ORIGINAL PAPER

Posterior SBC: simulation-based calibration checking conditional on data

Teemu Säilynoja[1] · Marvin Schmitt[2] · Paul-Christian Bürkner[3] · Aki Vehtari[1]

Abstract

Simulation-based calibration checking (SBC) refers to the validation of an inference algorithm and model implementation through testing self-consistency with repeated inference on data simulated from a generative model. In the original and commonly used approach, the generative model uses parameters drawn from the prior, and thus the approach is testing whether the inference works for simulated data generated with parameter values plausible under that prior. This approach is natural and desirable when we want to test whether the inference works for a wide range of datasets we might observe. However, after observing data, we are interested in answering whether the inference works *conditional on that particular data*. In this paper, we propose *posterior SBC* and demonstrate how it can be used to validate the inference conditionally on observed data. We illustrate the utility of posterior SBC in three case studies: (1) A simple multilevel model; (2) a model that is governed by differential equations; and (3) a joint integrative neuroscience model which is approximated via amortized Bayesian inference with neural networks.

Keywords Calibration · Diagnostics · Model evaluation · Bayesian inference

1 Introduction

To ensure trust on the results of the inference, different forms of calibration checking play important roles in the Bayesian workflow for model building (Gelman et al. 2020). Simulation-based calibration checking (SBC; (Cook et al. 2006; Talts et al. 2020; Modrák et al. 2023)), is one of these methods and validates the chosen posterior inference algorithm as well as checks for inconsistencies between the model implementation and a possibly separate implementation of the data generating process. Probabilistic programming software (Štrumbelj et al. 2024) provide probabilistic programming languages to make it easier to implement probabilistic models, and inference engines that implement various inference algorithms. As the probabilistic programming software and algorithms for probabilistic modelling advance, the threshold for specifying increasingly complicated models gets lowered. However, as the complexity of the models increases, so does the probability of human error when writing the code, which implements the model. In addition to such model code implementation mistakes, the inference algorithms used for obtaining posterior approximations—even when correctly coded—can work well for one posterior, but suffer from computational difficulties for another. By simply changing the observed data, one may obtain vastly different posterior geometries, favoring one implementation or algorithm over another.

SBC aims to check that the inference is calibrated, but miscalibration can be caused either a mistake in the model implementation code, a mistake in inference algorithm implementation code, or significant bias in the inference algorithm. The algorithm is considered to be calibrated, if the probability integral transformation (PIT) of the parame-

✉ Teemu Säilynoja
teemu.sailynoja@gmail.com

Marvin Schmitt
mail.marvinschmitt@gmail.com

Paul-Christian Bürkner
paul.buerkner@gmail.com

Aki Vehtari
aki.vehtari@aalto.fi

[1] Department of Computer Science, Aalto University, Espoo, Finland

[2] Cluster of Excellence SimTech, University of Stuttgart, Stuttgart, Germany

[3] Department of statistics, TU Dortmund University, Dortmund, Germany

ters is uniformly distributed, when transformed with respect to the posterior conditional on data generated using those parameters. We review SBC and the definition of calibration in more detail in Section 2.

In this paper we assume that the both the code for implementing the model, and the inference algorithm implementation are correct and focus on examples were the inference algorithm may produce significantly biased results with finite computation time. Even when the inference algorithm is implemented correctly, it may produce significantly biased inference for some posteriors. For example, many Markov chain Monte Carlo (MCMC) and variational inference approaches have challenges with funnel shaped posteriors (see, e.g. Neal 2003; Yao et al. 2018,), which are common in case of hierarchical models. The original SBC is limited to checking for calibration on the whole joint distribution of parameters and observations under the specified priors, and thus we call it *prior SBC*. Specifying prior distributions that accurately present the available prior information is often difficult. In some cases, the prior may have significant probability mass in parts of the parameter space that generate such data for which the inference algorithm produce significantly biased inference (in finite time). For example, the data generated from a hierarchical model with certain type of parameter values can lead to a funnel shaped posterior which has a challenging geometry for many inference algorithms to explore. Figure 1 illustrates two cases: a) in which the prior space includes two subspaces with canceling biases and SBC indicates that the calibration over the prior is good, and b) in which the prior space includes one subspace with biased inference causing SBC to indicate that the calibration over the prior is not good. However, after observing some

data we are not interested in calibration over the prior, but calibration within the region of the posterior.

In this paper, we introduce *posterior SBC*, a variant of SBC using the self-consistency of Bayesian posterior distribution to validate the model implementation and inference algorithm when conditioned on some fixed set of data. Figure 1 illustrates two cases where, if the posterior is much more concentrated than the prior, posterior SBC checks the calibration of the inference for those regions of the parameter space where it matters and does not care how well the inference algorithm works elsewhere. Prior SBC is useful for developers of software, algorithms, models, and priors who need their work to be applicable to a wide range of datasets, while posterior SBC is useful for modelers working on a specific dataset. As the number of modelers is several orders of magnitude bigger than the number of developers, we expect the posterior SBC to become more popular than prior SBC.

We next review the method of simulation-based calibration as it has traditionally been operationalized. In Section 3, we introduce posterior SBC. In Section 4 we illustrate the differences between prior and posterior SBC through three case studies including a simple hierarchical model, Lotka-Volterra model, and a drift-diffusion model. In the first two examples we use MCMC and in the last example we use amortized Bayesian inference for which there are no similar convergence diagnostics as for MCMC. Finally, in Section 5 we summarize the contributions of the paper and provide an outlook for future research.


(a) Two canceling biases   (b) One problematic region
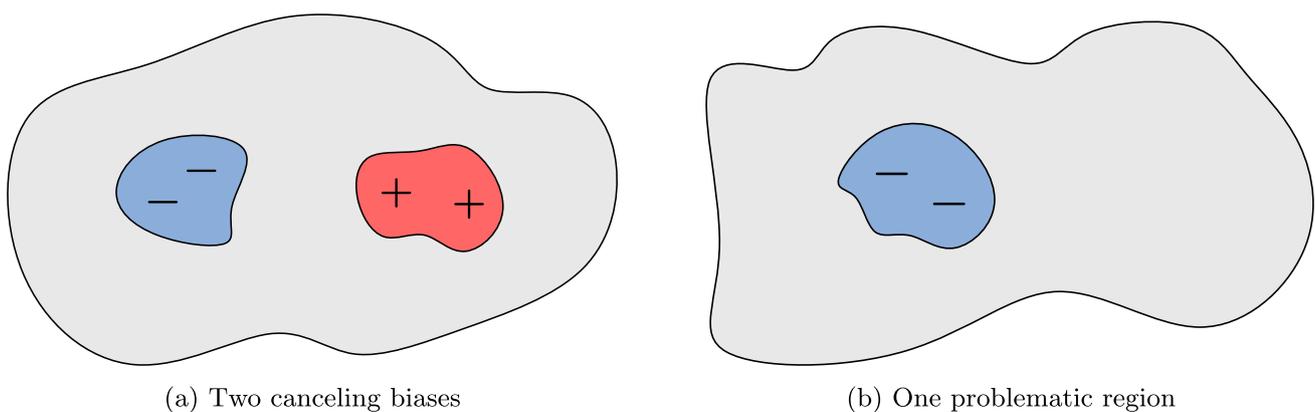
Fig. 1 Two conceptual illustrations of a model parameter space, with gray area denoting the prior. In (a), the colored regions are potential posteriors with bias in opposite directions, while the inference is well calibrated for parameter values outside these regions. Prior SBC will not show calibration issues due to cancellation of biases, while posterior SBC would indicate issues for posteriors intersecting the colored regions. In (b), only one region is problematic. Now prior SBC would indicate calibration issues, while posterior SBC would show that the inference is calibrated outside the colored region. In both cases posterior SBC has the benefit of focusing on the region, which matters given the observed dataset

2 Simulation-based calibration checking

Let us consider a Bayesian model of the joint distribution of the data y and parameters θ,

$$\pi(y, \theta) = \pi(y \mid \theta)\pi(\theta),$$

where $\pi(\theta)$ is the prior distribution of the parameters and $\pi(y \mid \theta)$ the likelihood of the data.

Cook et al. (2006) propose a simulation-based calibration checking method for validating Bayesian inference software designed to fit the model. The key to this method is the following self-consistency property of Bayesian posterior distributions. Let $\theta' \sim \pi(\theta)$, and $y' \sim \pi(y \mid \theta')$. Now the pair (y', θ') represents a draw from the joint distribution $\pi(y, \theta)$, and therefore θ' is a draw from the posterior distribution, $\pi(\theta \mid y')$. This self-consistency property, linking the prior, prior predictive, and posterior distributions, can be summarized with the following *SBC equality*:

$$
\begin{aligned}
\pi(y', \theta', \theta'') &= \pi(\theta'' \mid y')\pi(y' \mid \theta')\pi(\theta') \\
&= \pi(\theta'' \mid y')\pi(\theta' \mid y')\pi(y'),
\end{aligned}
\tag{1}
$$

where $\theta'' \sim \pi(\theta \mid y')$ is a posterior draw conditioned on y'.

Given an observation y' and the model implementation, the inference algorithm should be able to sample from the posterior, $\theta'' \sim \pi(\theta \mid y')$. Therefore, to verify that the inference algorithm is able to accurately sample from the posterior, we should assess whether the posterior draw θ'' and the prior draw θ', conditional on $y' \sim \pi(y \mid \theta')$, are from the same distribution.

To assess whether θ' and θ'' are from the same distribution conditional on y, Cook et al. (2006) repeatedly draw parameter vectors θ'_i from the prior $\pi(\theta)$, generate data from the observation model $y_i \sim \pi(y_i \mid \theta'_i)$, and then use the algorithm to be validated to sample $\theta''_{i,1}, \ldots, \theta''_{i,S} \sim \pi(\theta'' \mid y_i)$. The authors then propose to compute empirical probability integral transform (PIT) values,

$$u_i = p(\theta''_i < \theta'_i \mid y_i), \tag{2}$$

which prove to be uniformly distributed as $S \to \infty$. The process is repeated for $i = 1, \ldots, N$ and the empirical PIT values are used for testing. Figure 2 shows three iterations of prior draws and their respective PIT relative to the corresponding posterior distributions. Cook et al. (2006) further propose the use of the χ^2-test for the inverse of the normal cumulative distribution function (CDF) of the empirical PIT values. The inference is said to be calibrated if the PIT values pass a uniformity test.

Cook et al. (2006) did not take into account that the estimated PIT values are discrete when computed with finite sample size S, and how the possible correlation in Markov chains affects the results (Gelman 2017; Talts et al. 2020). Talts et al. (2020) propose testing for the discrete uniformity of the empirical PIT values (see also Modrák et al. 2023, forimprovedtheoryandproofs). By thinning $\theta''_{i,1}, \ldots, \theta''_{i,S}$ to remove possible autocorrelation, which can affect the estimated empirical PIT values (Säilynoja et al. 2022), the uniformity of these discrete empirical PIT values can be tested with the approach presented in Säilynoja et al. (2022).

Instead of only assessing the uniformity of the ranks of individual parameters, Modrák et al. (2023, Sections3.3–3.4) propose the use of the joint log-likelihood $\log p(y \mid \theta)$ as a test statistic for assessing the calibration. That is, $p(y \mid \theta')$ and $p(y \mid \theta'')$ are used instead of θ' and θ'' when computing PIT values with (2). This has the benefit of being a joint function of both data and all parameters. Using the data in the test quantity allows detecting calibration issues rising from the model partially ignoring the data. Modrák et al. (2023, Sections3.4and4.3) demonstrate how test quantities based on only parameters can not detect if the model completely ignores all data (the posterior is the prior) or part of the data (e.g. the first data point), but using the log-likelihood as a test quantity indicates the problem. Another advantage of using the log-likelihood as a test quantity is in checking calibration of models with high-dimensional parameter spaces, where modelers might not be particularly interested in gauging the calibration of all the individual parameters, and using the log-likelihood as a test quantity is a natural choice for a joint function of model parameters. Using the joint log-likelihood also avoids the need of correcting for multiple comparisons from simultaneously assessing the calibration of many individual parameters.

Algorithm 1 Prior SBC checking

Given: Inference algorithm $\mathcal{A}$, and a model implementing $\pi(y, \theta)$.
Sample $\theta'_1, \ldots, \theta'_N \sim \pi(\theta)$
for $i = 1$ to N **do**
    Draw $y_i \sim \pi(y \mid \theta'_i)$
    Apply $\mathcal{A}$ (and thinning) to obtain:
    $\theta''_1, \ldots, \theta''_S \sim \pi(\theta \mid y_i)$
    $u_i \leftarrow p(\theta''_i < \theta'_i \mid y_i)$
end for
Test if $u_1, \ldots, u_S$ uniformly distributed.

Performing SBC on the full prior parameter space has three fundamental drawbacks. First, as shown in the case study of Section 4.1, it is possible that only some parameter values generate such data that lead to a posterior from which the inference method is not able to sample faithfully. If such parameter values occur in a small region of the parameter space, and we average over a much wider prior, the effect in the calibration checking can be minor and prior SBC may miss the miscalibration completely. If such parameter values

Fig. 2 An illustration of how PIT values are computed in prior SBC. The model is a normal distribution $\mathcal{N}(\theta, \sigma)$ with a known standard deviation σ. The dashed line shows the prior distribution $p(\theta)$, and solid lines show three posteriors $p(\theta | y_i)$ conditioning on the predictive draws $y_i \sim \pi(y_i | \theta_i')$. The PIT value of each prior draw θ_i', shown as blue dots, equals the respective shaded blue area (0.03, 0.43, and 0.97 respectively). In practice the PIT values (2) are computed using Monte Carlo draws from the posteriors $\theta_{i,1}'', \ldots, \theta_{i,S}'' \sim \pi(\theta'' | y_i)$
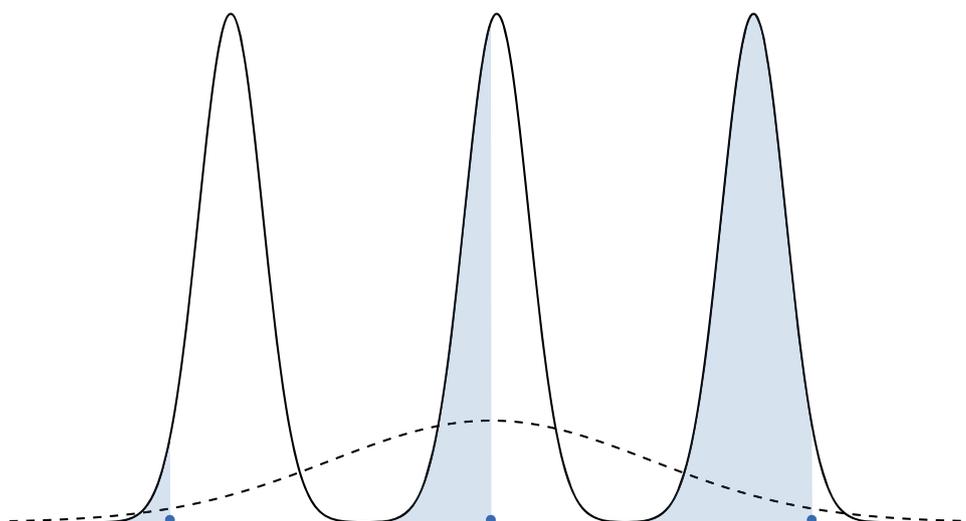
have high probability mass under the posterior, then posterior SBC shows the miscalibration.

Second, the time required for running the inference on each SBC iteration is often non-trivial and one needs to consider the tradeoff between their computational budget and the power of the calibration checking. Third, as visualized in Figure 1, even with unlimited computational resources the result of prior SBC does not guarantee calibration within a specific subregion of the inspected parameter space. There can always be another region with an equally strong but reverse effect on the overall distribution of the PIT values. As proposed by Modrák et al. (2023), one solution to this last problem is to employ additional test quantities using functions of parameters and non-monotonic transformations (such as folded rank statistics), that might detect these problematic regions, but coming up with these test quantities can be difficult.

3 Posterior simulation-based calibration checking – conditioned on observed data

Weakly informative priors are commonly used in Bayesian inference due to their desirable properties in real-life applications (Gelman et al. 2017). However, this class of priors can cause serious issues for interpreting the results of prior SBC. Even though a large portion of the prior mass may lie on a region of the parameter space where the inference works well, smaller regions may exist where the inference method can fail to faithfully present the posterior. On the one hand, if large enough, these issues may be detected through prior SBC. The modeler could then try to modify the priors or the model to reduce the prior probability of the problematic parameter values. On the other hand, if one already has access to the data that the model will later be used to run inference on, the assessment of the inference algorithm can

be focused on the region of the parameter space around the posterior. For inference on the given data, this is arguably the most important region to ensure proper calibration for.

Using the sequential Bayesian updating rule, one can view the posterior as the new prior and perform the entire SBC procedure conditional on the observed data y_{obs}. In other words, by expanding the joint distribution in (1) to include both the observed data and the predictive draw,

$$\pi(y_{\text{obs}}, y, \theta', \theta'') = \begin{aligned} &\pi(\theta' | y_{\text{obs}})\pi(y | \theta', y_{\text{obs}}) \\ &\times \pi(\theta'' | y, y_{\text{obs}}), \end{aligned} \quad (3)$$

we see that a sample from the original posterior, $\theta_i' \sim \pi(\theta | y_{\text{obs}})$, and $\theta'' \sim \pi(\theta | y_i, y_{obs})$ drawn from an augmented posterior, need to have the same distribution given the prediction $y_i \sim \pi(y | \theta_i')$. The identity (3) applies to Bayesian updating, and holds even if the model is misspecified. Thus, the reliability of inference can be diagnosed also for misspecified models. Thus, we operationalize the approach for data conditioned SBC by drawing $\theta_i' \sim \pi(\theta | y_{\text{obs}})$, generating posterior predictions $y_i \sim \pi(y | \theta_i')$, and then using the inference algorithm to be validated to draw from the new posterior $\theta_1'', \ldots \theta_S'' \sim \pi(\theta | y_i, y_{\text{obs}})$.

The calibration of inference can be checked by testing the uniformity of the PIT values, $u_i = p(\theta_i'' < \theta_i' | y_i, y_{obs})$, this time computed for the original posterior draws, θ_i', with regard to the (thinned) augmented posterior draws, $\theta_{i,1}'', \ldots, \theta_{i,S}''$. We call this variant of SBC using posterior draws *posterior SBC*.

When using prior SBC for validating the inference algorithm, drawing parameter values generatively from the prior is usually assumed to be trivial. In posterior SBC, we need to sample from two different posteriors, which would usually be achieved using the same algorithm that is being evaluated. Posterior SBC does not require that we are able to get draws

---

Algorithm 2 Posterior SBC checking

Given: Observation y_{obs}, inference algorithm $\mathcal{A}$, model implementing $\pi(y, \theta)$.
Sample $\theta'_1, \ldots, \theta'_N \sim \pi(\theta \mid y_{\mathrm{obs}})$
for $i = 1$ to N **do**
   Draw $y_i \sim \pi(y \mid \theta'_i)$
   Apply $\mathcal{A}$ to obtain:
     $\theta''_1, \ldots, \theta''_S \sim \pi(\theta \mid y_i, y_{\mathrm{obs}})$
     $u_i \leftarrow p(\theta''_i < \theta'_i \mid y_i, y_{obs})$
end for
Test if $u_1, \ldots, u_S$ uniformly distributed.

---

from the true posterior. If the sampling fails for either the original posterior or for the augmented posteriors, the SBC equality in Equation 3 does not hold, leading to non-uniform PIT values in posterior SBC. While the Bayesian updating is consistent when more data is observed, it is very unlikely that we would encounter such biased inference that would be consistent when conditioned on more data.

The three drawbacks listed in Section 2 for prior SBC are greatly reduced, but still present for posterior SBC. First, it is still possible, that there is a problematic region of the posterior with a small probability mass, that would not be detected with posterior SBC. However, unlike with prior SBC, passing posterior SBC would indicate that the inference is very likely to work for the inspected dataset. Second, posterior SBC can still be computationally intensive as one need to run multiple posterior inferences, but focusing on the relevant parameter values ensures that these computational resources are better focused for assessing whether the inference works for the observed data. Thirdly, it is still possible to have regions of the posterior with canceling biases, but as with the first drawback, the relative size of these to the posterior is greatly reduced.

Posterior SBC looks similar to some other methods using draws from the posterior predictive distribution. Posterior predictive checking (PPC; Box 1980; Rubin 1984; Gelman et al. 1996, 2013; Gabry et al. 2019,) compares draws from the posterior predictive distribution to the original data. PPC uses the data twice which can lead to optimistic checking unless the test statistic is completely ancillary which is unlikely especially for more flexible models. Still, we expect predictive checking or other forms of model checking to assess model misspecification to be a part of modeling workflows (Gelman et al. 2020). Parameter recovery experiments generate data from the model using some feasible parameter values, and the posterior is compared to those parameter values visually without a formal test. The crucial part of posterior SBC is to use the chain rule and compare the draws from the original posterior to the augmented posterior which is conditioned on both the original data and the posterior predictive draws.

A reader may wonder why do we need posterior SBC, when sampling from the original posterior with Markov chain Monte Carlo can be diagnosed with convergence diagnostics such as $\widehat{R}$ (Vehtari et al. 2021) and divergences (Betancourt 2018). Vehtari et al. (2021, AppendixC) demonstrate how $\widehat{R}$ fails to diagnose bias in the case of a funnel shaped posterior and very long chains. While in that example Hamiltonian Monte Carlo specific divergence diagnostic did work, it has also been observed to miss convergence issues. Furthermore, additional diagnostics similar to $\widehat{R}$ and the divergence diagnostic are not available for all MCMC methods. What is more, while convergence diagnostics can indicate issues, they do not indicate the direction or magnitude of the induced bias. For example, it is known that there are also false positive divergence warnings, and in such cases posterior SBC can be used to measure the amount of potential bias. While posterior SBC is operationalized by sampling from the posterior, the inference algorithms tested do not need to be (Markov chain) Monte Carlo methods. Any inference algorithm can be tested as long as we can also get draws from the approximated posterior. (Yao et al. 2018) demonstrate the challenges of diagnosing the reliability of variational inference in high dimensions and use prior SBC to check variational inference by drawing from the variational approximation. Schad et al. (2023) discuss problems in estimating Bayes factors and use prior SBC for checking computation of Bayes factors with bridge sampling and Savage-Dickey method. Considering Yao et al. (2018) and Schad et al. (2023) did consider reliability of computation given the observed data, it would have been sensible to use posterior SBC, instead. In Section 4.3 we illustrate the use posterior SBC for amortized Bayesian inference, for which before this there were no practical inference checking tools.

Finally, as with prior SBC, posterior SBC is highly parallelizable as after the initial posterior inference, the subsequent inferences on the augmented datasets can be performed independently, and one only needs to consider the results jointly when testing for uniformity of the PIT values.

4 Case studies

In this section, we present three case-studies demonstrating the differences of prior and posterior SBC. We demonstrate how posterior SBC is better suited for calibration assessment when the modeler is particularly interested in the trustworthiness of the inference with a given dataset. In this case, Prior SBC may be too computationally ineffective to detect calibration issues, or alert of issues that are not present after conditioning on the data.

In Section 4.1, we show an example where both of two alternative model implementations exhibit miscalibration. However, the problematic region is relatively small. Prior SBC does not imply calibration issues even with a large num-

ber of iterations, while posterior SBC is able to detect the issue.

In Section 4.2, we show an example where prior SBC indicates calibration issues with data plausible under prior beliefs of the parameter values. After conditioning the model on an observed dataset, these problems are no longer present. Additionally, in this example, the challenges caused by the chosen diffuse priors make the inference computationally heavy, and thus would make it impractical to iteratively improve the model.

Lastly, in Section 4.3, we illustrate the potential of posterior SBC in amortized Bayesian workflows. There, it can be employed as the default data-conditional diagnostic for assessing the trustworthiness of the learned neural posterior approximator with negligible computational overhead.

4.1 Simple hierarchical model – choice of model parameterization

This case study focuses on a situation where the best choice of model implementation to obtain trustworthy inference depends on the observed data. Hierarchical models often exhibit funnel shaped posterior geometries that are difficult to sample (Papaspiliopoulos et al. 2007). Depending on the observed data, the likelihood can contribute to the posterior in vastly different ways, making it near-impossible to choose a model parameterization before observing data.

Setup We study a simple hierarchical model for a dataset with 50 groups each producing five observations,

$$y_{i,j} \sim \mathcal{N}(\mu_j, \sigma^2),$$
$$\mu_j \sim \mathcal{N}(\mu_0, \tau^2),$$
$$\mu_0 \sim \mathcal{N}(0, 1),$$
$$\sigma, \tau \sim \mathcal{N}^+(0, 1),$$

where $j \in \{1, \dots, 50\}$, and $i \in \{1, \dots, 5\}$.

The standard normal prior for the population level mean μ_0, as well as the truncated standard normal priors for both the population level variance, τ, and the within group variance, σ, quantify relatively large prior uncertainty on both the similarity between the groups, and the similarity of the observations within any given group.

We first summarize the results of using prior SBC to investigate the calibration of the posterior inference with two alternative parametrizations of the joint likelihood: the centered parametrization shown above, and the mathematically equivalent non-centered parametrization:

$$y_{i,j} \sim \mathcal{N}(\mu_j, \sigma^2),$$
$$\mu_j = \mu_0 + \tau z_j,$$
$$z_j \sim \mathcal{N}(0, 1),$$

$$\mu_0 \sim \mathcal{N}(0, 1),$$
$$\sigma, \tau \sim \mathcal{N}^+(0, 1).$$

The non-centered parameterization exhibits a more convenient posterior geometry in case of a weakly informative likelihood.

For posterior SBC, we assess the calibration of posterior inference conditional on two datasets, the first generated with $\tau = 0.06$ and $\sigma = 1.96$, and the second with $\tau = 1.96$ and $\sigma = 0.06$. These correspond to the 5th and 95th percentile of the truncated standard normal priors. The first case has a strong prior and a weak likelihood leading to a funnel shaped posterior with centered parameterization. The second case has a weak prior and a strong likelihood leading to a funnel shaped posterior with non-centered parameterization. The population mean was fixed at $\mu_0 = 0$ for both observations.

We implement this model with centered and non-centered parameterizations using the Stan probabilistic programming language (Stan Development Team 2024) and run the posterior inference with its default MCMC sampling algorithm, the no-U-turn sampler (Hoffman and Gelman 2014; Stan Development Team 2024). Despite the slower sampling, we raise the target acceptance ratio, called adapt delta, from 0.80 to 0.99, to more reliably sample posteriors with highly varying curvature. In each iteration of both prior and posterior SBC, we sample four MCMC chains of 1000 warm-up draws and 1000 posterior draws. To automate SBC iterations, we used the sbc R package (Kim et al. 2024).

Results

To assess the calibration of the inference, we inspect the uniformity of the PIT values of the joint log-likelihood, and of the parameters (μ_0, τ, σ) shared between the groups. PIT values are computed as described in Sections 2 and 3. We employ the graphical uniformity test introduced by Säilynoja et al. (2022), see e.g., Figure 3. The test compares the empirical cumulative distribution function (ECDF) of the discrete PIT values to 95% simultaneous central confidence bands under the assumption of discrete uniformity. We use the implementation of the test available in the bayesplot R package (Gabry and Mahr 2024). To provide better dynamic range in the plots, we show the ECDF difference (the observed ECDF minus the expected CDF values for uniform distribution) as recommended by Säilynoja et al. (2022). The PIT values for a given quantity are obtained by transforming the quantity with respect to the corresponding posterior draws, see Figure 2. That is, for the population parameters, the prior draw is transformed with respect to the posterior draws. For the joint log-likelihood, we compare the log-likelihood of the sample, when evaluated with the known true parameter values, against the joint log-likelihood evaluated with the parameter posterior draws.

(a) Centered Parameterization

(b) Non-centered Parameterization
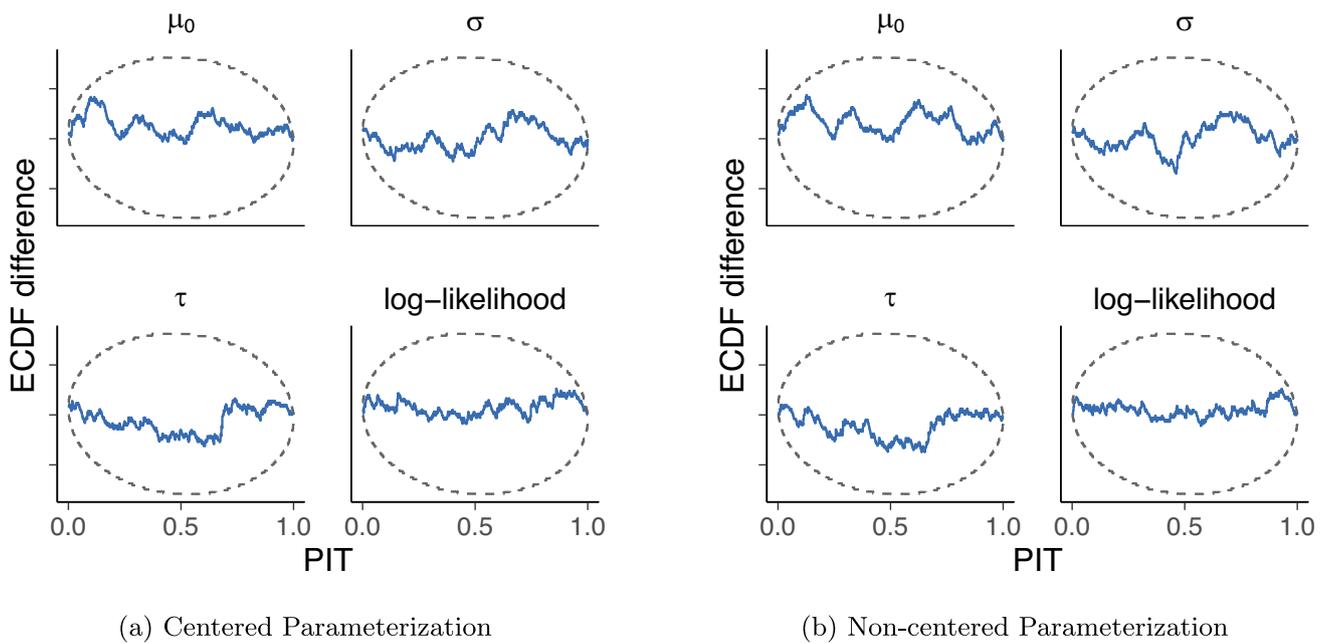
Fig. 3 *Prior SBC* checking for the hierarchical model with (a) the centered and (b) the non-centered parameterization using four different test quantities. Subplots show PIT-ECDF difference plots using four different test quantities and 95% simultaneous confidence intervals under the assumption of uniformity. As all blue PIT-ECDF difference lines are inside the 95% simultaneous confidence intervals, the inference seems to be calibrated when SBC iterations are averaged over prior draws

(a) Centered Parameterization

(b) Non-centered Parameterization
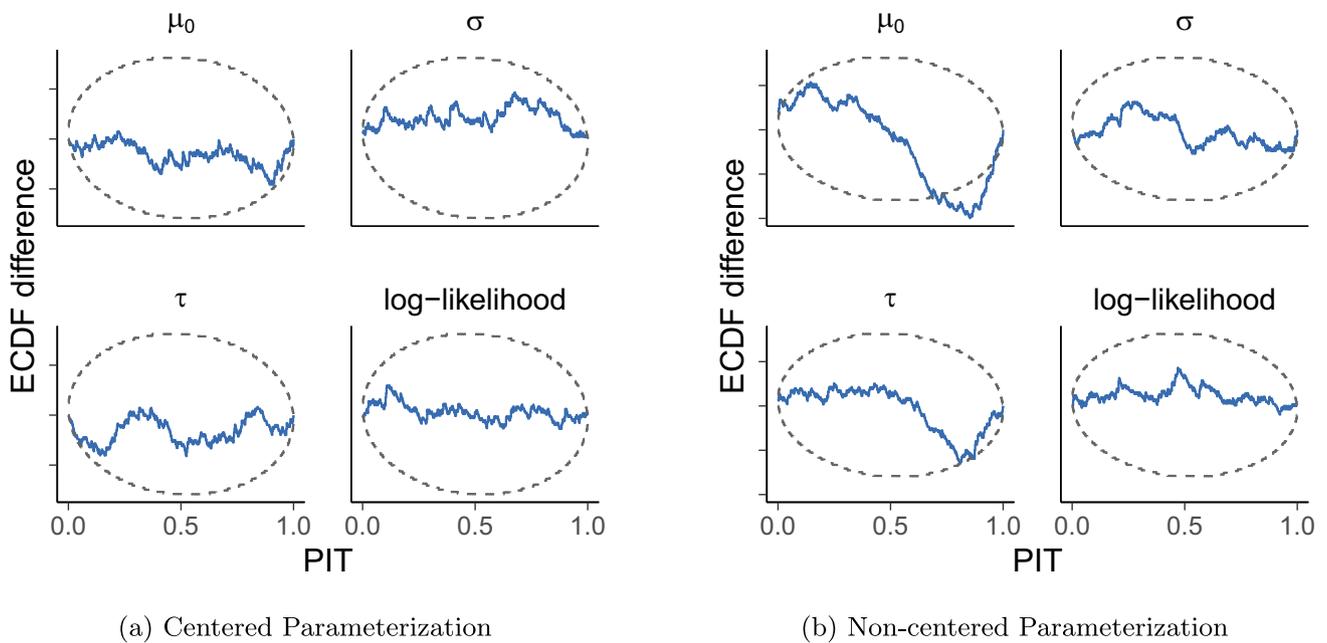
Fig. 4 *Posterior SBC* for the hierarchical model with an observed data with large τ and small σ, corresponding to *weak population prior and strong likelihood*, implying a close to normal posterior with (a) the centered parameterization, and a funnel shaped posterior with (b) the non-centered parameterization. Subplots show PIT-ECDF difference plots using four different test quantities and 95% simultaneous confidence intervals under the assumption of uniformity. The blue PIT-ECDF difference lines for the centered parameterization stay inside the envelope, indicating calibrated inference. The blue PIT-ECDF difference lines for the non-centered parametrization show that the right tail of the approximated posterior for μ_0 and τ tends to be thin (the line dips outside of the envelope) when SBC iterations are averaged over the observed data posterior draws

(a) Centered

(b) Non-centered
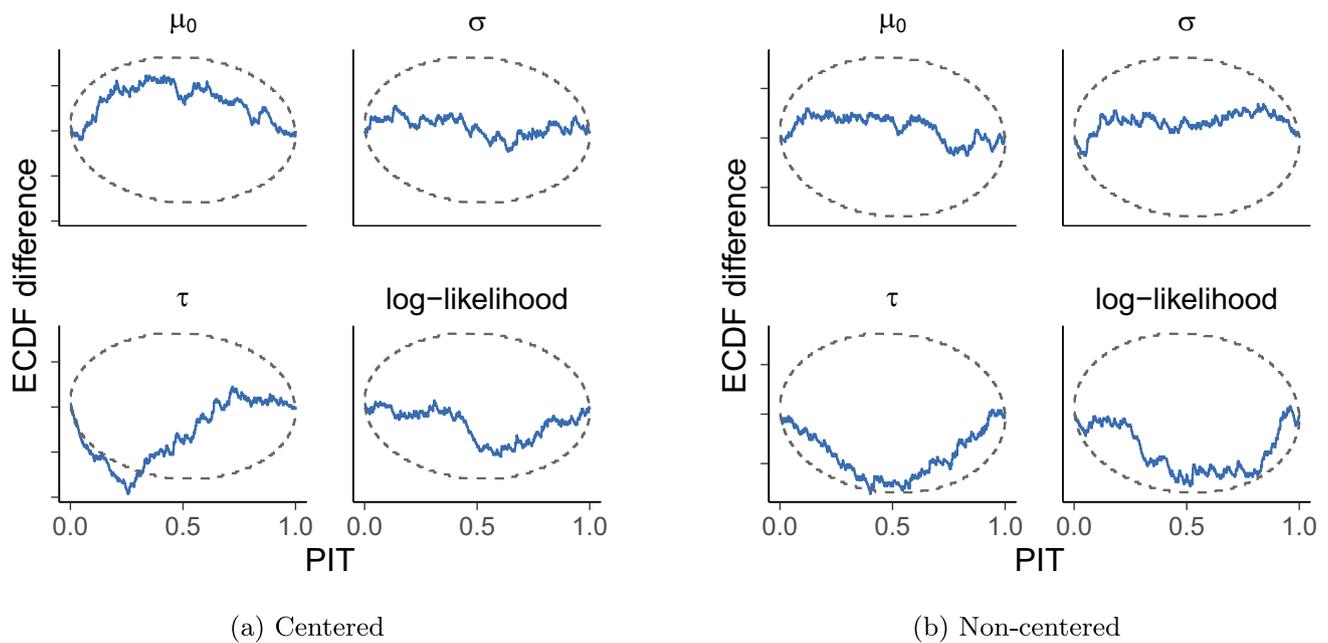
Fig. 5 *Posterior SBC* for the hierarchical model with an observed data with small τ and large σ, corresponding to *strong population prior and weak likelihood*, implying a funnel shaped posterior with (a) the centered parametrization, and a close to normal posterior with (b) the non-centered parametrization. Subplots show PIT-ECDF difference plots using four different test quantities and 95% simultaneous confidence intervals under the assumption of uniformity. The blue PIT-ECDF difference lines for the centered parameterization show that the inference often overestimates τ (the line dips below the envelope), when SBC iterations are averaged over the observed data posterior draws. The blue PIT-ECDF difference lines for the non-centered parameterization stay inside the envelope, indicating calibrated inference

After prior SBC with 500 iterations, Figure 3 shows no noticeable calibration issues in the inference with either parameterization. This would indicate to the modeler that everything is fine with either of these model implementations.

Next, we assess the calibration via posterior SBC by conditioning on the dataset with large τ and small σ. This corresponds to the weak population prior and strong likelihood case, implying a funnel shaped posterior with the non-centered parameterization. Running 500 iterations of posterior SBC reveals a very different story for the calibration of the two parameterizations. As shown in Figure 4, the centered parameterization shows excellent calibration, but the non-centered parameterization has calibration issues when evaluating the population level parameters μ_0 and τ.

The second dataset has been generated with small τ and large σ. This corresponds to a case of strong population prior and weak likelihood, implying a funnel shape for the posterior of the parameters in the model with centered parameterization. The posterior exhibits highly varying curvature, making it hard to explore for the inference algorithm. When we run posterior SBC conditional on this data, the presence of calibration issues with the centered parameterization of the model is confirmed. Figure 5b a shows the calibration assessment of posterior SBC, revealing the clear calibration issues with the centered parameterization, but also showing

possible overestimation of τ in the inference run with the non-centered parameterization. If we would like to inspect this effect more closely, we could run posterior SBC with more iterations to increase the sensitivity of the calibration checking.

To conclude, in this example, we showcased a model where prior SBC indicates no calibration issues with either of the candidate model implementations. In contrast, with posterior SBC, we observed that for both of the inspected datasets better calibration is achieved with one of the parameterizations, conditional on a given observed dataset.

In the presented cases, the Stan interface conducted automated inference diagnostics (partially in Stan itself and partially using the `posterior` R package, Bürkner et al., 2024), producing warnings for one or both of the parameterizations. In prior SBC both of the model parameterizations faced a considerable number of iterations with high $\widehat{R}$ values (Vehtari et al. 2021), implying bad mixing of the MCMC chains. Additionally, 12% of the iterations of the centered parameterization model had one or more divergent transitions, implying possibly biased posterior inference. In posterior SBC, the first dataset with strong likelihood produced warnings for the non-centered parameterization, and the second dataset with weak likelihood was problematic for the centered parameterization. These warnings align with our findings with SBC by also indicating the inference problems.

This is plausible since the issues are caused by the challenging posterior geometry. Although in this specific case the convergence diagnostics did indicate issues, the $\widehat{R}$ and divergence diagnostics do not quantify the direction or magnitude of bias for each parameter, which on the other hand can easily be read from the graphical PIT plots. For example, Figure 5b a shows that in the centered parameterization case the inference tend to overestimate τ (dip in the left hand side of PIT ECDF-difference plot indicates missing posterior mass for small values of τ), but the inference for σ is reasonably calibrated. Furthermore, good diagnostics are not necessarily available for new inference methods, which is demonstrated with amortized Bayesian inference in Section 4.3. Finally, when the model complexity grows, the inference issues might be more subtle, but still have a strong impact on the posterior estimates and calibration.

4.2 Lotka-Volterra model – focusing computational efforts

In this case study, we highlight how weakly informative priors can cause prior SBC to face computational issues, and how posterior SBC may provide substantial speed-ups for the calibration checking process. Additionally, we demonstrate a case where prior SBC indicates potential calibration issues for the model. However, these issues are not present in the parameter space explored by posterior SBC. Below, we show the results of performing both prior and posterior SBC for an inference task involving the Lotka-Volterra predator-prey model. We compare the findings of the calibration checking approaches, as well as the required computational efforts.

Setup The Lotka-Volterra predator-prey model consist of a system of first order non-linear differential equations, and describes the population dynamics of an interacting pair of species, a predator and its prey. Here we model the number of snowshoe hare and Canada lynx pelts collected by the Hudson Bay Company between years 1900 and 1920 (Odum and Barrett 2005). The expectations of the number of pelts collected are modelled as the solutions to the Lotka-Volterra equations,

$$\frac{d}{dt}H = \alpha H - \beta H L, \tag{4}$$

$$\frac{d}{dt}L = -\gamma L + \delta H L, \tag{5}$$

where H and L are the populations of hares and lynxes respectively, α is the rate at which the hare population would grow if no lynxes were present, and conversely γ is the death rate of lynxes without hares to hunt. The parameters β and δ characterize the effect that the interaction between the species has on their populations – some hares get eaten, and the lynx

population may grow. The signs in the equation system are chosen so that we expect each parameter to be positive.

We add an error term to model the measurement noise and the variation unexplained by the deterministic differential equation model. For both populations, the number of collected pelts is assumed to be log-normally distributed as

$$\log\left(\hat{H}_t\right) = \mathcal{N}\left(\log(H_t), \sigma_h\right),$$
$$\log\left(\hat{L}_t\right) = \mathcal{N}\left(\log(L_t), \sigma_l\right),$$

where $\hat{H}_t$ and $\hat{L}_t$ are the recorded numbers of pelts collected, H_t and L_t the latent true (trappable) populations of the species at that time, and σ_h and σ_l the error terms related to the species. All of these parameters are constrained to be positive. We do not have strong intuition on the interactions between the Lotka-Volterra model parameters, so we assign them the following independent priors, used for example in the case study of Carpenter (2018),

$$\alpha, \gamma \sim \mathcal{N}(1, 0.5)$$
$$\beta, \delta \sim \mathcal{N}(0.05, 0.05).$$

These priors are set with the aim to inform on the magnitude of the parameters and to keep the year-to-year variation of the populations from being too extreme. For both of the measurement error terms, we set a weakly informative log-normal prior:

$$\log\left(\sigma_i\right) \sim \mathcal{N}(-1, 1).$$

Figure 6 shows some prior predictive draws from the model together with the historical observations that will later use for posterior SBC. We describe these prior predictions more below, when discussing the results of the case study.

We implement the models using Stan and use NUTS for the posterior inference. This time, we initialize the Markov chains using the Pathfinder variational inference method (Zhang et al. 2022), which allows us to quickly obtain approximate draws close to the typical set of the posterior. This initialization step was added to the inference as the resulting posteriors are often multi-modal with a narrow high-probability mode corresponding to lower measurement error and a more informative likelihood for the ODE parameters, and a wider low-probability mode corresponding to high measurement error and low information on the ODE model parameters. With random initialization, some of the Markov-chains are prone to getting stuck on the low-probability mode, but Pathfinder can quickly find both modes and generate draws from the high-probability mode to be used as initial values for the Markov chains. Figure 7 shows an example of this posterior multimodality.

Fig. 6 Prior predictive draws of the observed pelt counts over time. Top left shows the true historical observation data. Many prior draws do not exhibit the observed periodicity, or are in other ways unrealistic
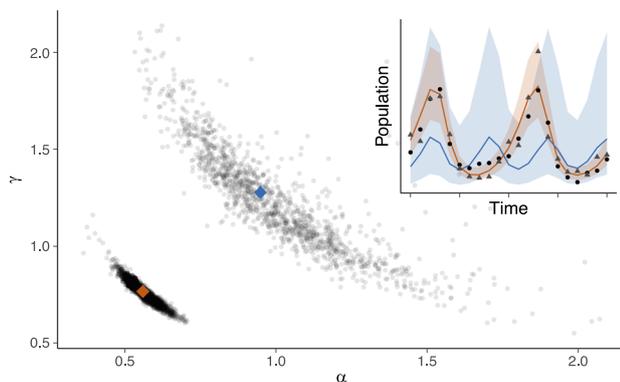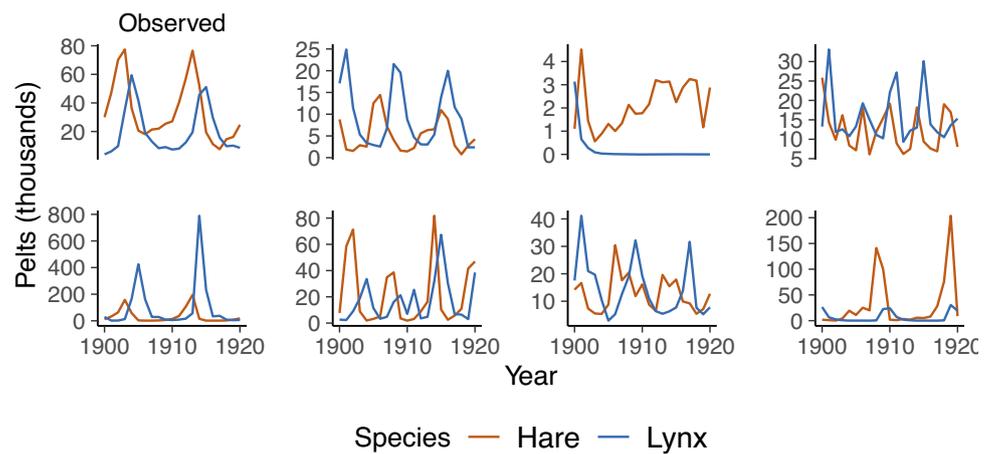

Fig. 7 Bimodal posterior of the two population growth parameters, inlaid with predictive means and predictive intervals of two posterior draws where (θ_1, θ_3) fall in two different modes. The concentrated mode has most of the posterior mass, and corresponds to lower predictive uncertainty, and a periodicity following the data closely. The more dispersed mode, having very small amount of the posterior mass, corresponds to high observation noise

Fig. 8 PIT-ECDF plots of the posterior joint log-likelihood test quantity in the Lotka-Volterra model for both prior and posterior SBC. Prior SBC uses only 250 iterations, resulting in wider confidence bands. Posterior SBC could include 500 iterations due to faster computation. The blue PIT-ECDF difference line for the prior SBC dips below the envelope, indicating the inference is not well calibrated when averaged over the prior draws. The blue PIT-ECDF difference line for the posterior SBC stays inside the envelope, indicating the inference is well calibrated when averaged over the observed data posterior draws

Results We set out to run 250 iterations of both prior SBC and posterior SBC, but could extend posterior SBC to 500 iterations, as the runtime per iteration was over five times faster for posterior SBC. The process of running 250 iterations of prior SBC took 6.5 hours when using parallel processing for sampling the individual MCMC chains, but sequential model fitting. Conversely, after using 10 seconds to run the initial posterior inference on the historical pelt dataset published by Mahaffy (2010), running 500 iterations of posterior SBC took only 2.5 hours.

In prior SBC, we observe some prior predictive draws resulting in posterior inference with severe convergence issues. When investigating the calibration of the inference, we look at the PIT-ECDF of the joint log-likelihood in Figure 8. There, we see some potential calibration issues manifesting as fewer than expected PIT values between 0.5 and 1. The calibration assessment for the individual parameters can be found in A.1, but shows no calibration issues

for the individual parameters before corrections for multiple testing.

Posterior SBC, in turn, indicates good calibration for the observed historical data. A graphical assessment of the uniformity of the PIT joint log-likelihood values is shown in Figure 8. The faster inference during the posterior SBC allowed us to iterate on the model and experiment on using Pathfinder for initializing the MCMC chains in the more desirable posterior mode. More details are shown in A.1.

To conclude, we have demonstrated a case, where the calibration issues indicated by prior SBC are not present once we condition on an observed dataset. Moreover, due to the problematic inference in prior SBC, posterior SBC turns out to be computationally less demanding and thus allows us to obtain more power on the calibration assessment while using a lower computational budget.

The observed inference problems during prior SBC are due to the independent weakly informative priors assigning a considerable prior probability to problematic parameter com-

binations, that are no longer plausible under the observation. Furthermore, when we inspect the prior predictive samples in Figure 6, we observe multiple samples that do not align with our expectations on the periodic nature of the data. We could try to improve the prior, but if the computation already works for the posterior it is not needed.

O'Brien et al. (2025) demonstrate a case where numerical error in the ODE solver distorts the posterior. From the usual inference diagnostic perspective, the bias due to the distortion is not distinguishable from the true posterior as the numerical error alters the target distribution itself. However, the distortion from the numerical error is unlikely to exhibit consistency over Bayesian updating, and thus posterior SBC would indicate the problem.

4.3 Posterior SBC in amortized Bayesian inference

In this case study, we show how posterior SBC can serve as a powerful diagnostic for amortized Bayesian inference workflows. Amortized Bayesian inference (see (Zammit-Mangion et al. 2024), for an overview) constitutes a new wave of Bayesian inference methods, where neural networks learn a direct surrogate for the posterior distribution. An overview diagram of amortized Bayesian inference is shown in Fig. 9. Normalizing flows (Papamakarios et al. 2021) are arguably the most prominent neural density estimator for amortized inference, but other architectures have recently been explored as well, such as score-based diffusion models (Geffner et al. 2023; Simons et al. 2023), flow matching (Dax et al. 2023), and consistency models (Schmitt et al. 2024). For amortized inference, there are no inference diagnostics similar to the convergence diagnostics for MCMC. While posterior SBC for MCMC has the high cost of running MCMC many times, in amortized inference the additional cost of running posterior SBC is negligible.

Amortized inference consists of two stages: (1) A training stage, where a generative neural network q_ϕ learns to distill relevant information from the probabilistic model based on synthetic data $(\theta, y) \sim p(\theta, y)$ from the joint model; and (2) the inference stage, where the trained neural networks approximate the posterior distribution for an arbitrary new dataset y_{obs}. Amortized inference casts fitting a probabilistic model as a neural network prediction task, which typically takes well under a second. This achieves near-instant approximate posterior sampling $\theta \sim q_\phi(\theta \mid y_{\text{obs}})$ and even direct posterior density evaluations for arbitrary parameter values θ.

In an effort to align amortized inference with the needs of common probabilistic modeling settings, there are two crucial extensions to the standard approach described above. First, in Bayesian inference, the data y can be replaced by *sufficient* summary statistics $h^*(y)$ without altering the posterior: $p(\theta \mid y) = p(\theta \mid h^*(y))$. In amortized inference, neural

networks are employed to learn embeddings of the data *in tandem* with the posterior approximator (Radev et al. 2022, 2020; Chen et al. 2021; Chan et al. 2018; Huang et al. 2023; Schmitt et al. 2024). These *summary networks* h_ψ are parameterized by learnable neural network weights ψ and learn a fixed-length representation of the data which is approximately sufficient for posterior inference (not necessarily for reconstructing the data). Second, datasets $y = y_1, \ldots, y_N$ can come with a varying number of observations N. While the summary network learns to compress datasets of varying length to a fixed-length vector $h_\psi(y)$, the Bayesian posterior is influenced by the number of observations N (e.g., as described by contraction). Hence, the neural network needs to be informed about the number of observations in the raw dataset, and we need to cover datasets with varying numbers of observations in the neural network training stage. Accordingly, we draw the number of observations in each *simulated* dataset from a distribution $p(N)$ and condition the generative neural network on N in addition to the learned summary statistics $h_\psi(y)$. In summary, the forward simulation process is formalized as

$$N \sim p(N), \quad \theta \sim p(\theta), \quad y \sim p(y \mid \theta, N). \quad (6)$$

The resulting neural network training for a conditional normalizing flow with learned summary statistics h_ψ and the number of observations N as additional conditioning variable minimizes the maximum likelihood objective,

$$\psi^*, \phi^* = \underset{\psi, \phi}{argmin} \, \mathbb{E}_p \Big[-\log q_\phi \Big(\theta \mid h_\psi(y), N \Big) \Big], \quad (7)$$

where the expectation is taken over the joint distribution $p(N, \theta, y)$.

However, there is no free lunch, and state-of-the-art amortized inference methods still lack the gold-standard guarantees from established algorithms like MCMC. The two most pressing issues of amortized neural posterior estimators are (1) the increased number of design choices from neural network architectures and training hyperparameters; and (2) the lack of performance guarantees under domain shifts between the simulated training data and the real data y_{obs}, for example resulting from model misspecification. As a consequence, there is a need for strong diagnostics to gauge the trustworthiness of amortized neural posterior approximators. In the context of non-amortized MCMC, one bottleneck of simulation-based calibration (both prior and posterior SBC) is the associated computational burden from repeated model re-fits on new simulated datasets. Yet, an *amortized* approximator can sample the required approximate posterior draws for new datasets in near-instant time, which reduces the runtime for SBC to only a few seconds. Therefore, posterior SBC naturally lends itself to amortized inference due to (1)
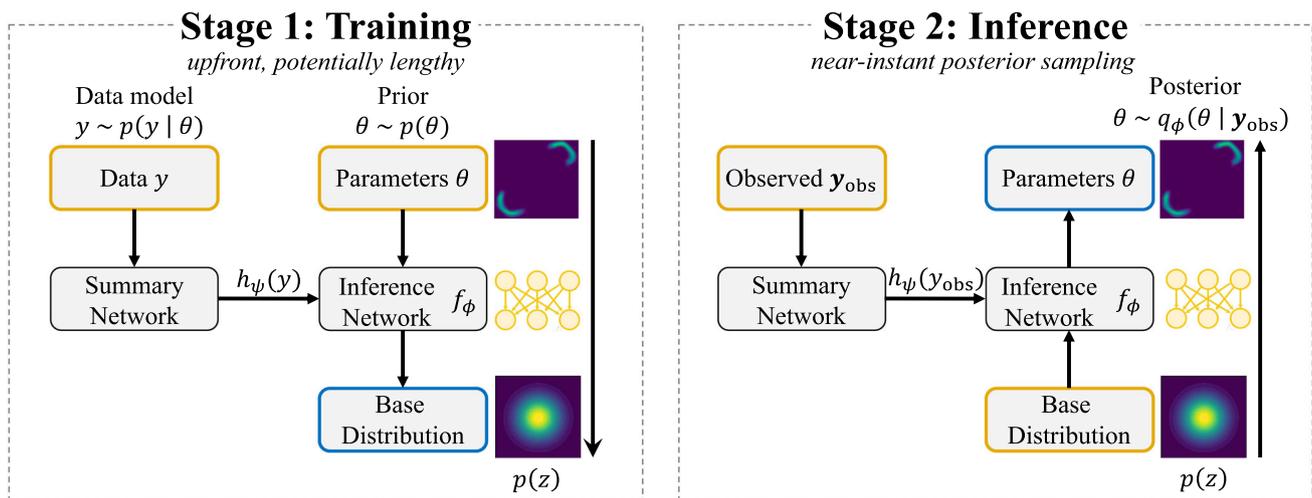
Fig. 9 Conceptual overview of amortized Bayesian inference with normalizing flows. **Stage 1: Training.** Based on samples from the joint model $p(\theta, y)$, a neural network tandem simultaneously learns to extract sufficient summary statistics $h_\psi(y)$ and establish a conditional mapping to a base distribution $p(z)$. **Stage 2: Inference.** Given a new unseen dataset y_{obs}, we can draw samples from the base distribution $p(z)$ and pass them through the inverted inference network conditional on the summary statistics $h_\psi(y_{obs})$ to obtain samples from the approximate posterior in near-instant time

the particular need for strong diagnostics; and (2) the stunningly quick runtime with an amortized approximator Fig 9.

Setup To improve our understanding of human cognition, researchers employ increasingly sophisticated statistical models to explore the neurological connections between cognitive processes and physical phenomena. From a mathematical perspective, human decision-making can be represented by a stochastic evidence accumulation process, specifically through a drift-diffusion model (DDM; Voss et al. 2004; Ratcliff and McKoon 2008). The DDM assumes that human decision-making is based on sequential integration of evidence over time until it reaches a threshold. The model is parameterized by an individual's cognitive parameters (e.g., information uptake speed and decision thresholds). In its general form, a DDM is mathematically described by a differential equation,

$$dy_t = v\delta dt + \sigma W_t, \tag{8}$$

where δ is the drift rate and σ controls the stochastic variance of the Wiener process W_t. Simulation programs use a discretized version of the process,

$$y_{t+\Delta t} = y_t + \delta \Delta t + \sigma \varepsilon \sqrt{\Delta t}, \tag{9}$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ is Gaussian noise and Δt controls the time resolution (5 milliseconds in our data model). The DDM is additionally parameterized by a decision threshold α, an evidence starting point β to represent biases, and a non-decision time τ to account for motor reaction times. We refer to A.2 for more details.

In a line of applied work, neurological research has identified neural markers that are associated with decision-making processes. In a recent paper, Ghaderi-Kangavari et al. (2023) propose a set of multiple statistical models that integrate both the cognitive drift-diffusion model and a compressed representation of neural measurements, as found in EEG or fMRI data. Such integrative models, which combine neural processes at the level of single experimental trials, represent the current state-of-the-art in cognitive modeling research. This case study implements two probabilistic models M_1 and M_2, which correspond to models #2 and #6 from Ghaderi-Kangavari et al. (2023). We use deep neural networks to fit an amortized posterior approximator for each model. Detailed specifications of models and neural networks can be found in A.2. We showcase how posterior SBC can yield important information to compare the neural approximations of the candidate models conditional on two real datasets from another study (Georgie et al. 2018).

Results In the following, we summarize the main results of the case study and refer to A.2 for additional details. In neural amortized inference, we must check the inference validity for different numbers of observations because the latter is passed as a conditioning variable to the inference network f_ϕ. To this end, we confirm that amortized inference on synthetic datasets simulated from the joint model $p(N, \theta, y)$ can recover the ground-truth values of the model parameters that are identifiable for both N_{obs} and $2N_{obs}$ (see A.2). Further, we can confirm that prior SBC checking for datasets with N_{obs}
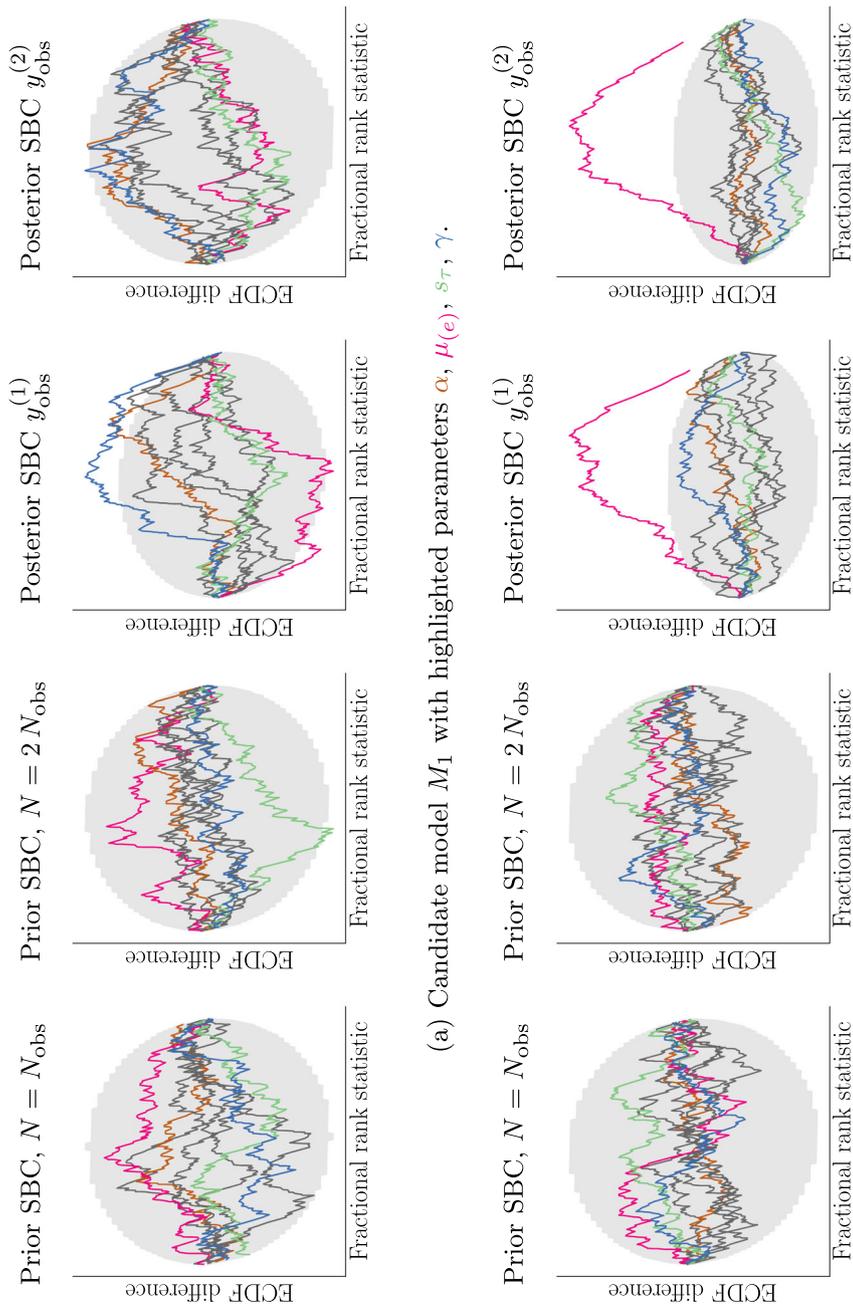
(a) Candidate model M_1 with highlighted parameters α, $\mu_{(e)}$, s_T, γ.

(b) Candidate model M_2 with highlighted parameters α, $\mu_{(e)}$, s_T, λ.

Fig. 10 Results of prior and posterior simulation-based calibration checking with amortized inference. The Bayesian models M_1 and M_2 are two joint integrative neuroscience models from Ghaderi-Kangavari et al. (2023), and the datasets for posterior SBC checking are two real datasets $y_{\text{obs}}^{(1)}$, $y_{\text{obs}}^{(2)}$ from Georgie et al. (2018)

as well as $2N_{\mathrm{obs}}$ yields satisfactory results (see Figure 10). However, this conclusion drastically changes when we perform SBC checking conditional on the real datasets $y_{\mathrm{obs}}^{(1)}$ and $y_{\mathrm{obs}}^{(2)}$. As shown in Figure 10a , posterior SBC indicates minor data-conditional calibration issues of model M_1 for the first observed dataset $y_{\mathrm{obs}}^{(1)}$ while calibration conditional on the second observed dataset $y_{\mathrm{obs}}^{(2)}$ is within the acceptance band. The second model M_2, however, shows pathologically bad calibration of the mean visual encoding latency $\mu_{(e)}$ conditional on either observed dataset (see Figure 10b). As a consequence, model M_2 might be refined according to the iterative Bayesian workflow (Gelman et al. 2020; Schad et al. 2020) with a particular focus on the mean visual encoding latency parameter $\mu_{(e)}$.

In this case study, posterior SBC could flag calibration issues of an amortized approximator that were not detectable with prior SBC, which illustrates its potential as a default data-conditional diagnostic in amortized Bayesian workflows (Radev et al. 2023; Schmitt et al. 2024).

5 Discussion

Prior SBC and posterior SBC have different roles in the Bayesian workflow. Prior SBC can be most useful for developers of probabilistic programming software, inference algorithms, models, and priors, who need their implementations to work for a wide range of datasets. Once data has been collected, posterior SBC can be used to diagnose whether the used inference algorithm is well calibrated, that is, sampling faithfully from the posterior and the augmented posteriors, which are close to the original posterior. In case of miscalibration, posterior SBC can indicate also the direction and magnitude of the bias for each parameter or other quantity of interest.

In many practical modeling tasks, the computational cost of the inference scales linearly when one increases the size of the data to condition on. In such cases, using the full observed data and an equally sized predictive sample should be feasible as long as running the original posterior inference is within computational budgets. However, for example with models involving superlinear matrix operations like Cholesky decomposition ($\mathcal{O}(n^3)$), the computational cost for the larger dataset can get too high. In this case, one can use S randomized subsets of the data, making the augmented datasets match the original data size, and proceed to perform both posterior fits independently for each iteration of posterior SBC. This approach can also be used if there is a worry that the augmented posterior is too different from the original posterior due to having double the data size. In practice, the approach to subsetting the data depends on the modeling context. For example, when predicting election results, one does

not expect the number of states to increase, but observations can be added to these groups. In clinical studies, however, one would not necessarily add observations to existing groups in the form of obtaining more measurements from the patients already included in the study, but the number of groups could increase as more patients enroll to the study.

Passing posterior SBC checking is a necessary but not sufficient condition to trust the posterior inference in the same way as most convergence diagnostics given finite computation time. If the common MCMC diagnostics, like $\widehat{R}$ and divergences, indicate big problems, it is good to first investigate the possible reasons before spending computation time for posterior SBC. Even if the convergence issues are not completely solved, it is possible to use posterior SBC to assess the direction and magnitude of the potential bias.

Posterior SBC is specifically useful when the usual diagnostics may have missed something, when the usual diagnostics may have false positive warnings, when using new algorithm implementations that are not yet trusted, or when using new inference algorithms which do not yet have faster diagnostics. Posterior SBC seems to be specifically useful for amortized inference, as the computational cost for repeated inference is low.

Appendix A case studies - details and additional results

This supplement provides details and additional results for the case studies of the main paper. The source code for reproducing these case studies is available at https://github.com/TeemuSailynoja/posterior-sbc.

A.1 Case study 2: Lotka-Volterra model

Below, we present calibration assessments, as well as parameter recovery results for the individual model parameters for both prior and posterior SBC of the Lotka-Volterra model implementation.

In Figure 11a we display the graphical calibration assessment of the individual parameters obtained with prior SBC for the Lotka-Volterra model. Even without correction for multiple comparison, the inference of the parameter posteriors looks well calibrated. In the parameter recovery plots for prior SBC, displayed in Figure 11b , we see no systematic biases, but there are some cases where the posterior is very far from the ground truth. As discussed in Section 3.2 of the article, these outliers are likely due to the posterior having distinct modes, which leads to a high probability of one or more MCMC chains not properly exploring the full posterior.

Figure 12a and Figure 12b show the calibration assessment and recovery of the individual parameters once we condition the inference on the historical observations. Again, we
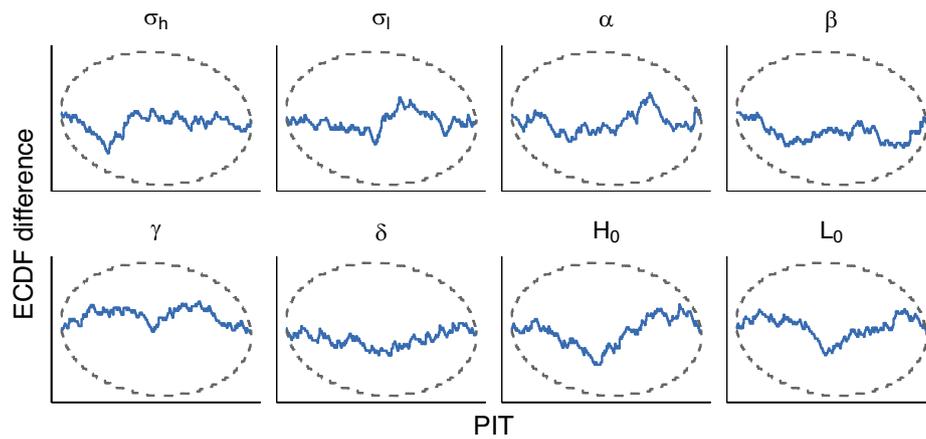
(a) Individual parameter calibration assessments.
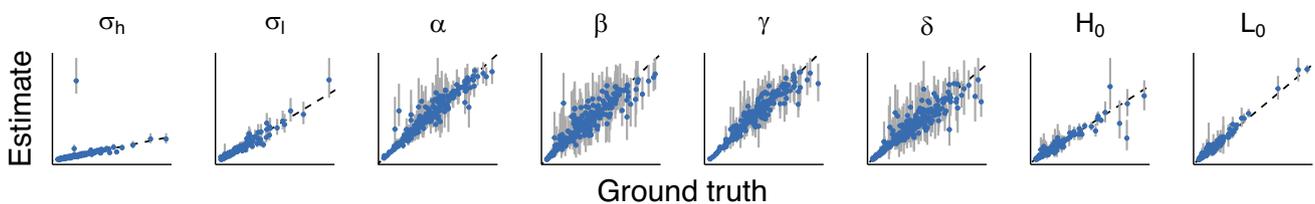
(b) Parameter recovery.

Fig. 11 Per parameter results of prior SBC for the Lotka-Volterra model

observe no calibration issues. The parameter recovery also exhibits no outliers similar to those observed in prior SBC.

A.2 Case study 3: Posterior SBC checking in amortized Bayesian inference

In the following, we detail the full neural network settings as well as model specifications in case study 3, and present additional results concerning the parameter recovery of the amortized approximators.

Training setup and neural network settings

The prior distribution $p(N)$ on the number of observations in the synthetic training datasets is an integer-valued uniform distribution $\mathcal{U}(50, 150)$. The summary network is a DeepSet (Zaheer et al. 2017) with mean-pooling, which combines equivariant and invariant neural network layers to extract a 32-dimensional summary vector $h_\psi(y) \in \mathbb{R}^{32}$ from each dataset $y \in \mathbb{R}^{N \times 3}$, where $N \in \{50, \ldots, 150\}$ as specified above. The inference network is an affine coupling flow with 6 coupling layers. We train the neural network tandem for a total of 300 epochs with a batch size of 64, 1000 iterations per epochs, and determine the learning rate based on a cosine decay schedule with initial value of $5 \cdot 10^{-4}$.

Model 1

The first probabilistic model M_1 corresponds to "model 2" in Ghaderi-Kangavari et al. (2023). It has the following data model,

$$
\begin{aligned}
r_i, y_i &\sim \mathrm{DDM}(\alpha, \tau_{(e)i} + \tau_{(m)}, \delta, \beta), \\
z_i &\sim \mathcal{N}(\gamma \cdot \tau_{(e)i}, \sigma^2), \\
\tau_{(e)i} &\sim \mathcal{N}(\mu_{(e)}, s_{(\tau)}^2),
\end{aligned} \tag{A1}
$$

where $\mathrm{DDM}(\alpha, \tau, \delta, \beta)$ denotes a Wiener drift-diffusion model with threshold α, non-decision time τ, average drift rate δ, and initial bias β. We use the following prior distributions from Ghaderi-Kangavari et al. (2023):

$$
\begin{aligned}
\delta &\sim \mathcal{U}(-3, 3), & \alpha &\sim \mathcal{U}(0.5, 2), \\
\beta &\sim \mathcal{U}(0.1, 0.9), & \mu_{(e)} &\sim \mathcal{U}(0.05, 0.6), \\
\tau_{(m)} &\sim \mathcal{U}(0.06, 0.8), & \sigma &\sim \mathcal{U}(0, 0.3), \\
s_{(\tau)} &\sim \mathcal{U}(0, 0.3), & \gamma &\sim \mathcal{U}(0, 3).
\end{aligned} \tag{A2}
$$

Figure 13 shows the parameter recovery of the amortized approximator across a range of settings. We observe that most parameters can be recovered with sufficiently
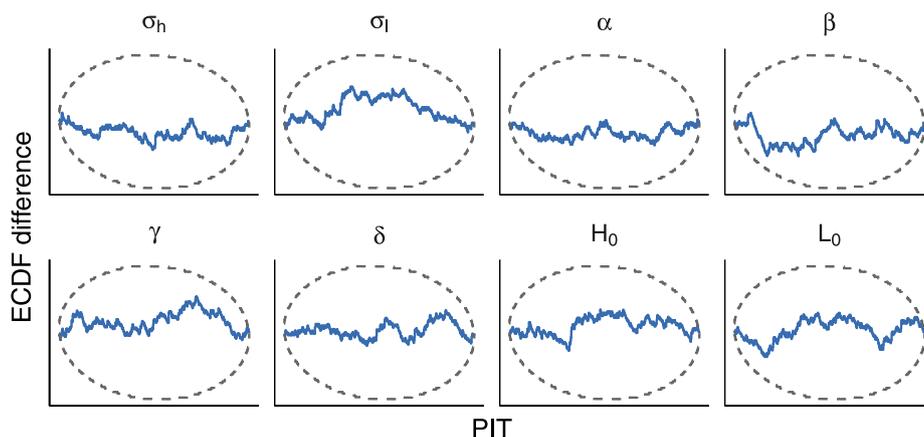
(a) Individual parameter calibration assessments.
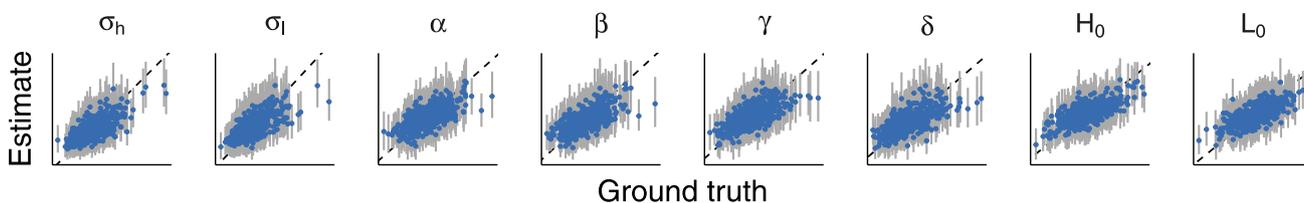
(b) Parameter recovery.

Fig. 12 Per parameter results of posterior SBC for the Lotka-Volterra model

high accuracy given the number of available observations in each dataset (see Figure 13a). Further, the epistemic uncertainty in the parameter estimates shrinks as the number of observations grows (see Figure 13b). Finally, errors in the posterior estimates propagate into the parameter recovery based on simulated data from the approximate posterior predictive distribution (see Figure 13c,d). This evaluation follows the same principle as posterior SBC but reports the data-conditional parameter recovery rather than the data-conditional simulation-based calibration results. In summary, the patterns from the empirical parameter recovery match the results from (prior and posterior based) simulation-based calibration checking in Section 2.3.

Model 2

The second probabilistic model M_2 represents "model 6" by Ghaderi-Kangavari et al. (2023), which implements a drift-diffusion model with collapsing boundary (Drugowitsch et al. 2012; Hawkins et al. 2015; Ratcliff et al. 2016). The collapsing boundaries are formalized through a scaled Weibull cumulative distribution function,

$$u(t) = \alpha - \left(1 - \exp\left(-\left(\frac{t}{\lambda}\right)^3\right)\right) \cdot \frac{\alpha}{2}, \tag{A3}$$

$$l(t) = \alpha - u(t),$$

with upper threshold $u(t)$ and lower threshold $l(t)$ as a function of the time passed in the experiment. Again, we follow Ghaderi-Kangavari et al. (2023) and use the following prior distributions:

$$
\begin{aligned}
\delta &\sim \mathcal{U}(-3, 3), & \alpha &\sim \mathcal{U}(0.5, 2), \\
\beta &\sim \mathcal{U}(0.1, 0.9), & \mu_{(e)} &\sim \mathcal{U}(0.05, 0.6), \\
\tau_{(m)} &\sim \mathcal{U}(0.06, 0.8), & \sigma &\sim \mathcal{U}(0, 0.3), \\
s_{(\tau)} &\sim \mathcal{U}(0, 0.3), & \lambda &\sim \mathcal{U}(0.5, 4).
\end{aligned} \tag{A4}
$$

Paralleling the previous evaluation for model M_1, Figure 14 reports the parameter recovery capabilities of the amortized approximator. Again, most parameters can be recovered with sufficiently high accuracy given the number of available observations in each dataset (see Figure 14a), and the uncertainty shrinks when the datasets contain more observations (see Figure 14b). The large epistemic uncertainty in the esti-
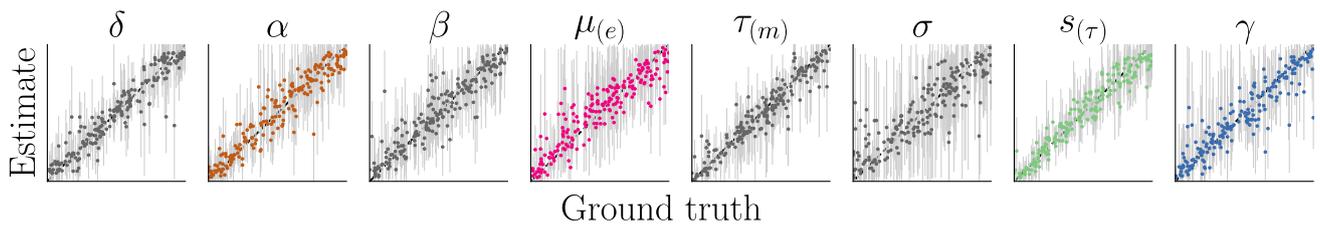
(a) Recovery on synthetic datasets with $N = N_{obs}$ observations each.
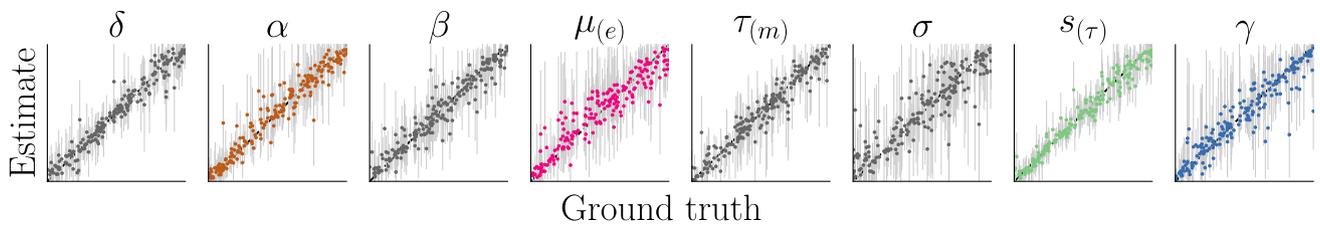
(b) Recovery on synthetic datasets with $N = 2\,N_{obs}$ observations each.
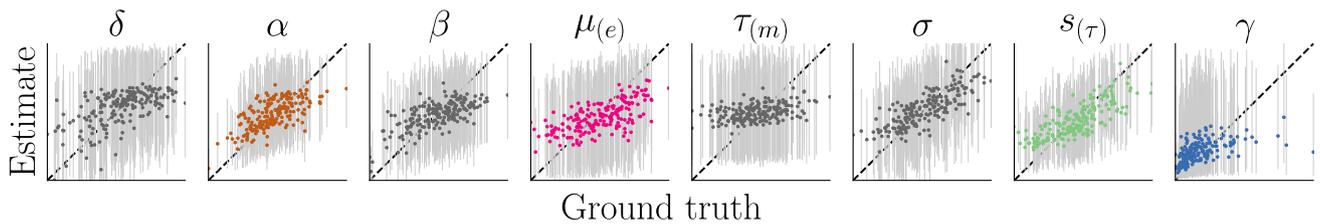
(c) Recovery on data from the posterior predictive distribution conditional on $y_{obs}^{(1)}$.
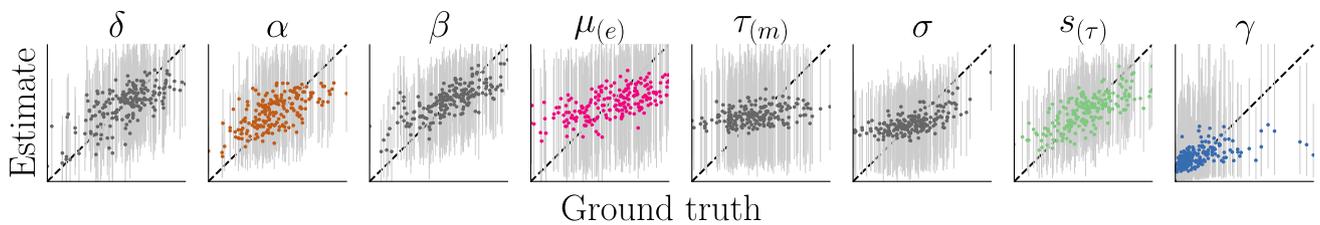
(d) Recovery on data from the posterior predictive distribution conditional on $y_{obs}^{(2)}$.

Fig. 13 Parameter recovery of the amortized approximator for model M_1

mate of the boundary collapse parameter λ has previously been reported by Ghaderi-Kangavari et al. (2023). Furthermore, the parameter recovery patterns mirror the results from (prior and posterior based) simulation-based calibra-tion checking in Section 3.3. Most notably, the previously reported data-conditional miscalibration in the parameter $\mu_{(e)}$ (Figure 10b in the article) manifests in a biased data-conditional parameter recovery in Figure 14c ,d.
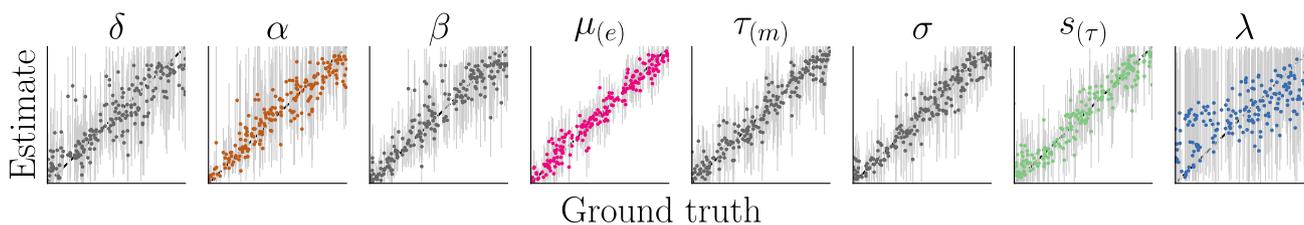
(a) Recovery on synthetic datasets with $N = N_{\text{obs}}$ observations each.
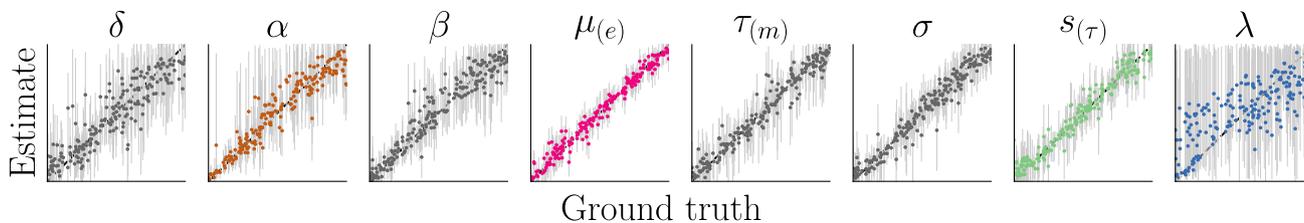
(b) Recovery on synthetic datasets with $N = 2\,N_{\text{obs}}$ observations each.
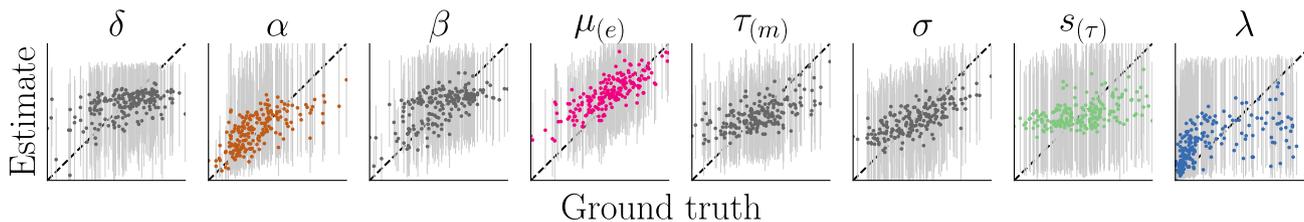
(c) Recovery on data from the posterior predictive distribution conditional on $y_{\text{obs}}^{(1)}$.
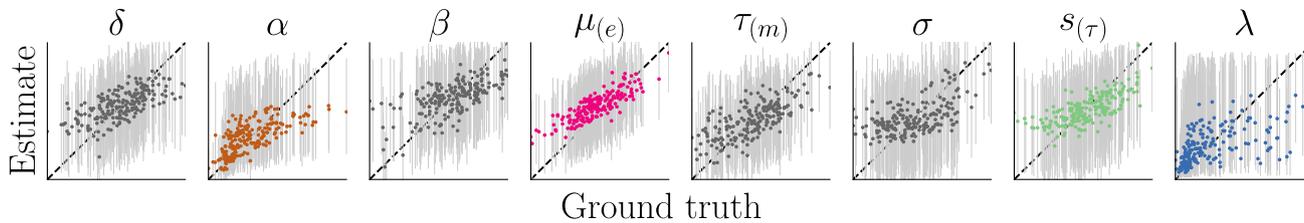
(d) Recovery on data from the posterior predictive distribution conditional on $y_{\text{obs}}^{(2)}$.

Fig. 14 Parameter recovery of the amortized approximator for model M_2

Code Availability Code to reproduce all experiments is available in the public repository at https://github.com/TeemuSailynoja/posterior-sbc.

Declarations

Competing interests The authors declare no competing interests.

References

Betancourt, M.: A Conceptual Introduction to Hamiltonian Monte Carlo (2018). arXiv:1701.02434

Bürkner, P.-C., Gabry, J., Kay, M., Vehtari, A.: posterior: Tools for Working with Posterior Distributions. R package version 1.6.0 (2024). https://mc-stan.org/posterior/

Box, G.E.P.: Sampling and Bayes' Inference in Scientific Modelling and Robustness. Journal of the Royal Statistical Society. Series A (General) **143**(4), 383 (1980) https://doi.org/10.2307/2982063

Carpenter, B.: Predator-prey population dynamics: the Lotka-Volterra model in Stan. Technical report, Columbia University (January 2018). https://mc-stan.org/learn-stan/case-studies/lotka-volterra-predator-prey.html Accessed 2025-01-23

Cook, S.R., Gelman, A., Rubin, D.B.: Validation of software for Bayesian models using posterior quantiles. J. Comput. Graph. Stat. **15**(3), 675–692 (2006). https://doi.org/10.1198/106186006X136976

Chan, J., Perrone, V., Spence, J., Jenkins, P., Mathieson, S., Song, Y.: A likelihood-free inference framework for population genetic data using exchangeable neural networks. Neural Information Processing Systems **31** (2018)

Chen, Y., Zhang, D., Gutmann, M.U., Courville, A., Zhu, Z.: Neural approximate sufficient statistics for implicit models. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=SRDuJssQud

Drugowitsch, J., Moreno-Bote, R., Churchland, A.K., Shadlen, M.N., Pouget, A.: The cost of accumulating evidence in perceptual decision making. J. Neurosci. **32**(11), 3612–3628 (2012). https://doi.org/10.1523/jneurosci.4010-11.2012

Dax, M., Wildberger, J., Buchholz, S., Green, S.R., Macke, J.H., Schölkopf, B.: Flow matching for scalable simulation-based inference. In: Neural Information Processing Systems (2023)

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. Chapman and Hall/CRC, (2013). https://doi.org/10.1201/b16018

Gelman, A.: Correction to Cook, Gelman, and Rubin (2006). J. Comput. Graph. Stat. **26**(4), 940–940 (2017). https://doi.org/10.1080/10618600.2017.1377082

Ghaderi-Kangavari, A., Rad, J.A., Nunez, M.D.: A general integrative neurocognitive modeling framework to jointly describe EEG and decision-making on single trials. Computational Brain and Behavior **6**, 317–376 (2023). https://doi.org/10.1007/s42113-023-00167-4

Gabry, J., Mahr, T.: bayesplot: Plotting for Bayesian Models (2024). https://mc-stan.org/bayesplot/

Gelman, A., Meng, X.-L., Stern, H.: Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**(4), 733–760 (1996)

Georgie, Y.K., Porcaro, C., Mayhew, S.D., Bagshaw, A.P., Ostwald, D.: A perceptual decision making EEG/fMRI data set. bioRxiv 253047 (2018) https://doi.org/10.1101/253047

Geffner, T., Papamakarios, G., Mnih, A.: Compositional score modeling for simulation-based inference. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 11098–11116. PMLR, (2023)

Gelman, A., Simpson, D., Betancourt, M.: The prior can often only be understood in the context of the likelihood. Entropy **19**(10), 555 (2017). https://doi.org/10.3390/e19100555

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A.: Visualization in Bayesian workflow. J. R. Stat. Soc. A. Stat. Soc. **182**(2), 389–402 (2019). https://doi.org/10.1111/rssa.12378

Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., Modrák, M.: Bayesian workflow. arXiv:2011.01808 (2020)

Huang, D., Bharti, A., Souza, A.H., Acerbi, L., Kaski, S.: Learning robust statistics for simulation-based inference under model misspecification. In: Thirty-seventh Conference on Neural Information Processing Systems (2023). https://openreview.net/forum?id=STrXsSIEiq

Hawkins, G.E., Forstmann, B.U., Wagenmakers, E.-J., Ratcliff, R., Brown, S.D.: Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. J. Neurosci. **35**(6), 2476–2484 (2015). https://doi.org/10.1523/jneurosci.2410-14.2015

Hoffman, M.D., Gelman, A.: The No-U-Turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. J. Mach. Learn. Res. **15**(47), 1593–1623 (2014)

Kim, S., Moon, A.H., Modrák, M., Säeilynoja, T., Fazio, L., Stenz, T.: Simulation-based Calibration: SBC (2024). https://hyunjimoon.github.io/SBC/

Mahaffy, J.M.: Math 636 - Mathematical Modeling Fall Semester, 2010 Lotka-Volterra Models (2010). http://jmahaffy.sdsu.edu/courses/f09/math636/lectures/lotka/qualde2.html Accessed 2025-01-24

Modrák, M., Moon, A.H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., Vehtari, A.: Simulation-based calibration checking for Bayesian computation: The choice of test quantities

shapes sensitivity. Bayesian Anal. (2023). https://doi.org/10.1214/23-BA1404

Neal, R.M.: Slice sampling. Ann. Stat. **31**(3), 705–767 (2003)

Odum, E.P., Barrett, G.W.: Fundamentals of Ecology, 5th edn. Thomson Brooks/Cole, Belmont, CA (2005)

O'Brien, T., Moores, M., Warton, D., Falster, D.: Here be dragons: Bimodal posteriors arise from numerical integration error in longitudinal models (2025) arXiv:2502.11510

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B.: Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research **22**(1) (2021)

Papaspiliopoulos, O., Roberts, G.O., Sköld, M.: A general framework for the parametrization of hierarchical models. Statistical Science **22**(1) (2007) https://doi.org/10.1214/088342307000000014

Ratcliff, R., McKoon, G.: The diffusion decision model: Theory and data for two-choice decision tasks. Neural Comput. **20**(4), 873–922 (2008). https://doi.org/10.1162/neco.2008.12-06-420

Radev, S.T., Mertens, U.K., Voss, A., Ardizzone, L., Köthe, U.: BayesFlow: Learning complex stochastic models with invertible neural networks. IEEE Transactions on Neural Networks and Learning Systems **33**(4), 1452–1466 (2022). https://doi.org/10.1109/TNNLS.2020.3042395

Radev, S.T., Mertens, U.K., Voss, A., Köthe, U.: Towards end-to-end likelihood-free inference with convolutional neural networks. Br. J. Math. Stat. Psychol. **73**(1), 23–43 (2020). https://doi.org/10.1111/bmsp.12159

Ratcliff, R., Smith, P.L., Brown, S.D., McKoon, G.: Diffusion decision model: Current issues and history. Trends Cogn. Sci. **20**(4), 260–281 (2016). https://doi.org/10.1016/j.tics.2016.01.007

Radev, S.T., Schmitt, M., Schumacher, L., Elsemüller, L., Pratz, V., Schälte, Y., Köthe, U., Bürkner, P.-C.: BayesFlow: Amortized Bayesian workflows with neural networks. Journal of Open Source Software **8**(89), 5702 (2023) https://doi.org/10.21105/joss.05702

Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics **12**(4) (1984) https://doi.org/10.1214/aos/1176346785

Štrumbelj, E., Bouchard-Côté, A., Corander, J., Gelman, A., Rue, H., Murray, L., Pesonen, H., Plummer, M., Vehtari, A.: Past, Present and Future of Software for Bayesian Inference. Stat. Sci. **39**(1), 46–61 (2024). https://doi.org/10.1214/23-STS907

Schmitt, M., Bürkner, P.-C., Köthe, U., Radev, S.T.: Detecting model misspecification in amortized bayesian inference with neural networks. In: Köthe, U., Rother, C. (eds.) Pattern Recognition, pp. 541–557. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-54605-1_35

Schad, D.J., Betancourt, M., Vasishth, S.: Toward a principled Bayesian workflow in cognitive science. arXiv:1904.12765 (2020)

Säilynoja, T., Bürkner, P.-C., Vehtari, A.: Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. Stat. Comput. **32**(2), 32 (2022). https://doi.org/10.1007/s11222-022-10090-6

Schmitt, M., LI, C., Vehtari, A., Acerbi, L., Bürkner, P.-C., Radev, S.T.: Amortized bayesian workflow (extended abstract). In: NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty (2024). https://openreview.net/forum?id=pKZp2JjC76

Schad, D.J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., Vasishth, S.: Workflow techniques for the robust use of bayes factors. Psychol. Methods **28**(6), 1404 (2023). https://doi.org/10.1037/met0000472

Schmitt, M., Pratz, V., Koethe, U., Bürkner, P.-C., Radev, S.T.: Consistency models for scalable and fast simulation-based inference. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024). https://openreview.net/forum?id=IHjKpKljyH

Simons, J., Sharrock, L., Liu, S., Beaumont, M.: Neural score estimation: Likelihood-free inference with conditional score based diffusion models. In: Fifth Symposium on Advances in Approximate Bayesian Inference (2023). https://openreview.net/forum?id=AVkJEb1ahOY

Stan Development Team: Stan Users Guide and Reference Manual. https://mc-stan.org (2024)

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., Gelman, A.: Validating Bayesian inference algorithms with simulation-based calibration (2020) arXiv:1804.06788

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., Bürkner, P.-C.: Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). Bayesian Anal. **16**(2), 667–718 (2021). https://doi.org/10.1214/20-BA1221

Voss, A., Rothermund, K., Voss, J.: Interpreting the parameters of the diffusion model: An empirical validation. Memory & Cognition **32**(7), 1206–1220 (2004). https://doi.org/10.3758/bf03196893

Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Yes, but did it work?: Evaluating variational inference. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 5581–5590. PMLR, (2018). https://proceedings.mlr.press/v80/yao18a.html

Zhang, L., Carpenter, B., Gelman, A., Vehtari, A.: Pathfinder: Parallel quasi-Newton variational inference. J. Mach. Learn. Res. **23**(306), 1–49 (2022)

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., (2017)

Zammit-Mangion, A., Sainsbury-Dale, M., Huser, R.: Neural methods for amortized inference (2024) arXiv:2404.12484

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.