed under a variety of names. Approximate Bayesian Computation proceeds by simulating data from the model at parameter locations selected from a proposal distribution and retaining the parameters as samples from the posterior if the simulated data is similar to the observed data (Beaumont et al., 2002; Lintusaari et al., 2017). Likelihood-free inference methods include approximate Bayesian computation as a special case but additionally include density estimation (Alsing et al., 2019) and optimization (Gutmann et al., 2016) strategies. Simulator-based inference methods are more general still (Cranmer et al., 2019). Collectively, these methods provide principled approaches to posterior inference on likelihood-free and simulator-based models.

Computational resources are a major concern when performing likelihood-free inference, with some methods requiring hundreds of thousands of simulations to converge (Lintusaari et al., 2017). Further, the same process often needs to be repeated for independent observations. Accordingly, sample efficiency and amortization of inference have been identified as principal shortcomings of contemporary likelihood-free inference (Cranmer et al., 2019). Performing statistical inference on computationally intensive simulators for multiple, independent observations is, in many cases, presently intractable.

In this Capstone, I present a novel workflow for performing multi-observation likelihood-free inference on expensive, simulator-based models for which tractable likelihoods are unavailable. The formulation of the workflow was first contemplated in collaboration with Henri Pesonen and Professor Samuel Kaski as part of the Aalto Science Institute internship program. Further development, implementation, and experimentation were additionally performed by the author. The workflow builds on developments in Bayesian optimization for likelihood-free inference (Brochu et al., 2010; Gutmann et al., 2016) and density estimation methods (Bishop, 1994; Alsing et al., 2019; Dinh et al., 2017). The workflow efficiently produces posterior inferences for multiple observations by learning an approximation to the intractable likelihood from accumulated data. The expensive simulator is then replaced by the surrogate likelihood to complete inference. I demonstrate the ability of the method to recover the target posteriors with order-of-magnitude fewer expensive simulations and note limitations.

The remainder of the paper is organized as follows. I begin by introducing Bayesian inference as a solution to inverse problems or parameter estimation. Next, a discussion of the various flavors of Bayesian inference leads into a review of likelihood-free inference techniques, with an emphasis on the tools and methods employed in the workflow. The workflow is then presented, along with accompanying implementation notes. Experiments follow, which first validate the workflow, then test behavior in higher dimensions, and lastly demonstrate its ability to, in select cases, fully amortize inference. The paper concludes with a discussion of the implications, limitations, and possible extensions of the work.

2. Background

2.1. Modeling

The language and procedure described below can be summarized as follows: forward modeling defines parameters and relational properties of a system; inverse modeling determines the values of parameters corresponding to observed data conditional on the pre-defined, forward relational properties. A useful criterion to apply to problems generally is that of well-posedness in the sense of Hadamard (Hadamard, 1902; von Würtemberg, 2011). Borrowed from mathematical physics, this criterion categorizes problems by their solutions. The solution to a well-posed problem satisfies 1) existence, 2) uniqueness, and 3) continuity with respect to initial conditions. Violation of any of these criteria yields an ill-posed problem. Many complex and all chaotic dynamical systems, for example, exhibit strong sensitivity to initial conditions and are therefore ill-posed.

It should be noted that many definitions of modeling are loose or domain-specific. What follows are definitions that facilitate understanding of the remainder of the paper but are not definintive. Much of the following discussion is informed and inspired by Kabanikhin (2008).

2.1.1. Forward Problems

Forward or direct problems arise naturally from a desire to explain or predict systems and phenomena around us. Fundamentally, the forward direction of inquiry proceeds from causes to effects. To that end, a general solution (valid for all parameters in the domain) to a forward problem is a model that describes the system or phenomenon in question as some function of parameters and inputs. Formulating the forward problem additionally entails specifying a domain and boundary conditions, features which we will assume to be implicit or automatic henceforth. In mathematical and statistical modeling, the function is commonly given by equations or systems of equations. I am careful, however, to not restrict the definition of the solution function to "equations" because a central objective of this project is to enable modeling of systems that are not best represented by equations. Tunable values in the model, which modulate the relationships between random or independent quantities without changing the governing logic of the relationships, are *parameters*, represented in the general case by θ. Parameters are generally unobserved or unobservable quantities, despite often entertaining natural interpretations. In equation-based models, common types of parameters are coefficients and initial conditions. A parameterized model serves to predict or explain observable quantities. Those observable quantities are *data*, $\mathscr{D}$, and include independent (explanatory) variables x and dependent variables y. I denote a model by $\mathscr{M}$, which represents one possible model from a potentially infinite space of models. In summary, models are sets of parameter-tuned relationships that define a forward map from parameters to data, $\mathscr{M} : (\theta, x) \rightarrow y$. What is commonly referred to as "modeling" is synonymous with solving the forward problem.

2.1.2. Inverse Problems

Definitions of inverse problems exist at various levels of granularity. For our purposes, the inverse problem corresponding to a given forward model $\mathscr{M}^*$ is: *what values θ^* explain or produce $\mathscr{D}^*$ conditional on $\mathscr{M}^*$?* Inverse problems prompt us to identify a parameterization or parameterizations of our model that correspond to observations. This task is contingent on the choice/specification of forward model. Thus, the terminology is perhaps misleading; from the outset, forward modeling precedes inverse modeling. Inverse modeling occurs only once assumptions about the forward model have been made. However, forward models can and should be iterated conditional on the solution to the associated inverse problem.

Forward modeling is often constructive–designing relationships based on knowledge about a system, which can often be done straightforwardly, even for chaotic systems. Inverse problems therefore bear the brunt of ill-posedness. For even moderately complicated models, it is common for errors to propagate through the model, compounding and interacting in complicated ways. As a result, models that are straightforward to implement but exhibit severe instability with respect to initial conditions are impossible to invert exactly and immensely difficult to invert approximately.

Though inverse problems are challenging, they are necessary for constructing reasoned, expressive models of systems. Because parameters are generally unobserved quantities, modeling without performing any sort of inverse or backward inference is simply guessing-and-checking. In general, solutions to inverse problems are valuable in two ways: 1) direct knowledge about semantically meaningful parameter values and 2) access to a parameterized model with which to reason about future or unobserved data. The first mode is explanatory; knowledge about the parameters translates directly to knowledge about the system. The second is predictive; the parameterized model can be used to predict past or future data. These modes are, of course, neither mutually exclusive nor exhaustive.

Solving inverse problems (inferring model parameters from data) is the focus of entire academic and professional disciplines and is the focus of this project.

2.2. Solving Inverse Problems

Inverse problems arise from observing data and endeavouring to learn more about the model that generated or explains them. Here I discuss the two general strategies for doing so, with an emphasis on the strategy predominantly employed in this paper: Bayesian inference. Statistical inference, learning about parameters based on observations, is the form of inverse modeling contemplated here.

2.2.1. Deterministic inference

Deterministic inference methods, such as maximum likelihood estimation and regularization, enforce criteria such as fit to data or smoothness. By doing so, the problem is simplified, endowing well-posedness. Under the enforced, altered condition, we can then approximately solve the inverse problem. The method of least squares, for example, casts statistical inference as an optimization problem wherein the minimization objective is the sum of squared residuals,

$$\sum_{i=1}^{N}(y^{i}_{obs} - \mathscr{M}(x^{i}_{obs}, \theta))^{2}. \tag{2.1}$$

To call statistical inference of this nature "deterministic" is reductionist. Many so-called deterministic methods employ stochasticity and other probability-theoretic tools. Maximum likelihood estimation, by definition, requires a likelihood, which is an eminently probabilistic object. Deterministic methods, whether they be stochastic gradient descent or interpolation, frame the task of parameter estimation as one of seeking and finding *the* parameter(s); solutions are point estimates.

Upon defining a model (e.g., linear regression: $y = mx + b$), a common, general strategy of optimization methods is to define a differentiable objective (e.g., sum of squared residuals), and employ an efficient optimum-finding algorithm. Only for simple models is the optimization convex; modern, flexible models often force us to settle for locally optimal solutions. Optimization-based statistical inference and maximum likelihood estimation (MLE) play an important, supporting role in this project.

Countless optimization objectives or regularization criteria can be employed to solve inverse problems deterministically. A major concern with these approaches, however, is how to select the type and strength of regularization. In machine learning, the choice of regularization is is synonymous with crafting a loss function. The strength of regularization can be modulated empirically by use of cross-validation and hyperparameter tuning. In any case, the result of deterministic inverse problem solving is a point estimate of the solution parameters.

2.2.2. Probabilistic Inference

The primary statistical inference perspective interrogated in this project is Bayesian inference. Regularization- and optimization-based inverse problem solving methods commonly incorporate probability. However, Bayesian inference formaly regards the unknown parameter values as belonging to probability distributions. Thus, the governing theorem of Bayesian inference (Bayes' theorem or Bayes' rule) is derived directly from statements of probability theory.

Bayesian inference is a formal approach to probabilistically solving ill-posed inverse problems. The strategy targets the conditional probability of parameters given data $p(\theta|y)$. Let's now examine Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2.2}$$

Working backward to arrive at this statement, consider two statements of conditional probability from the product rule of probability (we introduce A and B as any two arbitrary quantities):

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{2.3}$$

$$p(B|A) = \frac{p(B,A)}{p(A)} \tag{2.4}$$

Notice that joint probabilities are symmetric: $p(B,A) = p(A,B)$. Thus, (2.4) can be reexpressed and rearranged as

$$p(B|A) = \frac{p(A,B)}{p(A)} \tag{2.5}$$

$$p(A,B) = p(B|A)p(A). \tag{2.6}$$

We now have a statement that can be substituted into (2.3), arriving at Bayes' rule:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.7}$$

There are numerous ways to read and interpret the four terms of Bayes' rule. I present each in turn along with interpretations that facilitate the subsequent discussion. Upon developing familiarity with the terms, I proceed to a discussion of the practice of Bayesian inference.

PRIOR: $p(\theta)$ Prior–a fitting name for the term representing prior beliefs. Bayes' theorem requires that we formally, probabilistically express existing beliefs about the parameter values in the prior. This is a probability distribution over parameters, and, in non-hierarchical models, it is a marginal probability. If the system being modeled has never before been investigated and there is no information about possible values for the parameters, the prior should be constructed to express this uncertainty (i.e., uninformative prior). Alternatively, if the system has been extensively researched, the prior should convey the conclusions of the past research (i.e., informative prior).

A popular interpretation of the theorem is as an *update*. Whereas deterministic methods introduce assumptions via regularizers, Bayesian methods introduce assumptions about parameter values via probabilistic expressions of prior beliefs. Any analysis begins with a preconceived belief about the values of parameters (even if that belief is very uncertain). The belief may follow from previous analyses or from fundamental constraints, e.g., populations are positive. Upon observing some new data, we would like neither to replace old beliefs nor disregard the new data. Rather, we would like to incorporate new evidence into existing beliefs, thereby arriving at an updated expression for the belief. Because the prior seemingly entails subjectivity, it is a feature of Bayesian inference that often receives criticism. And, indeed, inappropriately crafting a prior distribution can lead to severely misleading results. However, the prior is, in large part, the defining feature of Bayesian inference and enables conclusive inference on underconstrained problems. Furthermore, priors should involve minimal, if any, subjectivity. Quantifying prior knowledge should be evidence-based or follow directly from previous inferences. See Chapter 1 of McElreath (2018) for a description of Bayesian analysis devoid of references to beliefs or subjectivity.

NORMALIZING CONSTANT: $p(y)$ Despite being a constant, this term is a source of unending angst. The normalizing constant (or marginal probability of data or model evidence) gives the marginal probability of the observed data conditioned only on the model and can be best understood from its definition:

$$p(y) = \int p(y|\theta)p(\theta)d\theta. \tag{2.8}$$

This term answers the question, What is the probability of the data, under the model but irrespective of the parameter values? A practical interpretation of this term is that it serves to normalize the invalid probability distribution expressed by the product in the numerator, thereby ensuring that the posterior is a valid probability distribution. This interpretation implies an alternative formulation of Bayes' theorem:

$$p(y|\theta) \propto p(y|\theta)p(\theta). \tag{2.9}$$

The angst that results from this term is due to the intractability of the integral in (2.8). The main thrust of mainstream Bayesian inference methods is to work around or approximate this integral.

LIKELIHOOD: $p(y|\theta)$ OR $\mathscr{L}(\theta|y)$ The likelihood is the term responsible for incorporating the evidence or data according to the model. It is also responsible for significant confusion: the likelihood function *is not a valid probability distribution.* It is represented as a conditional probability $p(\cdot|\cdot)$, but it is not a valid probability. Consider what constitutes a valid probability distribution: a function that assigns probabilities to values of random variables, the total probability of which is exactly 1. With likelihood functions, trouble arises from how we define the random variable. In Bayesian inference, we have observed y; the data is not random. θ is unobserved, so despite the likelihood describing the probability of y given θ, $\mathscr{L}$ is a function of θ: $\mathscr{L}(\theta)$. $p(y|\theta)$ for some fixed theta and unknown y would be a valid probability distribution. Consider the probability density function of the normal distribution:

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.10}$$

It is intentionally left undefined. The familiar normal distribution is defined in terms of the random variable x, $\mathcal{N}(x|\mu,\sigma)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.11}$$

A normal likelihood expresses our belief that the data is distributed normally given parameters $\theta = \{\mu,\sigma\}$. The data is known and fixed. The parameters, however, are unknown, resulting in a function of θ:

$$f(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.12}$$

Defining the function in terms of θ bereaves us of a normal distribution. The functional form is the same, but the variable designation is different. Recall that the total probability a random variable taking on a value is 1 (i.e. the function integrates to 1):

$$\int_{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = 1 \tag{2.13}$$

$$\int_{\theta} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \int_{\mu}\int_{\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \neq 1 \tag{2.14}$$

A likelihood function returns the *likelihood* that particular parameters θ correspond to y, a value analogous to probability but not probability.

Importantly, the way that $\mathscr{L}(\theta)$ relates data and parameters reflects the model. The normal likelihood in (2.12) assumes the data is symmetric and normally distributed around a mean μ. If, however, data is, for example, non-negative, such a model may be inappropriate. The likelihood function should express the forward model of the system.

POSTERIOR: $p(\theta|y)$ This is the quantity of interest: the distribution over parameters conditional on observed data. It provides estimates of parameter values *and* the associated uncertainty. Once a posterior has been computed, it represents the fully updated belief of parameter values. Thus, upon observing new data, it can be used as the prior distribution.

2.2.3. Hierarchical Modeling

When faced with data that has natural, grouped structure, such as cohorts of patients from different hospitals or words in sentences, a statistician ignorant of hierarchical or multilevel modeling has two options: perform independent analyses for each group (cohort, sentence) and apply post hoc averaging or similar procedure, or pool all observations (patients, words) into a single group and perform analysis for the entire population. At one extreme, each group is assumed to be independent of the others. At the other, there is assumed to be no group-specific information. Hierarchical modeling blends these approaches by treating each group as a population within a meta-population. Formally, a hierarchical setup models group-level parameters as random variables by placing hyperpriors $p(\theta|\phi)$ over the parameters and inferring the associated hyperparameters ϕ,

$$p(\phi,\theta|y) = \frac{p(y|\theta)p(\theta|\phi)p(\phi)}{p(y)} \tag{2.15}$$

A hierarchical model allows a researcher to ask questions about unobserved units within an observed group or about an unobserved group. The values of inferred hyperparameters can also indicate the similarity among groups.

Hierarchichal modeling is arguably the most important development in Bayesian statistics in recent decades, with some authors (e.g., Gelman et al. (2013), McElreath (2018)) suggesting that it be the default approach to many inference tasks.

2.2.4. Bayesian Methods for Probabilistic Inference

Here, I will briefly introduce strategies for computing posterior distributions. Many of the claims I make about each class of methods refer to the general case; for example, while sampling methods are *generally* slower and more accurate than variational methods, there are exceptions in both directions. Take each claim with a grain of salt. See Gelman et al. (2013) for more comprehensive and in-depth coverage of these topics.

MAP Though it does not represent a fully-Bayesian treatment of a problem, maximum a posteriori (MAP) estimation is nonetheless a mainstay of Bayesian methods. Related to maximum likelihood estimation, MAP seeks the maximum of the posterior distribution, following from the fact that,

$$\begin{aligned} \arg\max_{\theta} p(\theta|y) &= \arg\max_{\theta} \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \arg\max_{\theta} p(y|\theta)p(\theta), \end{aligned} \tag{2.16}$$

because the normalizing constant does not depend on θ. This has the advantage of obviating the evaluation of the integral in the numerator but reduces the posterior to a point estimate, thereby discarding many of the advantages confered by Bayesian inference.

CONJUGATE DISTRIBUTIONS: *exact, analytical* Though we often dismiss the integral in the denominator of Bayes' theorem as intractable, there are cases for which it is analytically, exactly evaluable. For trivial problems or problems with small, discrete data/parameter space (where the integral reduces to a sum), this is obviously true. However, there are particular combinations of well-known probability distributions that integrate nicely. For likelihoods belonging to standard distributional families, there usually exists what is known as a conjugate prior distribution. Evaluating Bayes' rule (including the integral) gives a posterior distribution belonging to the same family as the conjugate prior. The hyperparameters (parameters of the posterior distribution) can thus be updated according to a deterministic update rule informed by data. Conjugate priors are generally not solutions to arbitrary problems in Bayesian inference as they necessarily constrain the choice of prior. However, they are both an important fundamental, fully-Bayesian method and constitute the basis for some of the most expressive, performant Bayesian methods around (e.g., Gaussian processes).

Distributional Approximations: *approximate, analytical* Distributional and modal approximations attempt to approximate the posterior distribution or the respective factors of the posterior with tractable distributions. This often entails formulating an optimization problem. The Laplace approximation, which applies a Gaussian approximation to the posterior by matching the mean with the maximum of the unnormalized posterior and computing an approximation to the variance by use of second derivative information (Laplace's method), is a basic form of distributional approximation. The accuracy of this approximation plummets for non-normal posteriors (e.g., multi-modality, skewness, etc.) (MacKay, 2003). It can be, however, startlingly fast to compute. Variational inference (VI) is a general term that, in computational statistics, most commonly refers to approximating the posterior with variational distributions that minimize the KL divergence (Kullback and Leibler, 1951). The approximating, variational distributions are specified to come from a convenient family of distributions (usually exponential) and to follow a particular factorization (often fully-factorized). Even under these strong assumptions, the KL divergence is usually intractable. A tractable lower bound,

$$\mathrm{E}_q\left[\log p(\theta, y)\right] - \mathrm{E}_q\left[\log q(\theta)\right], \tag{2.17}$$

where $q(\cdot)$ is the approximating distribution, is instead maximized, leading to equivalent results. This lower bound objective can be seen to contain the joint distribution of parameters and data, which includes the likelihood. For appropriate models, the convenient choice of distributional family and factorization yields analytic updates that can be performed iteratively, updating one variational parameter at a time. This leads to the variational Bayes or coordinate ascent VI algorithm. Improvements in the optimization strategy lead to improvements in the inference. Stochastic VI (Hoffman et al., 2013) replaces coordinate ascent with stochastic optimization, which produces faster, better inference results.

The keen reader will notice that VI seems to have more to do with deterministic inference than probabilistic inference. Importantly, VI is a strategy for performing probabilistic inference over parameters. However, VI uses deterministic inference in the distribution space to do so. VI abstracts parameter inference to inference over functions and computes a point estimate of the approximating function.

SAMPLING METHODS: *exact*, numerical* The strategy of sampling-based methods is to generate samples from the target distribution i.e., $p(\theta|y)$. The breadth of these methods is staggering, but they adhere to principles of Monte Carlo sampling (transforming random numbers into quantities of interest). Monte Carlo methods that sample from Markov Chains, known as Markov Chain Monte Carlo (MCMC) simulation, also inherit convergence properties from Markov theory. Upon convergence, sampling methods produce samples from the posterior distribution. The outcomes of these methods are sets of uncorrelated or correlated samples from the posterior distribution. Sample statistics, with sufficiently many samples and a reasonable sampling scheme, correspond to statistics of the target distribution. Sample statistics of these samples are exact in the limit of infinite samples (*) and thus represent some of the most accurate Bayesian methods. However, they also rank among the slowest methods. As with all other classes of methods discusses here, they require access to the unnormalized posterior.

In summary, Bayesian inference is a principled framework for solving inverse-problems. However, it is a non-trivial task because computing the posterior analytically or exactly is often impossible. Importantly, the classical methods that exist for posterior computation are all reliant on access to the unnormalized posterior density.

2.3. Likelihood-Free Inference

Performing standard probabilistic inference by use of Bayes' rule requires access to the likelihood of the parameters in addition to the forward model. Often times, the function describing the likelihood of parameters naturally expresses the forward model. When researchers anticipate using Bayesian inference to formulate their inverse problem, it is common to simply define the forward model probabilistically from the outset. Other models, which define a map from parameter and input space to data space, can be straightforwardly written as a likelihood. For example, assuming a linear model with Gaussian noise, the likelihood is given by a Gaussian distribution with mean $\mathbf{w}^\intercal\mathbf{x}$ and variance σ^2, where $\mathbf{w}$ is the vector of weights. With appropriate priors on $\mathbf{w}$ and σ^2, we obtain an expression that conveys both $\mathscr{L}(\mathbf{w},\sigma^2 \mid y)$ and the assumed, underlying generative process. A large class of models, however, evades such treatment, *likelihood-free models*. Membership in this class is somewhat contrived in that any model for which the likelihood is intractable or unavailable is a likelihood-free model. Theoretically, computing a likelihood entails considering every possible observation for a fixed set of parameters. For deterministic models, this is trivial because there exists only one such mapping. For stochastic models, however, this requires integrating over the stochastic, latent variables z present in the generative process, as shown in the definition of the likelihood,

$$p(y|\theta) = \int p(y,z|\theta)dz. \tag{2.18}$$

Classical probabilistic models define $p(y|\theta)$ directly, thus obviating the integral in (2.18). For likelihood-free models, the integral in (2.18) is intractable (in addition to the evidence termn, (2.8)). In the absence of an evaluable likelihood, standard Bayesian inference methods fall flat.

In practice, likelihood-free models take the form of mechanistic, rule-based models such as simulations. Though these models can be highly expressive, they often possess only implicit likelihoods; samples can be generated from the model by running the simulator for a set of parameters and inputs according to well-understood, natural mechanisms, but the likelihood of any outcome relative to others is unknown.

Likelihood-free inference (LFI) is a collection of methods dedicated to enabling probabilistic inference on simulator-based models. Why is this task important? Often times, describing how data arises or how systems evolve is much more natural and justifiable than constructing a probabilistic model from the outset. This is the motivation for modeling strategies like agent-based modeling. Much of the original development of LFI methods was motivated by the size and complexity of data in population genetics. LFI is thus closely related to advancing science: it enables statistical inference on descriptive, mechanistic physical models.

Generally, LFI methods approximate the posterior distribution over parameters by comparing forward simulations to observed data. Consider the Rejection-ABC scheme (Pritchard et al., 1999), which generates N samples from the approximate posterior distribution:

```
1. For $i$ in $1 : N$
2.     Do:
3.         $\theta_* \sim p(\theta)$
4.         $y_{\theta_*} \sim \mathscr{M}(\theta_*)$
5.     Until: $\Delta(y_{obs}, y_{\theta_*}) \leq \epsilon$
6.     $\theta_i = \theta_*$
7.     $\Theta = \{\Theta, \theta_i\}$
```

In prose, sample parameters from the prior; simulate data for the sampled parameters; if the simulated data is epsilon-close to the observed data, add the data-generating parameters to the set of posterior samples. This procedure requires the user to specify a model $\mathscr{M}$, a prior distribution $p(\theta)$, a discrepancy function Δ, and threshold ϵ. Because rejection-ABC uses rejection sampling, the number of accepted samples N must also be specified. With the exception of N, these features are common to all LFI methods.

Pritchard et al.'s first innovation was sampling from the prior. Sampling from the prior rather than an arbitrary space (without appropriate re-weighting) yields samples from $p(\theta|y_{obs})$. However, this is only theoretically possible for discrete y because $p(y_\theta = y_{obs} \mid \theta) = 0$ for continuous distributions. Pritchard et al.'s second innovation, the introduction of an acceptance tolerance ϵ, enabled inference on continuous models. The drawback, however, is that generated samples are from the approximate posterior $p_\epsilon(\theta|y) = p(\theta|\Delta(y_{obs}, y) \leq \epsilon)$. This is the source of the classical nomenclature, approximate Bayesian computation (ABC). The approximate posterior converges to the true posterior as ϵ goes to zero. Therefore, ϵ is a trade-off parameter; the smaller the tolerance, the fewer samples are accepted and the more simulations need to be run to reach N posterior samples. Small values of ϵ quickly become computationally intractable. Furthermore, recall that sampling methods with access to the likelihood only converge in the limit of infinite samples. Therefore, ϵ adds a layer of approximation: $\lim_{\epsilon \to 0} p_\epsilon^n(\theta \mid y) = p^n(\theta \mid y)$, and $\lim_{n \to \infty} p^n(\theta \mid y) = p(\theta \mid y)$.

For high-dimensional y, computing the discrepancy Δ becomes increasingly difficult. Beaumont et al. (2002) propose the use of summary statistics $S(y)$ in place of uncompressed data, yielding a posterior $p_\epsilon(\theta|S)$. The rejection-ABC algorithm can be modified to accommodate this change by replacing 5. with $\Delta(S(y_{obs}), S(y_{\theta_*})) \leq \epsilon$. A posterior of this form is equivalent to $p_\epsilon(\theta|y)$ if and only if $S(\cdot)$ is statistically sufficient (e.g., mean and variance of Gaussian data). The use of summary statistics, however, enables the comparison of rich data formats, including high-dimensional and time series data, which would otherwise be unwieldy. Summarization is generally assumed to outweigh the bias induced by use of a reduced representation. Choice of S is non-trivial; designing and choosing S is usually left to domain-experts.

A suite of natural extensions to rejection-ABC emanating from the Monte Carlo and MCMC literature have been applied to LFI problems. The Metropolis-Hastings algorithm can be applied to ABC by slightly modifying the acceptance probability of a transition from x to candidate x^*:

$$A(x^*|x) = \mathbb{I}\left[\Delta(y, y_0) \leq \epsilon\right] \mathbf{min}\left(1, \frac{p(\theta)}{p(\theta^*)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right) \tag{2.19}$$

where $x = (\theta, y)$, and q is a proposal distribution that depends only on θ. Replacing the rejection step in the above algorithm with this Metropolis-Hastings step yields the MCMC-ABC algorithm described by Marjoram et al. (2003). It has the advantage of preventing the need to naïvely sample from the prior.

To further address the problem of tolerance setting, a number of papers propose iterative algorithms that reduce the tolerance with each iteration (e.g. Del Moral et al. (2012), Sisson et al. (2007), Beaumont et al. (2009)). Collectively, these are known as sequential Monte Carlo-ABC (SMC-ABC).

In contrast to sampling approaches to LFI, Wood (2010), attempts to approximate the unavailable, implicit likelihood. Wood finds the Gaussian approximation to the distribution of *summary statistics* by use of parameter-simulated data pairs. The choice to use a Gaussian approximation is justified by the central limit theorem but is only principled as $n \to \infty$. In possession of a Gaussian approximation to the likelihood, you can proceed with standard Bayesian inference. However, when n is finite, the Gaussianity assumption may become restrictive. This strategy, however, has important implications and characterizes a growing class of methods for LFI known as synthetic likelihood.

Synthetic likelihood can be generalized by replacing the Gaussian approximation with more flexible approximations to the likelihood. This generalization connects LFI with density estimation methods. For a set of samples $X = \{x_1 \dots x_N\}$ drawn from an arbitrary, unknown probability density p, the task of density estimation is to identify an estimate $\hat{p}$, usually belonging to a predetermined family of distributions, of the true density. Density estimation is the foundation of numerous statistical and machine learning methods because it bridges the gap between discrete data and continuous, evaluable distributions that can be used in arbitrary analyses. Evaluable estimators can be sampled from via MCMC methods, but it is often more convenient for density estimators to be directly sampleable (i.e., contain or define an accessible transformation of a uniform random number). Learning a flexible approximation to the distribution of summary statistics from simulated data samples and using the resulting synthetic likelihood for inference is a LFI strategy some (e.g., Alsing et al. (2019)) have suggested naming density estimation for likelihood-free inference (DELFI). It is an increasingly popular strategy (see e.g., Alsing et al. (2019), Lueckmann et al. (2017), and Papamakarios (2019)). In particular, neural density estimation provides a theoretically infinitely flexible estimator.

A neural density estimation method that is central to this work is real-valued non-volume preserving transformations (real NVP) (Dinh et al., 2017).

2.3.1. Real NVP

Recent developments in density estimation and probabilistic generative modeling (e.g., Goodfellow et al. (2014), Rezende et al. (2014)) employ approximate inference or auxiliary models, such as discriminator networks, to enforce learning conditions and arrive at a learned density. Real NVP is a relatively simple approach to density estimation, requiring neither approximate inference nor auxiliary models. Many sophisticated approaches to generative modeling, including real NVP, involve learning a latent representation of the data space, a latent space. Some feature of the latent space (e.g. dimension, functional form) is pre-specified and the model learns a transformation between the data space and the latent space. Dinh et al. propose specifying a unit Gaussian latent space and defining the transformations as invertible bijections, thereby allowing for the use of change of variables formula. The transformations are further defined in terms of a scaling function s and translation t. For example, a single transformation is given by

$$z = \exp(s(x)) + t(x). \tag{2.20}$$

Transformations of this form can be layered for added flexibility. In layering transformations, masks are applied such that only a subset of dimensions are transformed by each layer, thereby enforcing a strong non-linearity and preventing layers of linear tranformations from reducing to a single linear transformation. In layered, masked form, the transformations are given by

$$\begin{aligned} y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= x_{d+1:D} \circ \exp(s(x_{1:d})) + t(x_{1:d}), \end{aligned} \tag{2.21}$$

where ∘ is an element-wise product and y is intermediate output (between data x and latent variables z). Inversion of this transformation can be performed by use of the change of variable formula, which requires evaluating the Jacobian of each transformation. The clever construction of this transformation, however, yields a Jacobian that does not involve the Jacobian of the functions s or t themselves. Therefore, s and t can be arbitrary functions. Dinh et al. choose to use deep neural networks; that is also what I choose to use in this project. This collection of transformations can be trained using simulated data to recover the latent space using maximum likelihood learning. The result is a learned transformation from the N-dimensional data space to an N-dimensional unit Gaussian, which is both evaluable and sampleable. The likelihood of a datum in the original space can be evaluated by transforming the datum x^* into its unique counterpart in the latent space z^* and evaluating the Gaussian at z^*. Samples can be drawn by sampling a z^* from the latent Gaussian space and inverse-transforming the sample.

2.3.2. Bayesian Optimization for Likelihood-Free Inference

Bayesian optimization for likelihood-free inference (BOLFI) is another LFI method central to the workflow presented here. I will first introduce Bayesian optimization and the active learning paradigm then explain its use in LFI. Under favorable conditions, BOLFI is the most sample-efficient LFI method in widespread use.

BAYESIAN OPTIMIZATION Bayesian optimization can be succinctly described as probabilistic, derivative-free, global optimization of black-box functions. Black-box functions (referred to as the "objective" herein) are not available in closed-form and can only be queried/evaluated. Therefore, we have access to neither closed-from derivative information nor resort to automatic differentiation, thus precluding methods like gradient ascent, hence, "derivative-free." Simulators are often black-box functions. The goal of Bayesian optimization is to find a global maximum or minimum of the black-box objective function. Bayesian optimization first fits a surrogate model to data queried from the objective and optimizes the surrogate model. A common choice of surrogate model is a Gaussian process (GP).

A GP is a regression model that computes a fully-Bayesian posterior over *functions*. Accordingly, a prior distribution $p(f)$ over the space of possible functions f is combined with queried data of the form $\mathscr{D} = \{x_{1:N}, f^*(x_{1:N})\}$ via the likelihood $p(\mathscr{D}|f)$. This yields a posterior distribution over functions: $p(f|\mathscr{D})$. A GP is commonly expressed as $GP(m,k)$, where m and k are user-specified mean and covariance functions. The posterior defines a Gaussian distribution at every element of the regression domain. A noiseless GP is exact for $x \in \mathscr{D}$ and estimates and quantifies uncertainty for $x \notin \mathscr{D}$. For this reason, GPs are regarded as infinite-dimensional Gaussian distributions. Gaussian distributions are well behaved. One consequence of this is that fitting the model to data is performed exactly and analytically by a conjugate update.

Having selected and introduced the surrogate regression model to be used in the Bayesian optimization, we now consider strategies for fitting the model to data. Naïvely sampling $x_{1:N}$ from the domain, evaluating f^* at $x_{1:N}$, fitting the GP surrogate to samples, and computing the minimum (or maximum) of the surrogate (which is possible due to the properties GPs) would seem to solve the problem. The nuance and strength of Bayesian optimization, however, is its sample efficiency. Rather than sampling the domain uniformly, Bayesian optimization uses active learning. Active learning is a general machine learning paradigm wherein training data is requested by the model, the primary objective being to gather better, fewer data. Therefore, it is particularly suited to settings in which data is difficult or expensive to gather. Bayesian optimization uses active learning through acquisition functions. Acquisition functions take as input the current state of the model and output an optimal domain value, which maximizes the utility function defined in the acquisition function. The objective is then queried at the given domain value and the model updated accordingly.

To recap, Bayesian optimization seeks to maximize or minimize a black-box objective function. To do so, a well-behaved surrogate model is fit to the objective function. In the fitting process, acquisition functions determine where to evaluate the objective. Acquisition functions typically define and maximize a utility function; the maximum of the utility function gives the "optimal" evaluation location.

Acquisition functions generally balance two forms of search: seeking maximum uncertainty (exploration) and seeking maximum information (exploitation). For example, maximum variance or MaxVar is a pure exploration strategy. Let $V(x \mid \mathscr{D})$ be the point-wise or predictive variance under the surrogate GP, as defined, for example, in equation 10 of Järvenpää et al. (2019). Here, V is the utility function. Upon observing N data and fitting the GP to them, MaxVar can be used to select observation $N+1$. The optimal acquisition is given by

$$x^* = \arg\max_x V(x \mid \mathscr{D}_{1:N}). \tag{2.22}$$

The black-box function is queried at x^*, yielding y^*, and the dataset is extended such that $\mathscr{D}_{N+1} = (x^*, y^*)$. This acquisition function is rather rudimentary. Thanks to the properties of Gaussian distributions and GPs, acquisition functions can consider probabilistic utility. A classic example is the probability of improvement (PI) function, which quantifies the probability that a particular acquisition will give a more optimal (with respect to the primary objective) value than the current optimum. With surrogate $\mathscr{G}$ and current optimum located at x^*, PI is given by $p(\mathscr{G}(x^* \leq \mathscr{G}(x))$, which can be computed using the Gaussian cumulative distribution function Φ as

$$\mathrm{PI} = \Phi\left(\frac{\mathscr{G}(x^*) - \mu(x)}{\sigma(x)}\right), \tag{2.23}$$

where $\mu(x)$ and $\sigma(x)$ are the mean and variance of $\mathscr{G}$ at x. This formulation of PI will seek a global minimum. Reversing the position of μ and $\mathscr{G}$ will give a maximization-oriented utility function. For more detail on PI and variants, see Brochu et al. (2010). Maximizing PI with respect to x gives the MPI acquisition function.

The last acquisition function I'll mention here is the lower confidence bound (LCB) with maximization counterpart upper confidence bound (UCB). The LCB is given by

$$LCB = \mu(x) - \nu\sigma(x), \tag{2.24}$$

where ν is a positive, tunable parameter and $\mu(x)$ and $\sigma(x)$ are defined as above. This can serve as a good, general-purpose acquisition function and is used in this project.

To summarize, Bayesian optimization uses surrogate models to optimize black-box objective functions. Acquisition functions are used to identify domain values at which to evaluate the black-box functions. Once fit, the optimum of the surrogate approximates the optimum of the objective. This approach is derivative-free, for use of surrogate models, and data-efficient, for use of acquisition functions. Furthermore, it is fully probabilistic due to the choice of GPs as surrogates.

BOLFI Features of Bayesian optimization seem eminently applicable to LFI: sample-efficient probabilistic modeling of black-box functions. Gutmann et al. (2016) propose modeling and minimizing the LFI discrepancy function Δ by use of Bayesian optimization. Specifically, the black-box function considered combines the simulator and discrepancy function, so the intractable objective is a function of parameters; the response variable is discrepancy. Thus, data used in modeling the objective is of the form $\mathscr{D} = \{\theta_{1:N}, \Delta(y_{obs}, \mathscr{M}(\theta_i))_{1:N}\}$. BOLFI begins by uniformly sampling the domain and running the simulator to obtain initial data with which to fit the GP. After sufficient, random data have been queried, acquisition functions take control and dictate the parameters at which to simulate data. The minimum of the resulting GP, $\tilde{\Delta}^*$, fit to parameter-discrepancy tuples, approximates the minimum of the discrepancy function Δ^*. An approximate posterior distribution over parameters can be recovered by truncating the GP at $\tilde{\Delta}^* + \epsilon$ and sampling from the resulting truncated Gaussian distributions.

BOLFI is an extremely sample-efficient method and can produce high-fidelity posteriors with $\mathcal{O}(N^2)$ simulations. BOLFI also has a number of drawbacks, however. One key drawback is that it can tolerate only limited data. With N data, updating a GP necessitates inverting the $N \times N$ covariance matrix, an operation in $\mathcal{O}(N^3)$. This restricts the number of observations in $\mathscr{D}$ to the order of thousands. Additionally, BOLFI is excessively tunable. GPs entail a number of modeling choices, including mean and covariance functions, acquisition functions often have user-tunable parameters, and BOLFI itself adds tunability. This is all in addition to the tuning implied by LFI. All told, this amounts to more tuning parameters than a practitioner can be reasonably expected to attend. Nonetheless, BOLFI and variants exist as the only tenable LFI methods for high-cost simulators.

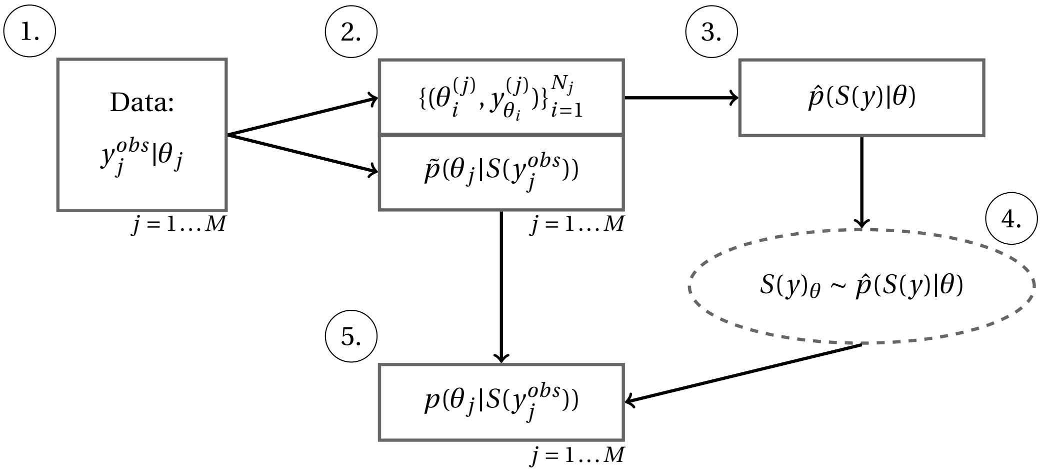
2.3.3. Engine for Likelihood-Free Inference

Engine for likelihood-free inference (ELFI) (Lintusaari et al., 2018) is a probabilistic program and Python package for LFI. It includes, among other things, implementations of rejection-ABC, SMC-ABC, BOLFI, and a number of simulators. This package serves as the basis for much of the LFI computation presented in this project. ELFI is by no means the only LFI package. However, it is an appropriate choice because it 1) is under active development, 2) implements the necessary LFI methods, and 3) has featured in high-impact journal articles (e.g. in Nature Microbiology, Shen et al. (2019)).

3. Methods: Workflow

The main contribution of this project is the introduction and validation of a general workflow for multi-observation likelihood-free inference on high-cost simulators by use of fast, global surrogate likelihood approximations (the surrogate likelihood described below is distinct from the surrogate regression model in Bayesian optimization). The workflow, in its entirety is presented first, followed by motivation and explanation of each component of the workflow. While being a general workflow, I attempt to offer grounding by way of implementation details for each component. I conclude by synthesizing the workflow and considering limitations, alternative interpretations, and possible extensions.

Full workflow:


1.
Data:
$y_j^{obs}|\theta_j$
$j = 1\dots M$
2.
$\{(\theta_i^{(j)}, y_{\theta_i}^{(j)})\}_{i=1}^{N_j}$
$\tilde{p}(\theta_j|S(y_j^{obs}))$
$j = 1\dots M$
3.
$\hat{p}(S(y)|\theta)$
4.
$S(y)_\theta \sim \hat{p}(S(y)|\theta)$
5.
$p(\theta_j|S(y_j^{obs}))$
$j = 1\dots M$

For observations $j = 1\dots M$ corresponding to unique sets of parameters of interest, the workflow proceeds as follows:

1. Observe $j = 1\dots M$ data corresponding to unique sets of parameters of interest:

$$y_j^{obs}|\theta_j$$

2. Compute crude, initial posterior updates for each observation:

$$\{\tilde{p}(\theta_j|S(y_j^{obs}))\mid j = 1\dots M\},$$

and pool the resulting M sets of N evidence points:

$$\Rightarrow \{(\theta_k, y_{\theta_k}) \mid k = 1 \ldots M \times N\}$$

3. Learn approximation to global likelihood (surrogate) of summary statistics:

$$\hat{p}(S(y)|\theta)$$

4. Replace expensive simulator with surrogate likelihood simulator:

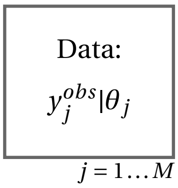
$$S(y)_\theta \sim \hat{p}(S(y)|\theta)$$

5. Use surrogate likelihood and initial posteriors together to complete local, posterior inference:

$$\{p(\theta_j|S(y_j^{obs})) \mid j = 1 \ldots M\}$$

3.1. Motivation

3.1.1. Multi-observation LFI



In the standard LFI setting, the target is $p(\theta|y_{obs})$ or $p(\theta|S(y_{obs}))$, where y_{obs} is a single realization of a data generating process. A complete "fit" of an LFI method, be it via MCMC, density estimation, or Bayesian optimization, returns a posterior belief about the values of parameters that probably generated the data, according to the model. Consider, now, collecting multiple observations. Each observation can be investigated using the same model, but each has a unique set of parameters of interest. In order to perform LFI in this setting, standard methods must be used sequentially or in parallel; a single fit returns a belief about a unique parameter vector. If this is impractical (e.g., too computationally costly), a natural solution would be to pool simulations across local inferences–use simulations from task A directly in task B.

This approach would immediately render parallelization non-trivial or impossible. More importantly, however, LFI methods with any level of sophistication rely heavily on locally-optimal acquisitions, simulations which offer the most information about a specific parameter vector. SMC-ABC, for example, focuses simulation resources and narrows the acceptance criterion in tandem, eventually producing only simulations in regions of high posterior probability. In high-dimensional space, most posterior mass exists in a minuscule region, so optimal acquisitions for one observation are likely to be rejected (i.e., $\Delta > \epsilon$) by other local inferences. Therefore, simply sharing simulations across tasks is not necessarily useful. Despite these challenges, information about the modeling task is contained in non-locally optimal simulations. This workflow endeavours to extract that information and share it among the local posteriors by learning a representation of the global likelihood. A global likelihood can be used to complete costly local inferences and quickly compute posteriors for future observations.

Thus, the first step in the workflow is observation of M data. One datum may be as complicated as a snippet of time-series recordings, samples from a process, or the state of a system after a fixed amount of time. In any case each datum corresponds to a single parameter setting of a model that describes all M observations–different parameters, common model.

3.1.2. Expensive Simulators

If the simulator in question is fast (order of seconds), this workflow is not relevant. Parallel or even sequential implementations are sufficient for high-quality inference in reasonable time. However, when a single simulation takes minutes or hours to run, and there are multiple, relevant posteriors of interest, LFI quickly becomes infeasible. Sophisticated MCMC-ABC methods may require hundreds of thousands of simulations per observation. This workflow is most relevant for tasks where the simulation budget is less than 10,000 simulations. BOLFI (and variants) is the only method I am aware of that is capable of computing a posterior with $< O(10^3)$ simulations. For complicated, realistic simulators, computing even a single posterior may take several weeks in a high-performance computing environment.

At its core, enabling inference on high-cost, complex simulators addresses the same problem as model selection. Instead of comparing the evidences of multiple, tractable models, LFI generalizes the definition of a model to accommodate mechanistic dynamics and convenient implementation strategies. The “best” model for describing a physical system can now come from a broader class of models without sacrificing any inverse problem solving abilities.

Though the strategy has been hinted at, for example by Alsing et al. (2019),

> However, in some scenarios we may run many “experiments” that generate independent realizations of data **d** from the same data generating process, and we want to analyze those data as they are taken. In these situations, it is desirable to abandon active learning and build a global emulator for $p(\mathbf{d}|\theta)$ over the full prior volume, that can then be used to analyze any subsequent data **d** as they are observed.

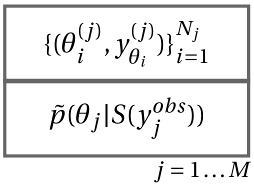
there is not been, to my knowledge, an end-to-end investigation. In their recent review article, Cranmer et al. (2019) detail three shortcomings of modern simulator-based inference: sample efficiency, quality of inference, and amortization (reducing repetition of computational steps). This workflow directly addresses amortization and sample efficiency. The third shortcoming, quality of inference, while of critical importance, is not the main objective of the workflow. The workflow aims to *enable* inference where it was previously impossible, thus *improving* inference is impossible. Application of this workflow to problems that can be solved with existing methods may result in a trade-off between accuracy and efficiency.

3.1.3. Implementation Details

In my experiments, data is obtained by first sampling M parameter vectors from the prior, then simulating data with those parameters.

A problem common to the validation of many LFI methods is how the “true” parameters are chosen. Because most methods are designed to infer parameters for a single observation, it is often the case that a specific parameter vector is chosen to be the test case. Commonly, this test case is consistent throughout the literature. This permits comparison across methods. However, the test parameters always live in a well-identified region of the space. They are often some of the easiest parameters in the prior to identify. As such, as soon as the test cases are generalized to multi-observation settings, where true parameters are sampled from the prior, the problem suddenly become much harder, both because there are multiple observations and because we must venture away from well-behaved regions of parameter space.

3.2. Initial Local Posteriors


$\{(\theta_i^{(j)}, y_{\theta_i}^{(j)})\}_{i=1}^{N_j}$
$\tilde{p}(\theta_j | S(y_j^{obs}))$
$j = 1 \dots M$

Upon observing multiple data, it is necessary to query the simulator to gather information (generate simulations) about the likelihood space, even in a sample-efficient scheme. In an idea world, we would construct a global acquisition rule that considers all observations, the states of all local inferences, and makes queries that are somehow optimal for all. However, such an acquisition rule does not yet exist and may be intractable for some problems given the potential for zero overlap among local feasible sets. Instead, I grant control of choosing simulation locations to the M local models. Each local inference selects N locally-optimal evidence points, where $N \times M$ is less than or equal to our simulation budget. For problems that merit this approach, the local posteriors that result will be crude, unconverged approximations to the posteriors. $N \times M$ evidence points may be sufficient to fully solve a single or even multiple posteriors, but are insufficient to solve all M posteriors. However, by performing initial LFI on *all* local posteriors, the evidence points should cover the likelihood space better than if all evidence were actively aquired by a single posterior (i.e. M weak local optima vs 1 strong local optimum), and we will have generated $N \times M$ evidence points with which to amortize inferences. We can likely also expect any pooling or compilation object to be more accurate near the M posteriors due to the concentration of training data chosen by the initial fits.

3.2.1. Implementation Details

As noted above, BOLFI is the LFI method of choice for the analyses performed here because is a natural choice in high-cost simulation settings. That being said, the workflow is not restricted to BOFLI-based inference. In practice, I compute M BOLFI posteriors with N locally-acquired evidence points per posterior. To do this, I pass M model objects to respective BOLFI inference objects using the ELFI package. I then fit the inference objects with N evidence and save the state of each. Additionally, I save all evidence points for later pooling.

3.3. Global Likelihood Surrogate

$$\boxed{\hat{p}(S(y)|\theta)}$$

In possession of $N \times M$ evidence points from across the parameter space but which slightly prefer the respective local optima, we can approximate the conditional distribution that describes the data: the global likelihood $p(y|\theta)$. Because LFI posteriors are fit in terms of summary statistics and because raw data is often unwieldy, we change the target distribution slightly to $p(S(y)|\theta)$, the likelihood of summary statistics. The switch to summary statistics is consistent with most practical LFI. Further, if the data is sufficiently compact, $p(y|\theta)$ can be approximated directly without deviating from the proposed workflow. The global likelihood differs from synthetic likelihoods appearing in other works in that it should return the likelihood for arbitrary data-parameter pairs; it is a free function of both y (or S) and θ. As such, it is a more difficult quantity to learn than a likelihood for fixed data. In the language of density estimation, this quantity corresponds to a conditional (on θ) density estimator. An appropriate estimator can be evaluated for a pair (θ, S) and yield samples $S_\theta \sim \hat{p}(S|\theta)$. In this way, we essentially have a surrogate simulator and evaluable likelihood function, valid for all data.

3.3.1. Surrogate Substrate

How to model the surrogate simulator or global likelihood ranked among the more important questions that arose in developing this workflow. Consider the role that the surrogate plays. The surrogate pools information from multiple, local posterior inferences, each with their own locally optimal acquisitions (i.e. simulator queries). The pooled acquisitions are used to learn a global representation of the likelihood, which can itself be sampled in place of the expensive simulator. These roles imply the necessary traits of a satisfactory surrogate substrate:

1. Generative (sampleable)
2. Sufficiently expressive
3. Inexpensive to train

A surrogate must be samplable in order to stand-in for the simulator. Because likelihood-free models are typically black-boxes, we are unable to constrain the complexity of the implicit likelihood and thus use a surrogate that is able to express highly abnormal distributions. Lastly, because we are operating in a high-cost computing environment (i.e., expensive simulator), learning the surrogate must be data-efficient.

3.3.2. Implementation Details

In validating this workflow, I experimented with a number of surrogate models. The original motivation was to build on recent developments in neural density estimation for likelihood free inference. The pydelfi (density estimation likelihood-free inference) package (Alsing et al., 2019) provides implementations of a number of NDEs (e.g. mixture density network (Bishop, 1994), masked autoencoder for density estimation (Germain et al., 2015), etc.) but results with these implementations were unsatisfactory. I simplified the surrogate substrate and instead assumed independence of the summary statistics and modeled each with a GP. This was simply to validate the method. I found this to perform well on a simple task, indicating the ability of the workflow. As the complexity and dimension of the simulator grew, however, this approach became worse, as was expected. Having loosely validated the method, I searched for a more flexible estimator that could more straightforwardly handle multidimensional output. This was satisfied by real NVP.

In order to use real NVP for this workflow, I adapted a publicly-available Python implementation (Arsenii, 2018). The significant changes made to the code were changing from marginal density estimation to conditional density estimation, enabling batch-wise sampling and simulation, and constructing a more efficient training strategy.

Due to the computational and statistical intractability of implicit likelihoods, it is impossible to comprehensively assess the estimator's goodness of fit to the likelihood. Therefore, great care must be taken to robustly train the estimator. The real NVP surrogate is trained with maximum likelihood learning. The loss function used is the negative log probability of data under the latent distribution. That is, the log probability is evaluated by first transforming data to the latent space and then evaluating the latent probability of the transformed data. This loss functions encourages the surrogate to assign greater probability to data-dense regions. I use batch-wise, stochastic gradient descent with a decaying learning rate and track the loss on a held-out validation set. The learning rate is reduced periodically by a user-specified multiplicative constant. Training terminates when the maximum number of epochs has been reached or the validation loss fails to improve for a user-defined number of consecutive epochs, at which point, the parameter values revert to the minimal validation loss setting. The period of learning rate decay should be shorter than the early stopping criterion so that the learning rate will always shrink at least once before early termination. Included in the attached Python file (`cnvp.py`) is also a simpler training scheme which optimizes the network parameters for a fixed number of epochs.

3.4. Surrogate Simulator

$$S(y)_\theta \sim \hat{p}(S(y)|\theta)$$

Stochastically generate synthetic data conditional on arbitrary fixed parameters. Given an appropriate choice of surrogate substrate, this task is equally applicable to the surrogate as to the original simulator. A learned, global likelihood that can be straightforwardly used as a simulation. Conditioning a perfect density estimator on arbitrary parameter values and sampling from the resultant conditional distribution will yield values indistinguishable from summarized simulations. We are forced to settle for an imperfect estimator. However, by carefully choosing the density estimator to satisfy the aforementioned necessary traits, sampling is trivial in any case and the samples are good approximations to summarized simulations. It is an exchange of accuracy for efficiency. For computationally intractable problems, however, it exchanges unattainable accuracy for usable results.

3.4.1. Implementation Details

Having adapted real NVP to be used for conditional density estimation, the latent space is conditionally s-dimensional Gaussian where s is the dimension of the summary statistics. To obtain samples, I condition on the chosen parameters, sample from the resulting s-dimensional Gaussian distribution, which is a fast and straightforward process, and map the Gaussian sample to the data space according to the learned transformation.

3.5. Posterior Inference

$$\boxed{p(\theta_j|S(y_j^{obs}))}$$

$$j = 1 \ldots M$$

All that remains is to link the incomplete inference objects with the surrogate simulator. Because sampling from the surrogate is nearly instantaneous, the surrogate simulation budget is larger than necessary to complete arbitrarily many inferences.

An important benefit of amortized inference, beyond solving pressing problems, is the object that is created as a result. In our case, this object is a surrogate likelihood. For future observations $M+1\ldots$, the surrogate can be used either in conjunction with the simulator to speed up inference or in isolation to provide a very fast approximation to the posterior. Further, any future simulations can be used to enhance the surrogate and past posteriors quickly recomputed.

3.5.1. Implementation Details

To obtain results, I decouple the simulator from the inference method and supply the surrogate sampling scheme in its stead. When a BOLFI posterior determines the location (parameter vector) of the next, optimal acquisition, the location is supplied to the density estimator, which is sampled conditional on that parameter vector. The sample is treated as a simulation, but because the surrogate is an estimate of $p(S(y)|\theta)$, no summarization is necessary. The sampled surrogate summary statistics are then used directly to update the state until we have achieved satisfactory convergence. Convergence of the BOLFI inference object concludes the workflow. However, to obtain samples and sample statistics from the posterior distributions, the GP underlying the BOLFI inference must be truncated at an appropriate threshold and sampled. Samples from the truncated GP model are samples from the approximate posterior.

To perform inference on new observations, I again sample true parameters from the prior and simulate true data at the parameters. I then build model and inference objects only around the surrogate simulator. In this way, we obtain fast approximations to the posterior with exactly zero expensive simulations.

3.6. Methods Discussion

The keen reader may notice strong similarities between this workflow and hierarchical Bayesian modeling. Though, in its present form, the formulation is not formally hierarchical (I do not infer a distribution over population-level parameters), the relationship is worth noting. Hierarchical LFI is not new. Turner and Van Zandt (2014) develop the Gibbs-ABC algorithm, which leverages the fact that the conditional posterior distribution over hyperparameters $p(\phi|y,\theta)$ does not depend on the likelihood, so ϕ can be sampled directly from $p(\phi)\prod_{i=1}^{N} p(\theta_i|\phi)$. While this approach performs well and infers population-level parameters, the number of simulations necessary is on the order of 10^5. In a different vein, Tran et al. (2017) build on developments in implicit variational inference (Titsias and Ruiz, 2018) to introduce hierarchical likelihood-free variational inference. Their approach weds variational inference with density ratio estimation, but shows some signs of instability. A natural, future extension to this work will be to formulate the task hierarchically, which may affect the choice of density estimation tool. One possible direction is to more formally consider the latent space of the density estimation; global latent variables that are shared across observations may possibly be interpreted as hyperparameters. Though real NVP has a latent space, they are conditional on parameters. The converse is true of hyperparameters: latent hyperparameters ϕ should satisfy $p(\theta|\phi) \neq p(\theta)$ (i.e., statistical dependence). Furthermore, hyperparameters should only affect the likelihood only via the parameters, $p(y|\theta) = p(y|\theta,\phi)$; the likelihood is independent of hyperparameters, conditional on the parameters.

Relatedly, this workflow begs the question, why not use the likelihood approximation as you would a standard likelihood function? Indeed, an accurate learned likelihood could obviate LFI methods. Once multiplied by the prior, standard MCMC methods could be used to sample from the unnormalized posterior density. This is another interesting future direction for this work. However, there may be problems with such a wholesale deviation from LFI. Theis et al. (2015) note that generative models can achieve high log-likelihood while producing samples that look nothing like training examples and can similarly produce high-fidelity samples without achieving good log-likelihood. In other words, dependence between the visual appearance of samples and log-likelihood is limited. LFI methods rely on the data-generating aspect of generative models whereas classical Bayesian methods rely on the likelihood evaluation aspect. Transitioning to, for example, MCMC would necessitate a careful analysis of the density estimation tool to ensure that it is achieving satisfactory performance on the correct metric.

4. Experiments

I test the ability of global surrogate likelihoods to accelerate multi-observation inference. When successful, this method will *recover reference solutions* with 5-10x fewer simulations, thus enabling inference in high-cost domains where reference solutions are unavailable. Because the surrogates are intended to be low-cost, high-fidelity replacements for simulators, non-identifiability or other flaws in reference solutions are expected and preferred to be reflected in surrogate solutions. Specifically, posteriors produced by the workflow should mirror the reference posteriors or exhibit bias in the direction of the true parameter value. Additional detail for all experiments can be found in the accompanying notebooks. Unabridged results are reported in the appendix.

The experiments are intended to test three characteristics of the workflow. The first is a proof of concept; can a global surrogate be used to accelerate multi-observation LFI? The second is an introductory analysis of scaling behavior; how does the workflow perform when the parameter space is higher-dimensional? The third and arguably most important is the ability of the artifacts of inference to perform standalone inference on future observations. I first demonstrate the viability of the method, using a g-and-k distribution as a likelihood-free model. A bivariate g-and-k distribution serves to double the dimensionality and demonstrate scaling behavior. Last, I present inference performed using no expensive simulations.

4.1. g-and-k

Whereas most familiar models in statistics are defined in terms of their likelihoods, quantile distributions are defined in terms their quantile functions or inverse cumulative distribution functions (inverse CDF). Because the inverse CDF is immediately available, quantile distributions are trivial to sample from by transforming random variables according to the given quantile function–immediate inverse transform sampling. Corresponding probability densities (likelihoods) are commonly unavailable. The g-and-k distribution is a five-parameter quantile distribution defined in terms of the pth standard normal quantile, $z(p)$:

$$
\begin{aligned}
y = Q_{gk}\left(z(p) \mid A, B, g, k\right) = & A + B\left(1 + c\frac{1 - \exp(-gz(p))}{1 + \exp(-gz(p))}\right) \\
& \times \left(1 + z(p)^2\right)^k z(p)
\end{aligned}
\tag{4.1}
$$

A controls location, B scale, g skewness, and k kurtosis. By convention, c, an asymmetry-related parameter, is fixed at 0.8 (Allingham et al., 2009; Rayner and MacGillivray, 2002). Q is simply a transformation of a normal random variable. Therefore, to sample from this quantile distribution, $z(p)$ is drawn from a standard normal and transformed according to (4.1). Because the quantile function is the inverse of a corresponding density function, inverting a quantile function yields the density. If we were to consider the standard normal quantile function $z(p)$, the standard normal density would simply be $z^{-1}(p)$. Because Q is a distribution defined as a transformation of a standard normal random variable, the density can be expressed in terms of the standard normal density. Doing so requires the change of variable formula. Thus, the density for an observation $p(y \mid \theta)$ is given by

$$
p(y \mid \theta) = \phi\left(Q^{-1}(y \mid \theta) \mid 0, 1\right)\left(\frac{dQ(z(p)|\theta))}{dz(p)}\right)^{-1}, \tag{4.2}
$$

where ϕ is the standard normal density. However, Q^{-1} is unavailable analytically, rendering the likelihood intractable. The absence of a tractable likelihood and the interpretability of the parameters make the g-and-k distribution a natual choice for LFI experimentation.

Quantile distributions are useful beyond statistical curiosity. By granting direct control over all first four moments, quantile distributions permit the modeling of highly abnormal (literally) data. By setting $g, k = 0$, we obtain the normal distribution as a special case. However, highly skewed and kurtotic data, the likes of which often motivates inquiry in the field of statistical robustness, is equally well-modeled by quantile distributions. Allingham et al. (2009), for example, use quantile distributions to model response times on the Stroop task (Heathcote et al., 1991) and find quantile distributions to outperform a Gaussian mixture model fitted using MCMC.

I adopt the simulation study setup of Drovandi and Pettitt (2011). Uniform priors on A, B, g, and k with respective ranges $(0, 5)$, $(0, 5)$, $(-5, 5)$, $(-0.5, 5)$ are used. Observations are each represented by 10000 samples from a uniquely parameterized g-and-k distribution and are summarized by robust sample-based estimates of the first four moments, as proposed by Drovandi and Pettitt. As they discuss, different regions of the parameters space respond differently to number of observations and choice of summary statistics. This setup is chosen because it performs well across the parameter space, rather than optimally for individual parameters of interest.

To validate the multi-observation framework, I first sampled 10 sets of true parameters from the prior. I then simulated one observation (10000 samples) from each of the resulting, parameterized g-and-k distributions to represent independent observations. For each observation, I performed BOLFI with 35 points, 20 of which were initial, random evidence and 15 of which were actively chosen, optimal evidence. The dotted lines in figure A denote the posterior distributions obtained by use of the 35 acquisition points alone. I then collected all evidence from all observations (350 points) and trained a real NVP surrogate likelihood. In posession of the surrogate, I replaced the g-and-k simulator with the sampleable, surrogate model and continued with BOLFI inference until reaching 250 total acquisitions. The posteriors resulting from this are displayed by dashed red lines. The reference solutions, solid black, are obtained by use of uninterrupted BOLFI for 250 simulations from the true data-generating mechanism, the first 35 of which are the same as the first 35 of the hybrid approach.

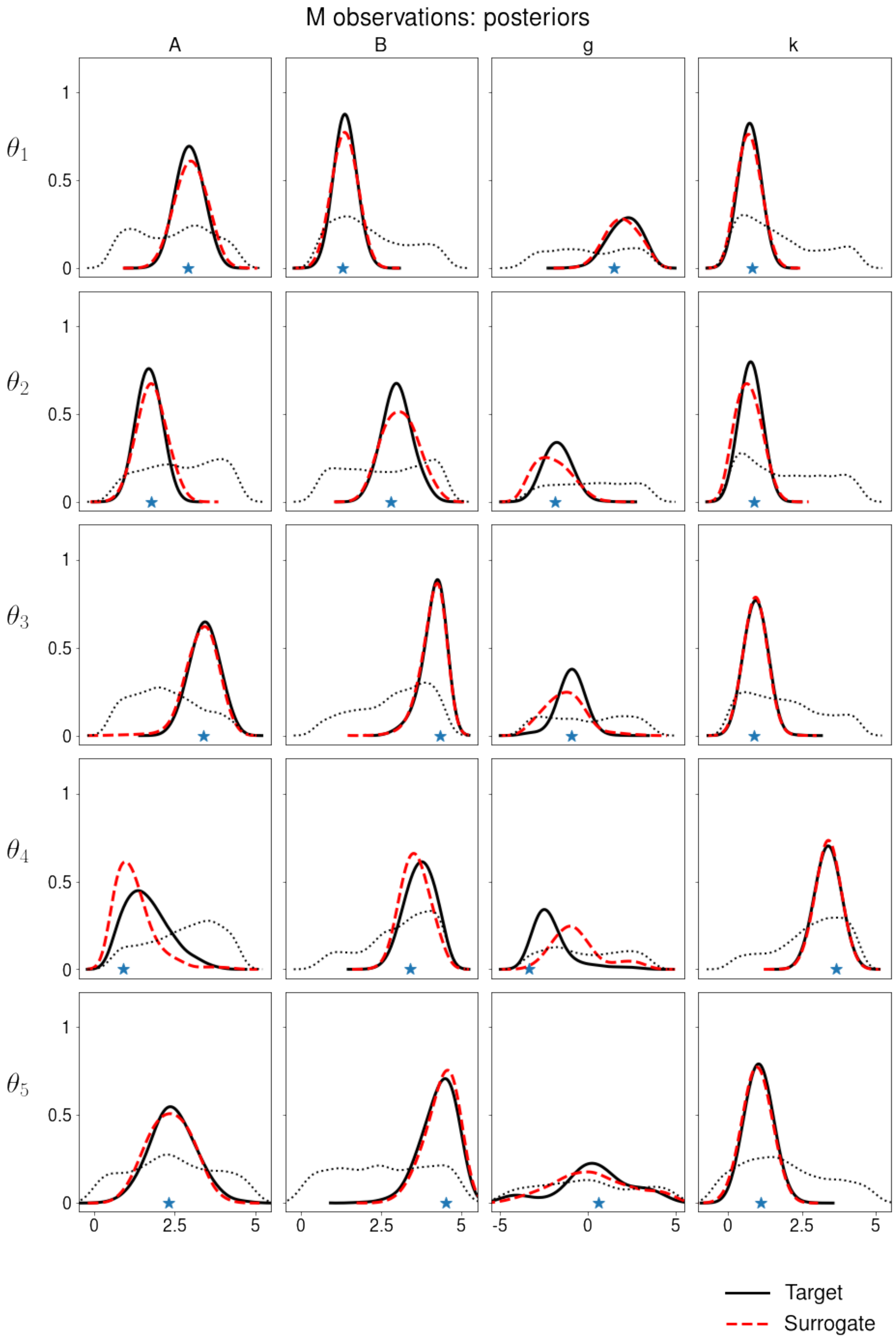


Figure 4.1: Marginal posteriors for each of the four parameters (A, B, g, and k) for first five observations (out of 10 total). True parameter values are indicated with stars. Reference solutions are recovered very well in most cases. Parameter g is recovered less well by both reference and surrogate methods.

4.1.1. Analysis

As noted above, the target of this method is the posterior distribution obtained by the underlying inference method, which is, in this case, BOLFI. High bias or variance in posteriors produced by the workflow is acceptable as long as they match the reference posteriors well. Matching but poor posteriors indicate shortcomings of the underlying inference rather than the surrogate strategy. In figure A, surrogate posteriors (red) can be seen to trace the reference posteriors (solid black) extremely well. In the vast majority of cases, the surrogate posteriors represents a dramatic improvement over the crude posteriors, more closely resembling the reference solutions than the crude solutions in all cases.

Parameter g is clearly solved less well. Though the prior support is double that of other parameters, the resulting 95% credible intervals are more than doubly wide. This resembles results obtained by Allingham et al. (2009); the 95% credible interval they obtain for g is at least an order of magnitude greater across all experiments. This may indicate insufficiency of the skewness summary statistic. All CIs reported here are considerably wider than those obtained by Allingham et al. (2009). To obtain their results, however, Allingham et al. (2009) perform 10^6 iterations of the ABC algorithm, necessitating as many calls to the simulator. Marginal CIs for all observations, parameters, and inference strategies can be found in appendix. However, to reiterate, while parameter recovery is important, *posterior recovery* is the primary objective. Here, peaked posteriors for 10 unique observations are obtained with 350 simulator queries, 35 each. The reference, pure BOLFI approach yields posteriors for all 10 observations with 2500 total simulator queries. This workflow enables posterior inference similar to and, in some cases, indistinguishable from pure BOLFI with nearly 10x fewer queries. For expensive simulators, this could mean the difference between several days and several months of computation. Furthermore, parameters drawn from the prior are recovered equally well to those hand-selected from well-behaved parameter space. Though there is undoubtedly room for improvement and robustness checking, these results are immensely encouraging and merit further investigation. We next double the dimensionality of the problem and examine its impact on performance.

4.2. Bivariate g-and-k

The g-and-k distribution can be straightforwardly generalized to multiple dimensions by letting $z(p)$ be multivariate normal and modeling each component of the observation vector y

as in (4.1). That is, $y_i = Q_i\left(z(p_i) \mid A_i, B_i, g_i, k_i\right)$, where $z(p_{i:N}) \sim \mathcal{N}(0, \Sigma)$. Naturally, the corresponding and unavailable density is

$$p(y \mid \theta) = \phi\left(Q_i^{-1}(y_i \mid \theta_i), \ldots, Q_N^{-1}(y_N \mid \theta_N)) \mid 0, \Sigma\right) \prod_{i=1}^{N} \left(\frac{dQ_i(z(p_i)|\theta))}{dz(p_i)}\right)^{-1} \tag{4.3}$$

Here, we assume unit variance and zero covariance among random variables z. Additionally, we assume each observation component is modeled by a g-and-k quantile distribution. Under these assumptions, for a bivariate distribution, we have $\theta = \{A_1, A_2, B_1, B_2, g_1, g_2, k_1, k_2\}$. Drovandi and Pettitt (2011) use a similar setup to model exchange rate returns of various currencies with autocorrelation structure.

Similarly to the univariate case, I model each parameter with a uniform prior and take 10000 samples from each distribution as an observation. The same moment-wise summary statistics are used in each dimension, yielding summarized data of dimension 8. Accordingly, more data is needed to arrive at useable results. I sampled 20 sets of true parameters from the priors and simulated observations. The initial fits again consisted of 35 expensive simulations, the last 15 of which were actively chosen. The surrogate likelihood was therefore trained on 700 evidence points. After the simulator-surrogate swap, inference proceeded until 350 acquisitions were made for each observation. Accordingly, reference solutions result from uninterrupted BOLFI with 350 expensive simulations. Abbreviated results are shown in figure 4.2; full results in appendix.

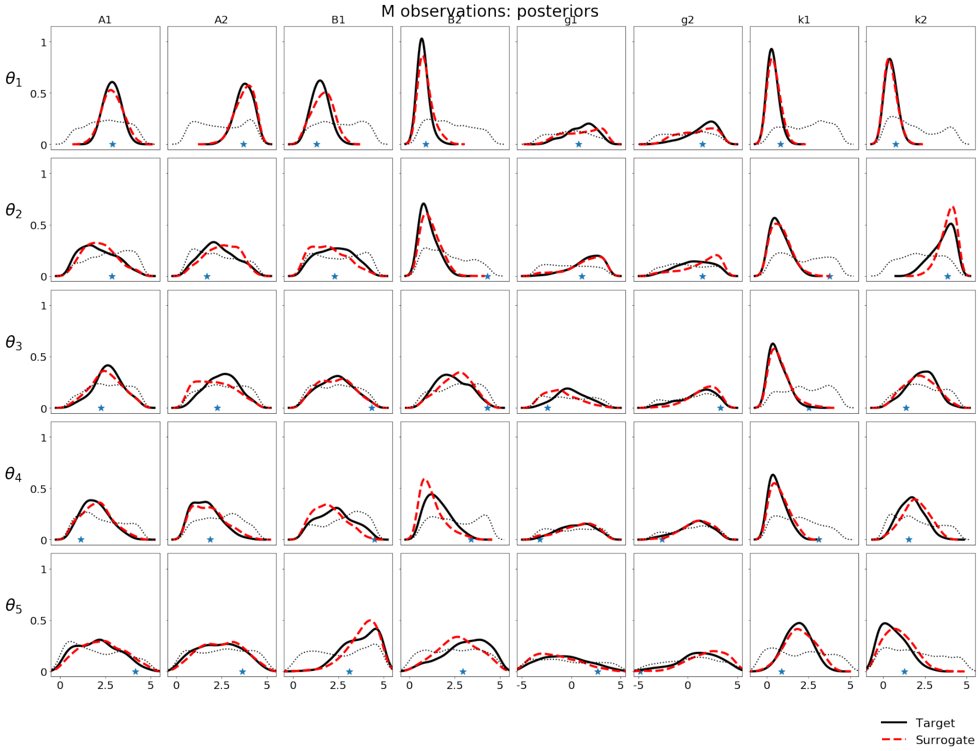


Figure 4.2: Marginal posteriors for each of the eight parameters (A, B, g, and k) for first five observations (out of 20 total). True parameter values are indicated with stars. Inferences can be seen to deteriorate in this higher dimensional space, but reference solutions are still well recovered in most cases.

4.2.1. Analysis

Inference on data generated by the bivariate g-and-k distribution are more diffuse and, in some cases, biased than univariate g-and-k data. The curse of dimensionality suggests that a parameter space with twice as many dimensions will require more than twice as many data to achieve the same quality inference. Here, we increased from four to eight parameters and merely doubled the number of expensive simulations. Unsurprisingly, the inferences deteriorated. However, recovery of the reference solutions by the surrogate remains largely unaltered; the 7000-simulation reference solutions, while unsatisfactory, are recovered by the 700-simulation surrogate method with similar fidelity to the univariate case. This is immensely encouraging. Though the underlying inference method suggests some difficulty with scaling behavior, the global surrogate appears to be more stable. I expect surrogate fidelity to scale non-linearly, however, because the volume of the global likelihood space will grow exponentially with parameter dimension. Accordingly, density estimation will become increasingly difficult. These results suggest that the limiting factor is the underlying inference method rather than the surrogate method. More analysis is merited.

4.3. New Observations

As noted above, the tremendous potential of amortizing inference in this way is the resulting artifact: a learned, approximate, global surrogate likelihood. A perfect surrogate can replace the simulator for all future computation. A poor surrogate should nonetheless provide fast, approximate computation. Having demonstrated that learned surrogates can *improve* inference with limited expensive simulations, we now examine the ability of the same surrogates to perform standalone inference.

To assess standalone inference on the univariate g-and-k distribution, I draw true parameters from the prior, simulate true data, and perform BOLFI with 250 surrogate simulations. The surrogate is the only source of likelihood information. The reference solution is BOLFI with 250 expensive simulations.

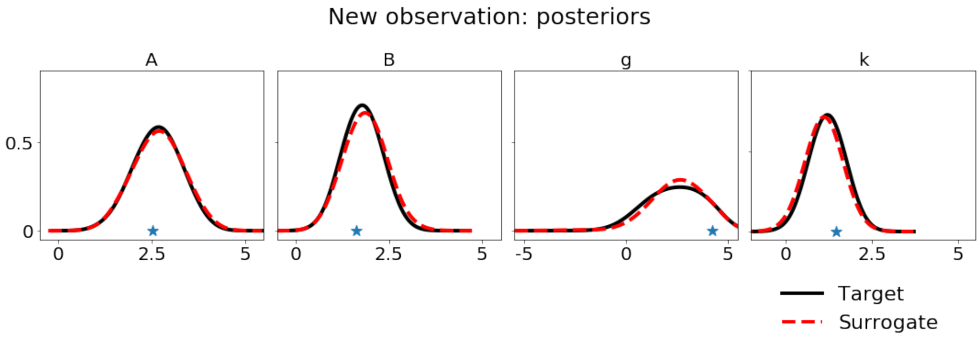


Figure 4.3: Marginal posteriors for each of the eight parameters. True parameter values are indicated with stars.

Similarly, to assess standalone inference on the bivariate g-and-k distribution, I draw true parameters from the prior, simulate true data, and perform BOLFI with 350 surrogate simulations; the reference solution is BOLFI with 350 expensive simulations.

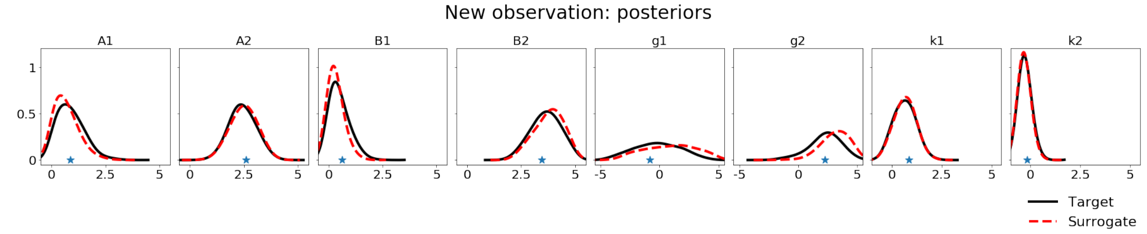


Figure 4.4: Marginal posteriors for each of the eight parameters. True parameter values are indicated with stars.

4.3.1. Analysis

With zero observation-specific simulations, the surrogate method can recover reference solutions. The resulting marginal posteriors are each unique and fit their respective marginal reference posteriors (i.e., the surrogate is not applying a standard approximation to all dimensions). These results suggest the true power of the method: after initial training, fast, high-fidelity inference can be performed with no expensive simulations, as may be required in online settings.

5. Discussion

A strong proof of concept with clear limitations and promising future directions.

5.1. Interpretations

Harnessing likelihood-free inference to a fast, global likelihood surrogate can recover posterior distributions for multiple observations with up to ten times fewer simulator queries; the workflow can enhance expensive, multi-observation LFI. Experiments performed on the g-and-k distribution, a practical and accessible generative process, indicate that the surrogate method is able to closely reproduce the reference solution with very few exceptions. The exceptions are mostly limited to those marginal posteriors that are poorly recovered by the reference method.

On a higher-dimensional problem, the bivariate g-and-k distribution, the workflow's performance appears to flag. Upon closer inspection, however, recovery of the reference solutions remains compelling. The reference solutions themselves deteriorate, resulting in deteriorating posteriors produced by the workflow. Though such deterioration is indication of limitations of the underlying inference method rather than the workflow, the choice of BOLFI as the underlying method is lightly baked into the design and could stand to be reconsidered. Strongly peaked, highly certain reference solutions appear to be recovered more consistently than broad and non-Gaussian solutions. The latter may simply not have converged and recovering not yet converged solutions may be a more difficult task.

High-quality inference on new observations, with likelihood information provided only by the surrogate, is evidently attainable. Reference solutions for sampled parameters not considered in the main run of the workflow are recovered as well or better than many of the original M solutions. Though the solutions produced exhibit relatively high uncertainty, the reference solutions are recovered well, suggesting that improvements in the tuning of the underlying inference method may improve the surrogate solutions.

5.2. Limitations

The most evident shortcomings are the limitations of BOLFI. Limitations of BOFLI include the restriction of the number of evidence points, decreasing performance in high-dimensional parameter space, and many tuning parameters. In general, LFI often requires extensive domain knowledge and familiarity with the problem at hand. Therefore, validating a method in the general case is a challenging task. BOLFI is particularly challenging. The obvious solution might be to choose an alternative underlying inference method that is 1) more general and robust and 2) can handle arbitrarily many simulated data so as to take full advantage of the fast, cheap surrogate simulator. Rejection-ABC, for example, fits the bill. BOLFI, however, was chosen for its immense computational savings over methods like rejection sampling. Though the workflow can reduce computational load by an order of magnitude, deferring to an alternative inference method, such as rejection-ABC, would discard the 2-3 orders of magnitude savings conferred by BOLFI. The workflow, in its present state, is therefore suited to problems that are known to be well-solved by BOLFI.

Consistent with expectations, inference deteriorates further in higher dimensions. However, reference and augmented solutions deteriorate similarly. Though this suggests the stable performance of the workflow in higher dimensions, usability remains dubious.

5.3. Implications

The proposed workflow *can* greatly accelerate multi-observation LFI. Similarly, complete amortization of inference is possible by use of this workflow. As it stands, statements about the generality and robustness of the approach are tenuous; generality and robustness of entirely amortized inference are more tenuous still. With these caveats, we can consider the potential for the development of this workflow. General, robust, accelerated multi-observation LFI could be a boon for the those working in the physical sciences, particularly life sciences. With standard approaches, inference must be performed anew for each observation, disregarding the scant possibility of reusable simulations. This approach is therefore applicable for any researcher who wants to use their model more than once. Because LFI is usually applicable only when models are too complex or expressive to be amenable to standard statistical treatment, significant development is usually involved in their formulation; using the model for multiple observations is therefore highly likely, if not expected.

As a tool for fully amortized inference, this method poses a unique opportunity to operationalize likelihood-free models at greater scales. Personalized medicine is a field that may benefit greatly. Lai et al. (2019), for example, have developed a complex simulation of multiscale cancer processes, which accounts for pharmacokinetics and pharmacodynamics. Inferring patient-specific parameters (i.e., treatments) for optimal outcomes would be of immense value in designing personalized treatment regimens. Even in its fastest setting, however, this model is prohibitively expensive. The use of a fast, global surrogate to amortize inference could enable on-demand inference and prescription for new patients.

6. Conclusion

Likelihood-free inference (LFI), which sits at the intersection of Bayesian inference and complex, mechanistic modeling, provides a principled, albeit exacting, approach to inference in the absence of the main workhorse: a tractable likelihood function. LFI methods generally proceed by forward simulating the mechanistic model and comparing simulations to observations; parameters that generate synthetic data most similar to observed data are assumed to be similar to true data-generating parameters. In many cases, these methods require prohibitively many simulations. Furthermore, modern methods are not designed with inference on multiple observations in mind. This capstone introduces, develops, and validates a novel workflow for performing LFI on high-cost simulators for multiple observations, which leverages the sample-efficiency of BOLFI and flexibility of neural density estimators.

The workflow is first situated in the relevant context, namely probabilistic inverse inference. Details of the progression and implementation of the workflow are then presented. Lastly, the workflow is validated on accessible, interpretable problems.

This capstone has succeeded in demonstrating a proof of concept for a novel likelihood-free inference workflow. In experiments, the workflow is shown to recover the target distribution with nearly ten times fewer calls to the simulator. This constitutes immense computational savings and could correspond to the ability to perform formerly intractable inference. I hasten to mention that, in its present form, the workflow has limitations and is not yet a general tool. Solutions provided by the workflow cannot reliably outperform the underlying inference method. The workflow is therefore presently suited only to problems that are known to be well-solved by BOLFI. Furthermore, fidelity of the density estimation can be expected to decrease with dimensionality in limited data settings. However, because BOLFI struggles similarly in higher dimensions, attributing decreased performance to one or the other is difficult. The complexity and extensive tuning of the workflow will also need to be reduced before it can be used widely.

As it stands, these results presented here suggest promising future directions. In a theoretical vein, connecting the formulation used here with latent variable modeling will enable use of the extensive, existing hierarchical modeling machinery. Methodologically, improvements to the generality and efficiency of LFI methods, chiefly BOLFI, will translate directly to improved performance of the workflow. Practically, LFI success remains largely dependent on individual problems and practitioner input; incorporating advances from meta-learning and experimental design may prove fruitful for breaking this dependence. As models evolve, so too must inference. This is an attempt at evolution.

References

Allingham, D., King, R. A. R., and Mengersen, K. L. (2009). Bayesian estimation of quantile distributions. *Statistics and Computing*, 19(2):189–201.

Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. (2019). Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*.

Arsenii, A. (2018). Real nvp pytorch a minimal working example. urlhttps://github.com/ars-ashuha/real-nvp-pytorch.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bishop, C. M. (1994). Mixture density networks.

Brochu, E., Cora, V. M., and De Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Cranmer, K., Brehmer, J., and Louppe, G. (2019). The frontier of simulation-based inference.

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.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55(9):2541–2556.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A. and Shalizi, C. R. (2012). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Gutmann, M. U., Corander, J., et al. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, pages 49–52.

Heathcote, A., Popiel, S. J., and Mewhort, D. (1991). Analysis of response time distributions: an example using the stroop task. *Psychological bulletin*, 109(2):340.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Järvenpää, M., Gutmann, M. U., Pleska, A., Vehtari, A., Marttinen, P., et al. (2019). Efficient acquisition rules for model-based approximate bayesian computation. *Bayesian Analysis*, 14(2):595–622.

Kabanikhin, S. I. (2008). Definitions and examples of inverse and ill-posed problems. *Journal of Inverse and Ill-Posed Problems*, 16(4):317–357.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lai, X., Geier, O. M., Fleischer, T., Garred, Ø., Borgen, E., Funke, S. W., Kumar, S., Rognes, M. E., Seierstad, T., Børresen-Dale, A.-L., et al. (2019). Toward personalized computer simulation of breast cancer treatment: A multiscale pharmacokinetic and pharmacodynamic model informed by multitype patient data. *Cancer research*, 79(16):4293–4304.

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.

Lintusaari, J., Vuollekoski, H., Kangasrääsiö, A., Skytén, K., Järvenpää, M., Marttinen, P., Gutmann, M. U., Vehtari, A., Corander, J., and Kaski, S. (2018). Elfi: Engine for likelihood-free inference. *Journal of Machine Learning Research*, 19(16):1–7.

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. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1289–1299. Curran Associates, Inc.

MacKay, D. (2003). Information theory, pattern recognition and neural networks.

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.

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Owen, J., Wilkinson, D. J., and Gillespie, C. S. (2015). Likelihood free inference for markov processes: a comparison. *Statistical applications in genetics and molecular biology*, 14(2):189–209.

Papamakarios, G. (2019). Neural density estimation and likelihood-free inference. *arXiv preprint arXiv:1910.13233*.

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.

Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T., editors, *Proceedings of* *the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Shen, P., Lees, J. A., Bee, G. C. W., Brown, S. P., and Weiser, J. N. (2019). Pneumococcal quorum sensing drives an asymmetric owner–intruder competitive strategy during carriage via the competence regulon. *Nature microbiology*, 4(1):198.

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.

Theis, L., Oord, A. v. d., and Bethge, M. (2015). A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Titsias, M. K. and Ruiz, F. J. (2018). Unbiased implicit variational inference. *arXiv preprint arXiv:1808.02078*.

Tran, D., Ranganath, R., and Blei, D. (2017). Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533.

Turner, B. M. and Van Zandt, T. (2014). Hierarchical approximate bayesian computation. *Psychometrika*, 79(2):185–209.

von Würtemberg, I. (2011). Ill-posed problems and their applications to climate research.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102.

Appendix 1: Capstone LOs

#connect

Connecting with seasoned researchers and developing a capstone from the resulting effort was not coincidence. As early as January 2019, I began seeking skilled research professionals with whom to connect. My plan was always to find a research group, collaborate to generate meaningful research, and adapt it into a capstone. I did so by joining the probabilistic machine learning group at Aalto University as a research assistant. In the interview, floated the idea crafting a bachelor's thesis from whatever progress we made, and the response was positive. As a result, in the early stages of this project, I was in close contact with two professional, domain experts, Prof. Samuel Kaski and Henri Pesonen. In addition leading the probabilistic machine learning group at Aalto University, Professor Kaski serves as the director of the Finnish Center for Artificial Intelligence (FCAI), of which Simulator-Based Inference is a primary research priority. His primary contribution was catalyzing and funding the initial stages of the research. In addition, he offered high-level guidance and connected me with other knowledgeable people (e.g., a PhD student who suggested trying out real NVP). Henri, formerly a postdoctoral researcher at Aalto University and currently a postdoctoral researcher at the University of Oslo, has been my closest collaborator. He first proposed an earlier version of the workflow and aided in my development of the work presented here. We are presently developing the hierarchical Bayesian formulation of this work.

#planningarchitecture

Three mechanisms have been instrumental in the completion of this project: a technique for maximizing productivity in short periods, consistent feedback with a familiar peer, and a big-picture planning and strategizing tool.

POMODORO The Pomodoro technique entails structured work time interspersed with structured breaks; 25 minutes on, 5 minutes off. I have employed this method extensively with great success, particularly when deadlines have been near. I thought myself above this sort of scheme, believing that I was capable of motivating and regulating myself. It was at the request/advice of a classmate that we give the method a try. It is the only way I have worked since. The method is particularly effective when used with others. We each state our goals for the 25 minutes, work, and share what we accomplished at the end of each work session. Strategic use of this method has proven particularly useful near the end of the project.

PEER FEEDBACK More details about the feedback itself and how it was used are available in the #feedback justification. The justification here concerns the mechanism itself and the accountability it entailed. Josh Broomberg was the peer reviewer of my project for both peer feedback assignments by design. He gives extensive, in-depth, critical feedback, much of it actionable. Electing to peer review works twice meant that I could implement the changes he suggested and then see if they satisfied his original concerns. Furthermore, stating plans and goal to one another in the fall and exchanging again in the spring creates an additional, automatic accountability mechanism. I can see his document on my Overleaf homepage and how long it has been since he edited it (currently 9 hours).

TRELLO BOARD As noted in past submissions, Trello has been an important tool for managing both the timeline and content of the final capstone submission. Each section and some subsection each had a corresponding deadline. Other general tasks like proof reading and organizing submission materials share a list. To ensure that I meet criteria for HC and LO tagging, HCs, capstone LOs, and project-specific LOs are each organized in their own list with cards for each #. This is an easier way to track, edit, and organize the different #s than in LaTeX.

#metrics

Different metrics have been useful at different times over the course of this project. Broadly, the metrics evolved to first assess completion, then refinement, and finally presentation.

During the main thrust of writing the capstone, deadlines were the most prevalent and useful guideline I used to evaluate my work products. During that phase, quantity *was* quality. I evaluated work products based on their existence and where their existence fell in my timeline. Deadlines were tracked using Trello cards.

In possession of complete sections, my attention turned to refining the explication, reasoning, and voice. Peer comprehension served as the primary assessment of explication. My main question to Josh in the second feedback assignment was "do you understand this?" The answer was a resounding "No." As a result, I extended the density estimation, Bayesian optimization, and Methods sections to explicitly, painstakingly grant the reader access to the workflow and its components. This extension totaled approximately 2500 words. To assess my reasoning and chosen voice, I turned to contemporary literature in the field. For a particular example, Lueckmann et al. (2017) is a is a paper that I used and cited in this project and is an example of the style of paper that this aspires to be. In determining whether the structure of my introduction was correct and sufficiently rigorous/comprehensive/readable, I referred to this paper many times and adopted some organizational conventions. The styling of papers in the same field and of the same scope was used as an organizational and tonal metric.

Near the end, comments and grades on the Full Draft proved to be the most definite and trustworthy metrics, as they came from the grader himself. Thus, for all HC and LO justifications, I am striving to give as much detail as I do in the #PlanningArchitecture tag, per "Aim for this level of detail on ALL of your HC and LO application descriptions."

#feedback

Feedback came from two primary sources, Josh Broomberg and Professor Wilkins.

Josh's feedback comprised formatting, prose signposting, questions about content. In response, I altered my formatting to make section designations clearer, added periodic introduction and summarization, and fleshed out sections he identified as hard to understand.

With respect to Professor Wilkins' feedback, the most significant changes made were greatly extending the discussion of active learning and acquisition functions for clarity, going into far greater depth on HC and LO justifications, and removing most meta-commentary. Of course, these are only a subset of the changes made in response to feedback.

#research

I have endeavoured to include my knowledge, textbook-level content (e.g., Gelman, Mackay), general resources from the field (e.g., Beaumont, Pritchard), and highly-relevant, recent developments (e.g., Alsing block quote).

In particular, the connections between Bayesian inference, inverse modeling, and well-posedness were new to me. I became aware of this constellation by contemplating how to situate/motivate my project. It ended up being a rather long section because I found it extremely interesting. There is surprisingly little literature on the topic, so I found what I could and connected it.

I demonstrate deep engagement with the resources central to the workflow, such as the paper introducing BOLFI. To explain BOLFI, I bring in other resources about Bayesian optimization and Gaussian processes.

To situate the workflow and its implications on the current LFI landscape, I draw on an article discussing the frontiers of LFI. The article describes three "frontiers," two of which are directly addressed by the workflow.

#accountability

The existence and submission of a complete, polished, and impactful Capstone project serves to demonstrate agency, stewardship, self-efficacy, and, most of all, ownership of this self-directed project. This is the justification.

COMMITMENTS AND DEADLINES I attended and contributed to every CP193/194 session and used no assignment extensions. I delivered all peer feedback in a timely manner and spent 4-5 hours sharing the feedback in person.

PLANNING I established and met deadlines for writing sections; e.g, "Background" due Jan. 23, "Methods" due Feb. 3. I used a centralized Trello board to track both the deadlines and progress on each section. The choice of this tool also enabled me to make notes about each section–notes like what to change in the next draft.

DEALING WITH SETBACKS I encountered setbacks and made progress despite them. This was possible because I designed my timeline to have some flexibility (i.e., completing all sections ahead of schedule). Though the neuron simulation worked well, its parameters were not well-solved by my underlying inference method of choice, which made use of the workflow impossible. Despite this, I have produced meaningful results, albeit in a different direction (testing scaling behavior and standalone inference), which are arguably more impactful.

EFFECTIVE PERFORMANCE Use of the Pomodoro method, as described in the #PlanningArchitecture justification helped me make the most effective use of my time, thereby mitigating behaviors that impaired effective performance.

#qualitydeliverables

Scope, depth, and rigor. These are the features I need to be sure to include. Justification:

SCOPE From the Capstone handbook: " it should be a product of professional quality and should make a meaningful contribution to your field."

I hope you'll agree that this Capstone achieves this. It addresses a well-known, contemporary challenge in the field and is written in a style reflecting the style of the field. Whether the contribution is meaningful remains to be seen, but it certainly exhibits potentiality.

DEPTH The recommended word counts for Statistics/Data Analysis and Research projects are 10,000-15,000 and 15,000-20,000 respectively. This is at the intersection of Statistics and Research and is squarely within the latter range accordingly. Introducing modeling and Bayesian inference is necessary for accommodating *all* competent, academic readers. I cover all of the necessary material for understanding the main research contribution.

RIGOR Rigor is maintained throughout by e.g., demonstrating the derivation of Bayes' rule, explaining the intractability of the probability densities associated with g-and-k distributions, and noting why the particular construction of real NVP allows for modeling scaling and translation functions arbitrarily.

APPENDIX 2: COURSE LOS

#ProbabilityTheory

Probability theory pervades this project. I deliver the most obvious application of this LO by applying rules of probability theory to derive Bayes' theorem; while perhaps uninventive, it is nonetheless rigorous and thorough. Subtlety, at no expense to rigor, arrives in my navigation of the likelihood term–a probability density when regarded as a function of data and a likelihood when regarded as a function of parameters. Perhaps the most nuanced foray into probability theory, however, is in my explanation that some latent variable models have natural hierarchical interpretation but that the latent variable model I have chosen, real NVP, does not entertain such an interpretation, citing the lack of conditional independence of the likelihood from hyperparameters.

#PythonImplementation (CS146)

I implement a novel workflow which leads to state of the art results in Python, making careful, practiced use of a specialized probabilistic program (ELFI) and supplementing with Python modules of my own creation. Great attention is paid to data types and dimension in order to pipeline data and results. Similarly, random seeds are used throughout and altered deterministically between observations and replications to ward off any unwanted determinism. Results are presented in a tidy, sophisticated format. Code is available in Jupyter notebooks with accompanying comments and prose descriptions to aid navigation. As a nifty bonus, I used a for loop to compute 95% credible intervals for each posterior and simultaneously printed LaTeX tabular syntax for typesetting the results directly to the document.

After receiving feedback that comments and annotations were too sparse, I fleshed out comments in every code file, with docstrings where appropriate and markdown cells explaining the procedure.

#QuantCommunication

Though this LO is applicable throughout the project, its application is particularly critical in some particular sections. In explaining the method, I layer explanatory tools in order to achieve clarity without sacrificing rigor. I introduce the workflow holistically by presenting a diagram and step-by-step procedure. I then decompose the diagram; each diagram component is revisited in turn and the associated theory and methodology is presented. Furthermore, I include implementation details to offer grounding for the practical-minded reader. The markdown detail included in the attached code notebooks completes the chain, theory to methodology to implementation to application.

Throughout, I include equations only when they aid the exposition. For example, I include the integral of the Gaussian distribution with respect to two different quantities to demonstrate the difference between likelihood functions and probability densities. I include the evidence lower bound of the KL divergence specifically to demonstrate that it requires access to the likelihood function. The acquisition functions I choose to include are either useful for building intuition about the active learning step or is used in the experimentation itself. I use notation that is consistent with the fields that I draw from, inventing only the layout of the workflow diagram.

#MaximumLikelihood

Maximum likelihood estimation is first explored theoretically as an alternative to Bayesian inference for inverse problem solving. It falls under the category of deterministic inverse inference. This exploration aids in the discussion of inverse problem solving and why the emphasis is on Bayesian methods. I go on to explain, implement, and employ a sophisticated maximum likelihood training framework for learning a global surrogate likelihood. Specifically, the loss function used to train the surrogate is the negative log likelihood of data under the model. I go on to outline how sample fidelity (visual appearance of samples) and log-likelihood are not interchangeable. This frames some possible concerns about employing classical Bayesian inference methods on a likelihood surrogate.

#RegressionAlgorithm

The sample efficiency and consequential success of the method proposed in this capstone is due in large part to the choice and use of a highly sophisticated extension (Bayesian optimization) to an infinite-dimensional regression algorithm (Gaussian process). I detail the workings of the algorithm and the cite various features, including sample efficiency and probabilistic modeling, that make it the only viable option for performing the type of inference I demonstrate. Though the implementation is handled by a sophisticated software package, deep knowledge of both the method and the particular implementation were necessary to use and monitor the performance of the algorithm. For example, I cite the computational complexity of updating the model to explain why the algorithm has trouble dealing with many data.

#InterpretResults

This LO is present throughout the paper. Chiefly I introduce and discuss the theory and methods for robustly inferring simulation parameters from data. This includes a discussion of summarizing simulation results (summary statistics), the difficulties associated with fully-Bayesian methods for simulation analysis, and how to interpret results from simulations as information about the implicit likelihood and posterior. In the first experiments, summarization of simulation results comes by way of robust estimates of mean, variance, skewness, and kurtosis discusses by Drovandi and Pettitt (2011). The workflow is employed to interpret “observed” simulation results by adaptively and optimally querying the simulator, pooling information, and replacing the simulator with an inexpensive surrogate.

The key application, however, is my interpretation of simulations as samples from the inaccessible, implicit likelihood, summarizing these simulations, and approximating the implicit likelihood.

Appendix 3: HCs

#Professionalism

This submission is typeset in LaTeX, as is consistent with convention for the field, and which serves to neatly communicate the content. Math and statistics notation is consistent throughout the document and with convention for the field. The document has been proofread several times in order to minimize typos and grammatical errors. Sources are used in a variety of ways and cited according to APA guidelines. In particular, I make use of inline and parenthetical citations and block quotations when necessary. When relevant content exceeds the scope of the project, the reader is directed to authoritative resources. Serif font is chosen over sans-serif to encourage comprehension and ease of reading rather than speed.

#Organization

Overarching organization of the capstone mimics the standard format of academic writing in the field. I have additionally striven to weave a narrative that persists throughout the paper, one highlighting the interplay between inference, science, model selection, and statistics. In this way, the sections should not only stand alone in the field-scripted order, but also build on one another to leave the reader with both a strong sense of the proposed workflow but also the significance of research in this field.

I took the time, this round, to slightly rethink and redesign the workflow diagram, which directly affected how I organized the *Methods* section. The arrangement of nodes and arrows corresponds to the logical organization of the section.

#EvidenceBased

Multiple arguments coalesce to form the main argument, that this workflow solves takes steps towards multi-observation LFI on expensive simulators. Each argument is evidenced, in turn.

The evidenced by the argumentation structure , which proceeds thematically as Inference–>Bayesian inference–>LFI–>multi-obs/expensive shortcomings–>Proposed solution validation. Evidencing these arguments proceeds in the same fashion, noting first the difficulty of inverse problems due to frequent ill-posedness, then enumerating Bayesian methods, all of which rely on the likelihood function, addressing this shortcoming of Bayesian inference by introducing the various flavors of LFI, and finally introducing additional shortcomings of modern LFI. The final entry in the sequence of argumentation, the proposed solution itself, is evidenced empirically.

Sub-arguments are also evidenced and justified. For example, the choice of BOLFI as the underlying inference method follows from its evidenced sample-efficiency; the choice to use real NVP follows from the requirements of surrogates and its empirical performance. Similarly, the choice to address amortized inference and sample-efficiency follows from contemporary literature, which cites these two things as "frontiers" of likelihood-free inference.

#SourceQuality

Introduction is densely cited with correct, relevant, high-quality sources, as is consistent with standard practice in the field. I styled the introduction after Lueckmann et al. (2017). Not only are claims and information gleaned from external sources cited, but experiment setups and data visualization design are adapted from literature. High-quality sources, which I have engaged deeply are used throughout.

Currency: Most papers used heavily (e.g., , ,) are less than five years old. This is a rapidly evolving area of research. Since the commencement of this project, similar approaches have cropped up. I took the time to address those that I have come across in an effort to be as current as possible. For more fundamental topics, I referred primarily to a well-respected textbook () that has been expanded and re-released with more content and more authors as a third edition within the last ten years.

Relevance: No articles are included arbitrarily. Citations simply indicate the resources I used to answer questions I had during the research process. They are relevant by virtue of my seeking them out. Resources that did not direct or aid the research are not included.

Authority: The heavily used resources are all either journal or conference publications. BOLFI, for example, was published in the Journal of Machine Learning Research (JMLR), which is a well respected, open-access journal with one of the highest h-indexes among machine learning and artificial intelligence journals.

Accuracy: Though the reliability of academic papers is often hard to assess, particularly for recent publications, I pointed out what I believe is a shortcoming of many LFI papers: inferring hand-selected parameters rather than sampling parameters from the prior. I chose to infer sampled parameters partly because there was no other principled way of generating multiple "observations."

Purpose: I assumed all authors' motivations included (but were not limited to) getting published. Results, therefore, are assumed to best case scenarios obtained after sufficient, undocumented tuning.

#Responsibility

From the Capstone handbook: "In some cases, pairs or groups of HCs and LOs may function together. If so, you may explain their relevance in a single, longer appendix entry. Similarly, some of the HC/LO annotations could reference other parts of the appendix to minimize redundancy."

My application of this HC is best summarized by the justifications for #Accountability, #PlanningArchitecture and #Strategize. In particular, #PlanningArchitecture describes the planning and effective working tools I used, #Accountability describes how I managed to use the tools effectively, and #Strategize describes my strategy for driving effective social interaction in the interest of this Capstone.

#Probability

Because this submission is ripe with applications of this HC, I've organized a few highlight according to parts of the HC rubric:

Accurately and effectively applies an appropriate probability or interprets a probability in a complex or sophisticated context.: After introducing the likelihood term of Bayes' theorem, I explain why it is *not* a probability density, offering mathematical and intuitive justification. I also explain that probability density functions can be natural ways to express likelihood functions, which is on possible source of confusion. Later, I expand the definition of the likelihood to define likelihood-free models.

Accurately calculates an appropriate probability with clear detailed steps: I offer a clear, tractable derivation of Bayes' theorem, with detailed interpretation of each term. The result is the conditional probability of parameters conditioned on data.

Accurately applies appropriate probabilities with well justified reasoning: While explaining a number of classical Bayesian methods, I specifically shoehorn in math and probability statements containing the likelihood term (e.g., MAP, ELBO). Not only does this give insight into the role that the likelihood plays in computation, but it sets up the subsequent sections: why inference is hard without a likelihood.

Uses probability in a creative and effective way, relying on a novel perspective: A global, surrogate likelihood has not, to my knowledge, been implemented or applied in the manner presented here. While it is related to other synthetic likelihood approaches, this the first to learn a conditional, global surrogate likelihood for enhancing likelihood-free inference. Inferring the posterior distribution over parameters conditional on data *without a likelihood* is challenging. Doing so by use of an approximation to the distribution of summary statistics conditional on parameters is cutting edge.

#Distributions

This HC justification is the same as in the full draft, with additional applications from the project highlighted. Identification, application, and justification of distributions and their characteristics also pervades this capstone. Highlighted applications:

PROBABILISTIC INFERENCE: A detailed and rigorous explanation of the likelihood term in Bayes' theorem revolves around the use of the familiar normal distribution. This distribution and its characteristics are used to show how altering the argument of the function (random variable vs parameters) impacts whether the quantity is a valid probability density.

REAL NVP: I enter into a discussion about distributions over latent spaces. The ease of sampling and likelihood evaluation conferred by the choice to model the latent space as unit Gaussian is highlighted.

G-AND-K DISTRIBUTION: This is the most in-depth and non-standard contemplation of distributions. I use distributions that are not defined in the standard way (by their probability mass/density function), explain their relationships to standard distributions, and show why this yields both flexible generative processes and intractable likelihoods. I go on to use these distributions for validating a novel workflow.

#Modeling

In this capstone, I apply every aspect of the #modeling HC.

To begin, I offer a fundamental, nuanced take on modeling, citing forward modeling (classical "modeling") and inverse modeling (parameter estimation) as the two fundamental and complementary steps. I then frame various aspects of probability theory as tools for performing inverse modeling. To conclude the theoretical and methodological exposition of modeling, I introduce likelihood-free inference as inverse modeling for arbitrary, realistic forward models.

In an applied fashion, I introduce and develop a novel strategy for inverse modeling on statistically intractable models. In doing so, I both use a variety of appropriate models (e.g., GP regression, neural networks) and apply the strategy to relevant models (e.g., bivariate g-and-k).

Some particular applications: - I thoroughly motivate the choice of real nvp as the surrogate model of choice in the methods section, citing the important features of an appropriate surrogate model. - I weave the likelihood, both in prose and statements of math, into my explanation of Bayesian inference so that the impracticality of classical Bayesian methods without tractable likelihoods is clear. - In addition to computing posteriors for M observations using both expensive and surrogate simulations, I demonstrate the ability of the workflow to perform complete, fully-amortized likelihood-free inference with no expensive simulations. This is something I didn't think would be possible upon the commencement of this project but is extremely exciting.

#Context

The Abstract, Introduction, and Background sections give the relevant context.

Justification: Beginning with Bayesian inference, proceeding to likelihood-free inference, and then introducing the specific problem would be sufficient to present the workflow. However, I choose to additionally situate these topics within the context of forward and inverse problem solving. This context is not typically associated with statistical inference. I have included it because it 1) uniquely situates deterministic inference alongside probabilistic inference as competing solutions to the same problem, 2) defines "models" as general solutions to forward problems rather than equations or probability distributions, which accommodates regarding simulations as models, and 3) lays the foundation for relating likelihood-free inference to enabling science. Indeed, likelihood-free inference may be most interesting and transformative for its accommodating arbitrary complex models of the world. Providing this context therefore further illustrates the significance of the development presented later in the paper.

#Estimation

The Background section contains an in-depth and non-standard introduction to statistical inference or parameter estimation. I describe two different paradigms before providing a survey of methods within the paradigm that I primarily use. In my explanation of distributional approximations, I also offer a nuanced distinction between inference at the parameter level and inference at the function level. Variational inference performs deterministic approximate inference at the function level in order to obtain probabilistic estimates at the parameter level.

Real NVP is my clearest and most substantive application of estimation. The motivation for using density estimation as a tool for surrogate inference was born of developments in density estimation for likelihood-free inference. Real NVP is shown to be an appropriate choice of density estimation tool due to its 1) flexibility, 2) sampleability, and 3) sample efficiency. A sophisticated version (conditional density estimation with real NVP) of the estimator is implemented with a similarly sophisticated training strategy. Combining density estimation with other LFI methods in this way yields a unique and performant workflow.

#GapAnalysis

In the Motivation section, I identify and evaluate two shortcomings of modern LFI: solving multiple observations and inference on expensive generative processes. With respect to multiple observations, I thoroughly describe the gap (that most methods inefficiently solve multiple observations) and provide reasoning for why they fail to address the gap (locally optimal simulations, small regions of posterior mass/density). I again describe the difficulty of inference on expensive simulators and describe how the problem is compounded by multiple observations.

In this case, however, I recognize that an existing solution (BOLFI) returns promising results; thus, I utilize BOLFI as a feature of my proposed solution. The nuance to my application is 1) identifying non-obvious gaps, 2) considering and addressing their interplay, and 3) recognizing that a viable if imperfect solution exists and adapting it into the solution.

#Purpose

A quote from the Milestone 7 assignment of CP192 (spring 2019):

> *My intended capstone project: expand on summer undergraduate research sponsored by another university. However, the nature of my intended capstone bakes in the necessary critique and validation: from the research supervisors. My earliest capstone intentions were 1) that I create/do something and 2) that I do it under the auspices of a skilled professional from whom I can learn. This evolved into seeking research positions that could be extended in such a way.*

In designing my capstone, original research was always the objective. Given the difficulty of that task, I sought research positions with the intention of adapting the work that resulted into a capstone project. When consulted on this strategy, my academic advisor and CP192 instructor expressed that this is one of the best ways to undertake a capstone. I acknowledged my own limitations, supplemented them with the knowledge and experience of an external professor and postdoc, and have, as a result, managed to develop a capstone consisting of original, impactful research.

#InterpretiveLens

In the Background section on $p(\theta)$, I regard the prior distribution as a prior experience or expectation. This feature of Bayesian inference, as I indicate, receives criticism for being subjective–choice of prior can significantly impact inferences. I take the time to address and begin to dispel these criticisms by noting 1) assumptions are always present and the prior is one form they take in Bayesian inference, 2) they are sometimes necessary when problems are under-constrained, and 3) that they really shouldn't be subjective at all.

The main thrust of the application is explaining that these "prior experiences" are actually not altogether subjective or terribly impacted by practitioners' biases. There is also an implicit second application: the assumption that people regard Bayesian inference with undue skepticism, acknowledging a potential lens of the reader. The commentary responds to this directly.

#GapAnalysis

In the Motivation section, I identify and evaluate two shortcomings of modern LFI: solving multiple observations and inference on expensive generative processes. With respect to multiple observations, I thoroughly describe the gap (that most methods inefficiently solve multiple observations) and provide reasoning for why they fail to address the gap (locally optimal simulations, small regions of posterior mass/density). I again describe the difficulty of inference on expensive simulators and describe how the problem is compounded by multiple observations. In this case, however, I recognize that an existing solution (BOLFI) returns promising results; thus, I utilize BOLFI as a feature of my proposed solution. The nuance to my application is 1) identifying non-obvious gaps, 2) considering and addressing their interplay, and 3) recognizing that a viable if imperfect solution exists and adapting it into the solution.

#Audience

Formality in writing is useful to the extent that it preserves rigor, but is adverse to the extent that it hinders understanding. The size of this paper, in contrast to, for example, a conference paper, permits added emphasis on the understanding of the reader. Therefore, rather than something rigid and dogmatic, I have adopted a tone closer to what might be found in a textbook and packaged it in a long-form paper. In this way, both the pedagogical intentions and form factor are preserved. This caters to the likely readership while respecting the conventions of the field that it belongs to. I include a discussion of Bayesian inference for those unfamiliar but, in describing the organization of the paper, suggest that readers familiar with it proceed to the likelihood-free inference section. This avails the paper to the academic audience generally without miring domain experts in background and review.

A. Full Results

A			
True θ	Crude	Reference	Augmented
$\theta = 3.00$	(0.13, 4.81)	(2.18, 3.95)	(2.06, 4.14)
$\theta = 1.59$	(0.16, 4.86)	(0.67, 2.28)	(0.54, 2.65)
$\theta = 3.61$	(0.12, 4.84)	(2.65, 4.68)	(2.37, 4.60)
$\theta = 0.50$	(0.19, 4.86)	(0.11, 3.56)	(0.05, 2.87)
$\theta = 2.30$	(0.10, 4.82)	(1.11, 3.80)	(1.01, 3.61)
$\theta = 2.16$	(0.13, 4.85)	(0.43, 4.59)	(0.39, 4.32)
$\theta = 3.68$	(0.12, 4.89)	(0.67, 4.91)	(1.65, 4.84)
$\theta = 0.08$	(0.09, 4.85)	(0.17, 1.84)	(0.07, 1.51)
$\theta = 4.32$	(0.13, 4.90)	(3.05, 4.83)	(3.23, 4.85)
$\theta = 3.81$	(0.13, 4.83)	(1.69, 4.84)	(2.60, 4.69)
Avg width	4.72	2.66	2.41
Avg Reduction	-	2.06	2.31

B			
True θ	Crude	Reference	Augmented
$\theta = 1.00$	(0.10, 4.83)	(0.43, 1.79)	(0.34, 1.92)
$\theta = 2.86$	(0.12, 4.89)	(2.11, 4.08)	(2.00, 4.45)
$\theta = 4.77$	(0.18, 4.91)	(3.49, 4.98)	(3.52, 4.99)
$\theta = 3.61$	(0.17, 4.90)	(2.94, 4.90)	(2.88, 4.80)
$\theta = 4.52$	(0.15, 4.89)	(3.25, 4.97)	(3.28, 4.97)
$\theta = 4.89$	(0.22, 4.92)	(3.11, 4.98)	(3.46, 4.98)
$\theta = 2.48$	(0.15, 4.88)	(1.77, 4.01)	(1.75, 3.79)
$\theta = 0.50$	(0.10, 4.84)	(0.04, 1.19)	(0.05, 1.10)
$\theta = 4.42$	(0.10, 4.90)	(3.00, 4.79)	(3.13, 4.82)
$\theta = 2.22$	(0.11, 4.87)	(1.29, 3.68)	(1.41, 3.08)
Avg width	4.74	1.79	1.71
Avg Reduction	-	2.95	3.03

g			
True θ	Crude	Reference	Augmented
$\theta = 2.00$	(-4.61, 4.80)	(0.38, 4.80)	(0.35, 4.81)
$\theta = -2.55$	(-4.61, 4.79)	(-4.18, -0.50)	(-4.81, -0.11)
$\theta = -1.27$	(-4.77, 4.79)	(-3.27, 0.57)	(-4.48, 1.34)
$\theta = -4.60$	(-4.74, 4.73)	(-4.68, 3.53)	(-3.73, 4.09)
$\theta = 0.62$	(-4.77, 4.78)	(-4.35, 4.25)	(-4.51, 4.51)
$\theta = 4.08$	(-4.81, 4.79)	(-4.26, 4.62)	(-4.22, 4.88)
$\theta = -1.65$	(-4.66, 4.81)	(-4.28, 4.71)	(-3.82, 4.45)
$\theta = -4.05$	(-4.73, 4.79)	(-1.64, 1.84)	(-4.77, 0.83)
$\theta = -3.74$	(-4.66, 4.68)	(-4.90, 0.04)	(-4.91, -1.15)
$\theta = -0.94$	(-4.76, 4.75)	(-4.84, 3.90)	(-4.62, 3.72)
Avg width	9.48	6.38	6.69
Avg Reduction	-	3.11	2.80

k			
True θ	Crude	Reference	Augmented
$\theta = 0.50$	(-0.42, 4.76)	(-0.33, 1.09)	(-0.38, 1.20)
$\theta = 0.56$	(-0.41, 4.88)	(-0.29, 1.15)	(-0.44, 1.32)
$\theta = 0.58$	(-0.40, 4.81)	(-0.13, 1.40)	(-0.14, 1.38)
$\theta = 4.00$	(-0.28, 4.90)	(2.74, 4.54)	(2.78, 4.56)
$\theta = 1.13$	(-0.34, 4.78)	(0.33, 1.82)	(0.19, 1.80)
$\theta = 2.72$	(-0.39, 4.84)	(1.48, 3.57)	(1.81, 3.36)
$\theta = 3.63$	(-0.37, 4.74)	(2.19, 4.00)	(2.32, 4.08)
$\theta = 3.85$	(-0.40, 4.77)	(3.11, 4.20)	(3.14, 4.21)
$\theta = 0.46$	(-0.35, 4.88)	(-0.11, 1.44)	(-0.20, 1.32)
$\theta = 3.51$	(-0.36, 4.85)	(1.95, 3.98)	(2.47, 3.87)
Avg width	5.19	1.62	1.55
Avg Reduction	-	3.57	3.64

A1				A2			
True θ	Crude	Reference	Augmented	True θ	Crude	Reference	Augmented
$\theta = 3.00$	(0.16, 4.86)	(1.86, 3.99)	(1.60, 4.32)	$\theta = 4.00$	(0.14, 4.90)	(2.82, 4.94)	(2.52, 4.94)
$\theta = 2.95$	(0.13, 4.92)	(0.14, 4.37)	(0.17, 4.39)	$\theta = 1.48$	(0.13, 4.90)	(0.29, 4.66)	(0.18, 4.47)
$\theta = 2.19$	(0.17, 4.91)	(0.59, 4.65)	(0.37, 4.67)	$\theta = 2.17$	(0.13, 4.86)	(0.49, 4.62)	(0.13, 4.60)
$\theta = 0.82$	(0.12, 4.84)	(0.19, 3.86)	(0.18, 4.26)	$\theta = 1.71$	(0.13, 4.82)	(0.06, 3.31)	(0.09, 3.91)
$\theta = 4.16$	(0.11, 4.83)	(0.16, 4.55)	(0.19, 4.60)	$\theta = 3.64$	(0.13, 4.88)	(0.19, 4.53)	(0.24, 4.70)
$\theta = 3.70$	(0.14, 4.87)	(0.37, 4.75)	(0.20, 4.53)	$\theta = 3.21$	(0.13, 4.87)	(0.23, 4.72)	(0.20, 4.68)
$\theta = 2.06$	(0.14, 4.88)	(0.96, 4.81)	(0.82, 4.86)	$\theta = 1.75$	(0.08, 4.81)	(0.42, 4.50)	(0.39, 4.58)
$\theta = 2.36$	(0.11, 4.87)	(1.38, 4.87)	(1.52, 4.89)	$\theta = 1.11$	(0.20, 4.89)	(0.38, 4.32)	(0.21, 4.67)
$\theta = 4.81$	(0.15, 4.88)	(0.33, 4.88)	(0.17, 3.77)	$\theta = 1.33$	(0.13, 4.82)	(0.17, 4.41)	(0.55, 4.62)
$\theta = 1.83$	(0.13, 4.91)	(0.54, 4.91)	(1.12, 4.94)	$\theta = 2.06$	(0.14, 4.83)	(0.08, 4.37)	(0.17, 4.26)
$\theta = 4.58$	(0.07, 4.84)	(0.08, 4.15)	(0.34, 4.29)	$\theta = 3.52$	(0.14, 4.86)	(0.23, 4.60)	(0.26, 4.29)
$\theta = 4.94$	(0.16, 4.89)	(0.45, 2.93)	(0.15, 2.90)	$\theta = 0.28$	(0.16, 4.85)	(0.59, 2.99)	(0.37, 3.22)
$\theta = 2.33$	(0.13, 4.87)	(0.09, 3.90)	(0.07, 3.19)	$\theta = 0.94$	(0.14, 4.88)	(1.48, 4.91)	(1.49, 4.95)
$\theta = 4.37$	(0.09, 4.84)	(0.40, 4.64)	(0.48, 4.92)	$\theta = 1.59$	(0.12, 4.87)	(0.27, 4.62)	(0.30, 4.56)
$\theta = 0.84$	(0.09, 4.84)	(0.67, 4.83)	(0.76, 4.83)	$\theta = 4.95$	(0.12, 4.87)	(0.33, 4.09)	(0.45, 4.43)
$\theta = 0.14$	(0.13, 4.83)	(0.24, 4.72)	(0.16, 4.64)	$\theta = 4.38$	(0.12, 4.87)	(0.10, 4.28)	(0.10, 4.67)
$\theta = 2.89$	(0.20, 4.93)	(0.28, 4.76)	(0.25, 4.71)	$\theta = 4.80$	(0.12, 4.87)	(0.06, 3.88)	(0.08, 3.87)
$\theta = 3.21$	(0.11, 4.81)	(0.05, 3.32)	(0.05, 3.31)	$\theta = 2.81$	(0.17, 4.86)	(0.10, 3.87)	(0.08, 4.04)
$\theta = 3.85$	(0.12, 4.91)	(0.07, 3.96)	(0.09, 4.03)	$\theta = 0.67$	(0.18, 4.90)	(0.72, 4.89)	(0.80, 4.84)
$\theta = 2.25$	(0.17, 4.85)	(0.30, 4.39)	(0.34, 4.60)	$\theta = 1.59$	(0.11, 4.85)	(0.20, 4.13)	(0.12, 4.45)
Avg width	4.74	3.94	3.87	Avg width	4.73	3.92	3.97
Avg Reduction	-	0.80	0.86	Avg Reduction	-	0.81	0.76

B1				B2			
True θ	Crude	Reference	Augmented	True θ	Crude	Reference	Augmented
$\theta = 1.00$	(0.12, 4.81)	(0.27, 2.36)	(0.17, 2.63)	$\theta = 0.50$	(0.12, 4.92)	(0.01, 1.08)	(0.01, 1.56)
$\theta = 2.25$	(0.15, 4.88)	(0.29, 4.48)	(0.13, 4.51)	$\theta = 4.74$	(0.07, 4.82)	(0.03, 2.09)	(0.04, 2.18)
$\theta = 4.78$	(0.08, 4.84)	(0.17, 4.68)	(0.29, 4.63)	$\theta = 4.73$	(0.17, 4.91)	(0.49, 4.60)	(0.31, 4.50)
$\theta = 5.00$	(0.12, 4.91)	(0.19, 4.62)	(0.19, 3.96)	$\theta = 3.61$	(0.16, 4.91)	(0.07, 3.17)	(0.04, 3.32)
$\theta = 3.11$	(0.10, 4.88)	(1.58, 4.95)	(1.89, 4.95)	$\theta = 2.94$	(0.13, 4.82)	(0.62, 4.87)	(0.56, 4.69)
$\theta = 1.22$	(0.14, 4.91)	(0.61, 4.75)	(0.76, 4.69)	$\theta = 0.52$	(0.10, 4.91)	(1.27, 4.94)	(1.41, 4.96)
$\theta = 1.07$	(0.10, 4.87)	(0.02, 2.28)	(0.03, 3.39)	$\theta = 4.47$	(0.13, 4.90)	(2.47, 4.96)	(1.97, 4.95)
$\theta = 4.88$	(0.13, 4.89)	(0.64, 4.84)	(0.43, 4.01)	$\theta = 4.55$	(0.11, 4.92)	(1.88, 4.94)	(2.91, 4.98)
$\theta = 1.52$	(0.10, 4.74)	(0.67, 4.42)	(0.94, 4.83)	$\theta = 0.73$	(0.11, 4.90)	(1.08, 4.17)	(1.28, 4.42)
$\theta = 4.25$	(0.14, 4.90)	(0.03, 2.08)	(0.05, 2.13)	$\theta = 1.28$	(0.12, 4.89)	(0.05, 3.47)	(0.03, 2.53)
$\theta = 2.20$	(0.11, 4.87)	(0.18, 4.25)	(0.62, 4.90)	$\theta = 2.74$	(0.14, 4.87)	(1.46, 4.88)	(2.53, 4.96)
$\theta = 4.54$	(0.12, 4.89)	(1.44, 3.85)	(1.35, 3.65)	$\theta = 4.68$	(0.19, 4.86)	(3.52, 4.99)	(3.28, 4.99)
$\theta = 3.33$	(0.10, 4.88)	(0.24, 3.90)	(0.13, 3.98)	$\theta = 0.08$	(0.15, 4.90)	(2.94, 4.98)	(2.73, 4.98)
$\theta = 3.05$	(0.13, 4.83)	(0.19, 4.11)	(0.07, 3.89)	$\theta = 2.33$	(0.17, 4.88)	(0.01, 1.32)	(0.01, 1.01)
$\theta = 1.44$	(0.11, 4.87)	(2.23, 4.96)	(2.11, 4.95)	$\theta = 4.90$	(0.10, 4.88)	(0.22, 4.51)	(0.34, 4.39)
$\theta = 0.35$	(0.15, 4.86)	(0.05, 3.38)	(0.04, 2.83)	$\theta = 4.54$	(0.17, 4.90)	(0.30, 4.71)	(0.32, 4.65)
$\theta = 3.34$	(0.13, 4.91)	(0.10, 3.97)	(0.07, 3.83)	$\theta = 0.91$	(0.11, 4.85)	(0.49, 4.28)	(0.62, 4.71)
$\theta = 4.15$	(0.12, 4.86)	(0.10, 3.51)	(0.27, 3.85)	$\theta = 4.43$	(0.16, 4.85)	(1.82, 4.81)	(1.91, 4.91)
$\theta = 3.34$	(0.09, 4.82)	(0.73, 4.81)	(0.64, 4.78)	$\theta = 4.96$	(0.18, 4.91)	(1.73, 4.91)	(1.22, 4.91)
$\theta = 0.52$	(0.23, 4.92)	(0.40, 4.36)	(0.27, 4.18)	$\theta = 0.66$	(0.16, 4.89)	(1.09, 4.63)	(1.30, 4.75)
Avg width	4.74	3.57	3.44	Avg width	4.74	3.04	2.96
Avg Reduction	-	1.17	1.30	Avg Reduction	-	1.71	1.78

g1			
True θ	Crude	Reference	Augmented
$\theta = 1.00$	(-4.63, 4.79)	(-2.48, 4.45)	(-3.56, 4.87)
$\theta = 1.45$	(-4.75, 4.71)	(-2.91, 4.86)	(-4.68, 4.91)
$\theta = -3.29$	(-4.77, 4.75)	(-4.02, 4.02)	(-4.74, 3.57)
$\theta = -4.32$	(-4.78, 4.72)	(-3.96, 4.45)	(-4.42, 4.61)
$\theta = 2.65$	(-4.78, 4.70)	(-4.54, 3.69)	(-4.70, 3.72)
$\theta = 3.00$	(-4.81, 4.67)	(-4.62, 3.31)	(-4.34, 4.24)
$\theta = 4.06$	(-4.75, 4.80)	(-4.67, 3.58)	(-4.60, 3.80)
$\theta = 4.08$	(-4.72, 4.77)	(-3.81, 4.42)	(-4.30, 4.33)
$\theta = 0.68$	(-4.73, 4.75)	(-4.67, 3.57)	(-4.82, -0.03)
$\theta = -3.76$	(-4.74, 4.77)	(-4.68, 4.75)	(-4.71, 4.80)
$\theta = -3.21$	(-4.76, 4.64)	(-3.51, 4.40)	(-4.79, 3.31)
$\theta = 4.56$	(-4.72, 4.72)	(-4.58, 1.38)	(-4.79, 1.67)
$\theta = 4.06$	(-4.77, 4.71)	(-4.66, 4.39)	(-4.70, 2.88)
$\theta = -1.32$	(-4.82, 4.74)	(-4.73, 4.55)	(-4.60, 4.50)
$\theta = 1.40$	(-4.74, 4.64)	(-4.32, 3.79)	(-4.26, 4.68)
$\theta = 0.93$	(-4.77, 4.66)	(-4.80, 3.29)	(-4.89, 4.09)
$\theta = 3.20$	(-4.79, 4.71)	(-4.84, 2.51)	(-4.87, 4.61)
$\theta = 4.54$	(-4.68, 4.76)	(-3.04, 4.70)	(-4.04, 4.85)
$\theta = 2.87$	(-4.87, 4.69)	(-4.75, 2.01)	(-4.79, 2.79)
$\theta = -4.35$	(-4.82, 4.78)	(-4.52, 4.58)	(-4.68, 4.51)
Avg width	9.48	8.00	8.29
Avg Reduction	-	1.48	1.20

g2			
True θ	Crude	Reference	Augmented
$\theta = 2.00$	(-4.75, 4.71)	(-3.37, 4.75)	(-3.30, 4.78)
$\theta = 2.01$	(-4.76, 4.80)	(-3.73, 4.87)	(-4.56, 4.93)
$\theta = 4.47$	(-4.70, 4.78)	(-4.09, 4.71)	(-3.11, 4.84)
$\theta = -3.55$	(-4.76, 4.74)	(-3.96, 4.42)	(-3.58, 4.59)
$\theta = -4.82$	(-4.73, 4.76)	(-4.27, 4.62)	(-3.65, 4.74)
$\theta = -1.26$	(-4.76, 4.75)	(-3.75, 4.14)	(-4.39, 4.05)
$\theta = 4.07$	(-4.73, 4.68)	(-4.53, 4.45)	(-3.44, 4.79)
$\theta = -3.80$	(-4.76, 4.74)	(-4.49, 3.10)	(-4.79, 4.12)
$\theta = -2.27$	(-4.76, 4.60)	(-3.45, 3.79)	(-0.95, 4.81)
$\theta = 1.12$	(-4.69, 4.81)	(-4.80, 4.23)	(-4.90, 3.96)
$\theta = -3.98$	(-4.76, 4.64)	(-4.69, 3.06)	(-4.41, 4.43)
$\theta = 2.21$	(-4.75, 4.74)	(-1.01, 4.66)	(-0.37, 4.80)
$\theta = -0.65$	(-4.78, 4.70)	(-4.43, 4.76)	(-4.14, 4.34)
$\theta = -1.66$	(-4.74, 4.61)	(-4.55, 4.71)	(-4.78, 4.81)
$\theta = 4.86$	(-4.72, 4.66)	(-3.31, 4.86)	(-1.17, 4.89)
$\theta = -1.03$	(-4.76, 4.66)	(-4.03, 4.87)	(-4.48, 4.90)
$\theta = 4.53$	(-4.81, 4.78)	(-4.74, 4.70)	(-4.85, 4.40)
$\theta = 0.76$	(-4.72, 4.70)	(-4.17, 4.45)	(-4.77, 3.64)
$\theta = -4.36$	(-4.60, 4.78)	(-0.07, 4.89)	(-0.77, 4.84)
$\theta = -4.64$	(-4.77, 4.83)	(-4.52, 4.39)	(-4.55, 4.75)
Avg width	9.47	8.26	8.09
Avg Reduction	-	1.21	1.38

k1				k2			
True θ	Crude	Reference	Augmented	True θ	Crude	Reference	Augmented
$\theta = 0.50$	(-0.35, 4.87)	(-0.48, 0.73)	(-0.48, 0.92)	$\theta = 0.40$	(-0.39, 4.87)	(-0.47, 1.06)	(-0.48, 0.98)
$\theta = 4.11$	(-0.39, 4.83)	(-0.45, 1.91)	(-0.46, 2.26)	$\theta = 4.23$	(-0.32, 4.78)	(1.95, 4.96)	(2.62, 4.98)
$\theta = 2.57$	(-0.32, 4.90)	(-0.47, 2.02)	(-0.47, 2.12)	$\theta = 1.15$	(-0.31, 4.89)	(0.01, 4.26)	(-0.14, 4.48)
$\theta = 3.30$	(-0.37, 4.86)	(-0.47, 1.82)	(-0.45, 2.01)	$\theta = 1.35$	(-0.37, 4.84)	(-0.05, 3.28)	(-0.09, 3.57)
$\theta = 0.90$	(-0.40, 4.83)	(0.34, 3.39)	(0.38, 3.76)	$\theta = 1.25$	(-0.37, 4.83)	(-0.46, 2.54)	(-0.41, 2.68)
$\theta = 3.84$	(-0.31, 4.88)	(1.17, 4.82)	(1.19, 4.85)	$\theta = 2.63$	(-0.37, 4.82)	(-0.08, 3.96)	(0.04, 3.74)
$\theta = 0.17$	(-0.40, 4.80)	(-0.47, 1.51)	(-0.48, 1.57)	$\theta = 3.48$	(-0.39, 4.86)	(-0.08, 2.21)	(-0.23, 2.53)
$\theta = -0.40$	(-0.37, 4.78)	(-0.47, 1.61)	(-0.44, 2.09)	$\theta = 2.71$	(-0.25, 4.90)	(0.32, 2.70)	(-0.02, 2.28)
$\theta = 2.34$	(-0.36, 4.85)	(-0.47, 2.06)	(-0.48, 1.55)	$\theta = 1.74$	(-0.31, 4.85)	(2.74, 4.97)	(2.94, 4.98)
$\theta = 0.66$	(-0.35, 4.88)	(3.13, 4.99)	(3.75, 4.99)	$\theta = 1.66$	(-0.35, 4.88)	(-0.18, 3.56)	(0.29, 3.84)
$\theta = 0.38$	(-0.38, 4.90)	(-0.45, 2.91)	(-0.47, 1.83)	$\theta = 1.37$	(-0.35, 4.88)	(1.73, 4.95)	(2.84, 4.97)
$\theta = -0.41$	(-0.35, 4.88)	(-0.49, 0.44)	(-0.49, 0.57)	$\theta = 2.78$	(-0.38, 4.83)	(-0.50, 0.20)	(-0.49, 0.31)
$\theta = 1.65$	(-0.40, 4.75)	(-0.48, 1.49)	(-0.47, 1.70)	$\theta = 3.93$	(-0.39, 4.80)	(-0.19, 2.01)	(-0.20, 2.15)
$\theta = 4.10$	(-0.34, 4.87)	(-0.48, 1.20)	(-0.48, 1.25)	$\theta = -0.33$	(-0.31, 4.93)	(2.10, 4.97)	(3.01, 4.97)
$\theta = 2.90$	(-0.30, 4.91)	(-0.40, 1.77)	(-0.43, 1.82)	$\theta = 0.39$	(-0.35, 4.85)	(-0.45, 2.03)	(-0.46, 1.81)
$\theta = 2.75$	(-0.37, 4.88)	(-0.01, 4.75)	(0.77, 4.85)	$\theta = 4.65$	(-0.37, 4.88)	(-0.42, 2.74)	(-0.44, 2.62)
$\theta = 2.88$	(-0.41, 4.81)	(-0.31, 2.79)	(-0.27, 3.10)	$\theta = 3.06$	(-0.34, 4.89)	(-0.41, 2.18)	(-0.46, 1.88)
$\theta = 2.64$	(-0.39, 4.84)	(-0.48, 1.41)	(-0.47, 1.37)	$\theta = -0.01$	(-0.36, 4.87)	(-0.27, 2.33)	(-0.23, 2.14)
$\theta = 1.69$	(-0.40, 4.87)	(-0.34, 2.80)	(-0.39, 2.84)	$\theta = 4.17$	(-0.32, 4.89)	(-0.04, 3.38)	(-0.22, 3.38)
$\theta = 0.24$	(-0.38, 4.83)	(-0.47, 1.65)	(-0.48, 1.68)	$\theta = 0.43$	(-0.38, 4.89)	(-0.02, 2.78)	(-0.17, 2.88)
Avg width	5.22	2.57	2.39	Avg width	5.21	2.81	2.62
Avg Reduction	-	2.65	2.83	Avg Reduction	-	2.41	2.59

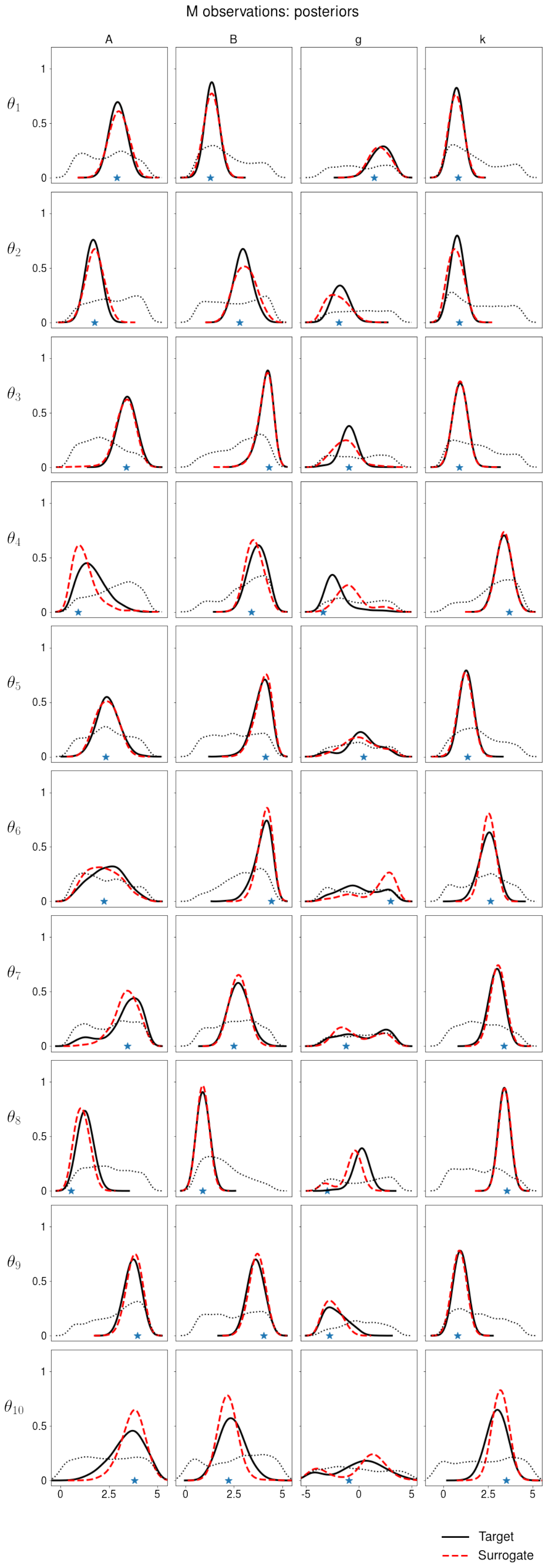


Figure A.1: Marginal posteriors for each of the eight parameters. True parameter values are indicated with stars.

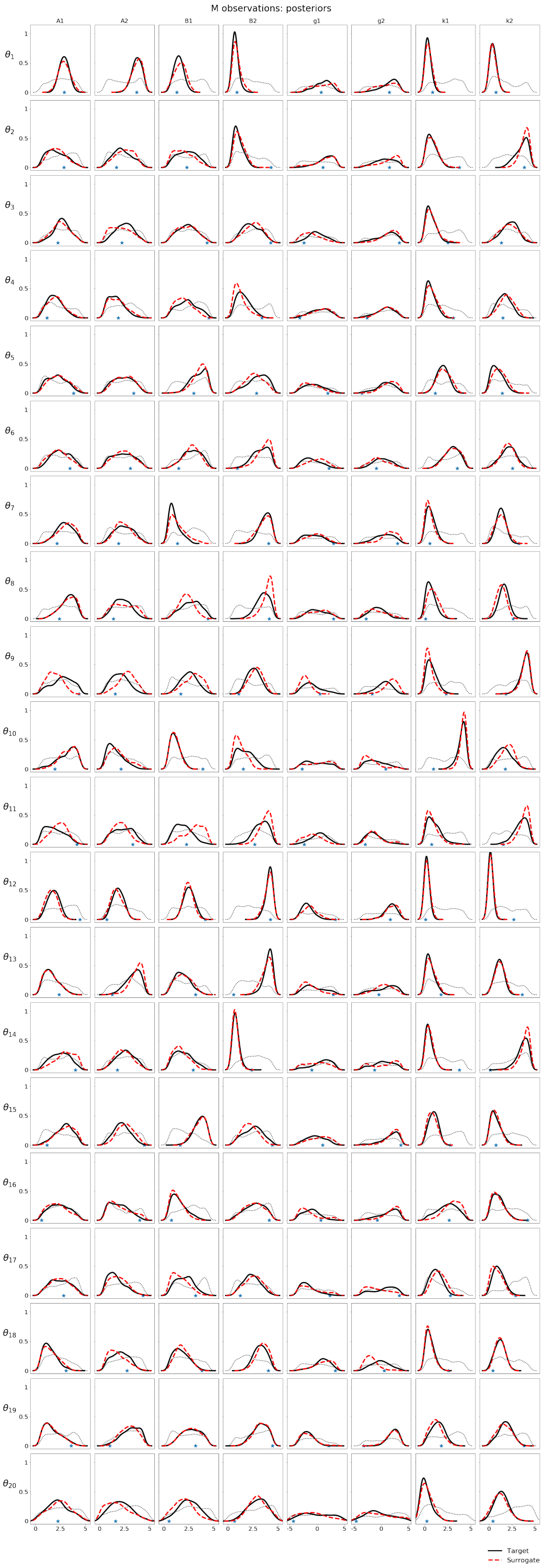


Figure A.2: Marginal posteriors for each of the eight parameters. True parameter values are indicated with stars.